Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review

Susan Armijo Olivo, Luciana Gazzi Macedo, Inae Caroline Gadotti, Jorge Fuentes, Tasha Stanton, David J Magee

Background and Purpose

The methodological quality of randomized controlled trials (RCTs) is commonly evaluated in order to assess the risk of biased estimates of treatment effects. The purpose of this systematic review was to identify scales used to evaluate the methodological quality of RCTs in health care research and summarize the content, construction, development, and psychometric properties of these scales.

Methods

Extensive electronic database searches, along with a manual search, were performed.

Results

One hundred five relevant studies were identified. They accounted for 21 scales and their modifications. The majority of scales had not been rigorously developed or tested for validity and reliability. The Jadad Scale presented the best validity and reliability evidence; however, its validity for physical therapy trials has not been supported.

Discussion and Conclusion

Many scales are used to evaluate the methodological quality of RCTs, but most of these scales have not been adequately developed and have not been adequately tested for validity and reliability. A valid and reliable scale for the assessment of the methodological quality of physical therapy trials needs to be developed.

S Armijo Olivo, BScPT, MSc, is a PhD candidate in the Faculty of Rehabilitation Medicine, University of Alberta, 3-50 Corbett Hall, Edmonton, Alberta, Canada, T6G 2G4. Address all correspondence to Ms Armijo Olivo at: sla4@ualberta.ca.

LG Macedo, BScPT, MSc, is a PhD student in the Back Pain Research Group, Faculty of Health Sciences, University of Sydney, Lidcombe, New South Wales, Australia.

IC Gadotti, BScPT, MSc, is a PhD candidate in the Faculty of Rehabilitation Medicine, University of Alberta.

J Fuentes, BScPT, MSc, is a PhD student in the Faculty of Rehabilitation Medicine, University of Alberta.

T Stanton, BScPT, MSc, is a PhD student in the Back Pain Research Group, Faculty of Health Sciences, University of Sydney.

DJ Magee, BPT, PhD, is Professor, Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta.

[Armijo Olivo S, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. *Phys Ther*. 2008;88:156–175.]

The medical literature is an important resource to guide clinical decision making and research. The evaluation of the methodological quality of studies is an essential step in the process of selecting the best clinical literature. According to Verhagen et al,[1] assessment of methodological quality involves evaluation of internal validity (the degree to which a study's design, conduct, and analysis have minimized biases) and external validity (the extent to which the results of a study can be generalized outside the experimental situation) as well as statistical analysis of primary research. Taken together, these validity constructs are important in determining the methodological quality of primary research. Khan et al[2] pointed out that some reasons for performing quality assessment include: to determine a minimum quality threshold for the selection of the primary studies for a systematic review; to explore differences in quality as an explanation for heterogeneity in study results; to weigh the results in proportion to the quality in meta-analysis; and, more importantly, to guide interpretation of findings, help determine the strength of inferences, and guide recommendations for future research and clinical practice.

The assessment of the quality of controlled trials is essential because variations in the quality of trials can affect the conclusions about the existing evidence.[3] In a review of trials evaluating primarily medical treatments, Moher and colleagues[4,5] demonstrated that trials that did not include features such as blinding and allocation concealment tended to report an exaggerated treatment effect compared with trials that did include these features. These facts emphasize the importance of methodological quality assessment in order to provide accurate information on therapeutic effects.

Trial quality can be divided into 2 categories (which overlap to some degree): methodological quality and reporting quality. *Methodological quality* is defined as "the confidence that the trial design, conduct, and analysis have minimized or avoided biases in its treatment comparisons."[6(p63)] *Reporting quality* is defined as "the provided information about the design, conduct and analysis of the trial."[6(p63)] Inadequate reporting makes the interpretation of studies difficult if not impossible.

Scales and checklists are 2 types of instruments that may be used to assess the methodological quality of clinical trials. These 2 types have been used interchangeably; however, they are actually quite distinct. Scales and checklists both include items measuring quality; however, with a scale, the responses to the individual items are summed to create an overall summary score representing trial quality. For example, with the Physiotherapy Evidence Database (PEDro) scale, a summary quality score can be created by determining the number of "yes" responses to items 2 through 11. A single score of trial quality is obviously appealing because it seems easier to interpret than a series of ticks on a checklist. However, unless accepted guidelines have been followed in scale development and the scale has performed well in subsequent psychometric testing (panel of experts; Delphi procedure; and tested for reliability, responsiveness, and content, construct, and concurrent validity),[7] scale scores may provide a false impression of meaningfulness.

The identification of a reliable and valid scale to assess the literature on a specific topic minimizes the chances of errors when determining the quality of the scientific literature. Thus, the purposes of this systematic review were: (1) to summarize the content, construction, areas of development, and psychometric properties of scales used to evaluate the quality of the randomized controlled trials (RCTs) in health care research and (2) to identify an appropriate scale to evaluate methodological quality of RCTs in the physical therapy and rehabilitation research field.

Method

Search Strategy

A computerized database search was performed to identify relevant articles. For this review, the literature was searched for published studies describing or using a scale to evaluate the methodological quality of RCTs in health care research.

The studies were searched in all languages according to the search strategy of Dickersin and Lefebvre.[8] The search included studies from 1965 up to March 2, 2007, which were obtained through an extensive search of bibliographic databases, including MEDLINE (1966–February 2007, week 4); EMBASE (1988 to 2007, week 8); CINAHL (Cumulative Index to Nursing and Allied Health Literature) (1982–February 2007, week 3); ISI Web of Science (1965–March 2, 2007); EBM (Evidence-Based Medicine) Reviews-Cochrane Central Register of Controlled Trials (CCTR), Cochrane Library, and Best Evidence (1991–first quarter, 2007); All EBM Reviews, comprising the Cochrane Database of Systematic Reviews (CDSR), ACP (American College of Physicians) Journal Club, Database of Abstracts of Reviews of Effects (DARE), and CCTR (1991–first quarter 2007); Global Health (1973–present); and HealthSTAR (1910–February 2007). Key words used in the search were: "scale," "critical appraisal," "critical appraisal review," "appraisal of methodology," "research design review," "quality assessment," "randomized controlled trial," and "RCT." Subject subheadings and some word truncations, ac-

cording to each database, were used as well to map all possible key words. Table 1 provides details on the specific search terms and combinations. The selection of these terms was made with the help of a librarian specializing in health sciences databases. In addition, the literature search also involved manual search of bibliographies of the identified papers, looking for key authors (ie, Jadad, Moher, and Chalmers) and relevant information to meet the objectives of this study. In addition, each study in which the original scale development was described was tracked through the Web of Science database in order to access all studies that referenced the original scale development.

Inclusion and Exclusion Criteria

Published studies reporting on scale development or the psychometric evaluation of a scale were eligible for inclusion. The inclusion criterion was: published scales developed to evaluate methodological quality of RCTs in any area of medical research. No unpublished scales were included. Scales were excluded if they were developed for the analysis of the methodological quality for only one specific systematic review or if the development of the scale was not described and the psychometric properties of the scale were not tested. Based on this information, we believe that, although the inclusion of these excluded scales would greatly increase the number of scales in this review, these scales would not contribute to the results because they were most likely not developed systematically. Checklists that clearly were not designed to be summed also were excluded from this systematic review.

Data Extraction

Five independent reviewers (SAO, LGM, ICG, JF, and TS) screened abstracts and titles for eligibility. When the reviewers felt that the abstract or title was potentially useful, full copies of the article were retrieved and considered for eligibility by all reviewers. When discrepancies occurred between reviewers, the reasons were identified and a final decision was made based on the agreement of all reviewers. STATA software (version 9.0)* was used to calculate kappa agreement between raters (multiple raters) in selecting the studies for this review.

The next step involved extracting the information regarding the content, construction, special features (eg, area of development, number of items, how items were selected for inclusion, time to complete, how scales and items were scored, the use of guidelines), and psychometric properties for each scale. Psychometric properties that were extracted and analyzed were: face validity, content validity, construct validity, concurrent validity, internal consistency, and reproducibility (intrarater and interrater reliability/agreement). We used the definitions of Streiner and Norman[9–11] and the guidelines established by Terwee et al[12] to determine quality of measurement properties. In short, quality of measurement included internal components of validity (ie, content validity: internal consistency, relevance of items and representativeness of items of the scale) as well as the external component of validity (ie, construct validity: the relationship with other tests in a manner that is consistent with theoretically derived hypotheses). In addition, intrarater and interrater reliability (ie, repeatability of measurements taken by the same tester at different times and repeatability of measurements taken by different testers, respectively) also were considered.

Scales were identified as being important to physical therapy if the authors specifically stated that the scale was developed for the physical therapy practice area or was developed by a group of physical therapist researchers, or if the Web of Science search identified that the scale was used in at least 2 physical therapy reviews.

Results

The initial electronic database search of the literature resulted in a total of 7,720 articles (Tab. 1). Of these, 49 were selected as potential studies based on their titles and abstracts. After the complete article was read, however, only 19 of these actually fulfilled the initial criterion.[3,6,13–29] Thirty papers were excluded after reading the complete article.[1,30–58] The main reasons for exclusion were: (1) the tool was a checklist and not a scale, (2) the tool was developed for a single systematic review, and (3) information regarding the scale's construction, development, and psychometric properties was missing or impossible to obtain. The agreement between the reviewers in selecting these articles after applying the inclusion and exclusion criteria was analyzed with a kappa statistic for multiple raters, which resulted in a value of $\kappa = .90$.

Each original scale was tracked in the Web of Science database in order to find any additional information that could add to the psychometric properties of the selected scales. A total of 3,158 articles were found by tracking each scale. From these, 56 new articles were selected from the Web of Science.[59–114]

Thirty-six articles also were obtained through a hand search (ie, bibliographies of the identified papers, key authors).[115–144] Thus, a total of 105 studies were finally included in the study and analyzed. The Figure details the searches.

* Stata Corp, 4905 Lakeway Dr, College Station, TX 77845.

Table 1.
Search Results From Different Electronic Databases[a]

Database	Keywords	Results
MEDLINE	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	2,417
EMBASE	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	1,695
CINAHL	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	858
Web of Science	(1) scale; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9 DocType=Article; Language=All languages.	2,086
EBM Reviews-Cochrane Central Register of Controlled Trials	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	161
CDSR, ACP Journal Club, DARE, CCTR	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	381
Global Health	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	80
HealthSTAR	(1) scale$; (2) critical appraisal tool; (3) critical appraisal; (4) critical appraisal review; (5) appraisal of research methodology; (6) research design review; (7) quality assessment; (8) 1 or 2 or 3 or 4 or 5 or 6 or 7; (9) randomized controlled trial; (10) 8 and 9	312
Total number of citations retrieved by electronic database searches		7,720[b]

[a] CINAHL=Cumulative Index to Nursing and Allied Health Literature, EBM=Evidence-Based Medicine, CDSR=Cochrane Database of Systematic Reviews, ACP=American College of Physicians, DARE=Database of Abstracts of Reviews of Effectiveness, CCTR=Cochrane Central Register of Controlled Trials.
[b] 270 articles were duplicates.

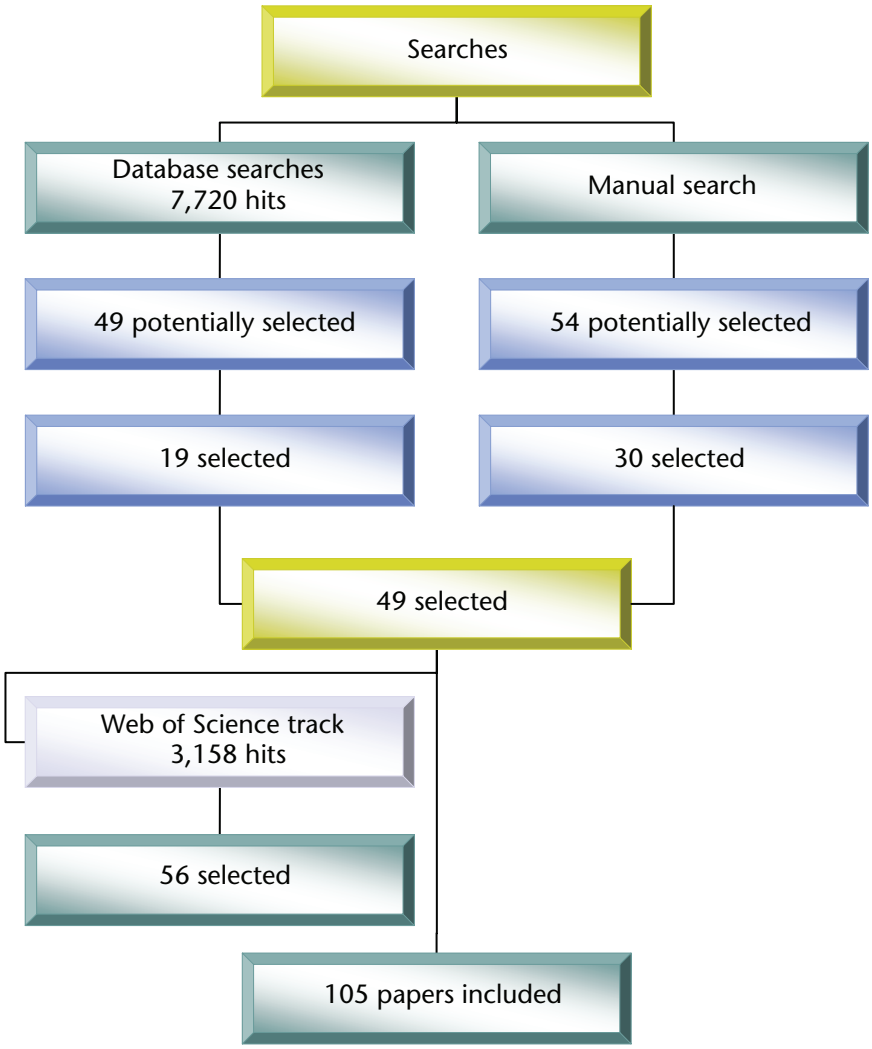

Figure.
QUORUM (Quality of Reporting of Meta-analyses) statement flow diagram of the literature search.

The included studies accounted for 21 scales and their modifications including: Jadad,[15,17,19,25,27,60–62,64,67–69,71,72,78,80,118,120,121,144] Maastricht,[3,62,126] Delphi List,[23,28,79,81–83,90,92,99,107–110] PEDro,[13,74,77,130,131,140,141] Maastricht-Amsterdam List (MAL),[29,85–90,93–97,99–102,104–106,112,113,143] Van Tulder,[81,82,84,91,92,98,103,107,108,110,111,114,142] Bizzini,[26] Chalmers,[14,16,22,63,65,70,117,124,125,127–129,138,139] Reisch,[122,123] Andrew,[115,116] Imperiale,[135] Detsky,[16,59] Cho and Bero,[119] Balas,[133] Sindhu,[20] Downs and Black,[134] Nguyen,[137] Oxford Pain Validity Scale (OPVS),[21] Arrivé,[76] CONSORT,[18,66,73,132] and Yates[75] scales. Details of each scale and their modifications regarding content, construction, special features, and psychometric properties are provided in Tables 2 and 3.

The majority of the adapted scales were based on the Chalmers[14,16,22,63,65,117,124,125,127–129,138,139] and Jadad[15,17,19,25,60,61,64,68,69,71,80,118,120,121,144] scales. The Chalmers Scale was developed to assess clinical trials on the use of aspirin in coronary heart disease studies, and the Jadad Scale was developed for pain research. The Chalmers Scale was modified for different areas and topics: abdominal surgical practice,[127] alcoholism,[129] maxillary sinusitis,[22] breast cancer,[128] periodontal research,[117,124] psychology,[65] pulmonary rehabilitation,[63] and lung cancer.[138] The Jadad Scale also has been adapted for use in many health care areas such as medicine, dentistry, psychology, and physical therapy[15,17,19,25,60,61,64,68,69,71,80,118,120,121,144] (Tab. 2). In addition, according to our Web of Science search, the Jadad Scale was by far the most frequently cited and the most commonly used scale by the health care community.

In the majority of cases, the process of constructing these scales was not

Table 2.
Characteristics of the Scales

Study (Authors, Year)	Area	No. of Items	How Items Were Selected for Inclusion	Validity	Reliability	Time to Complete	Use of Guideline
Scales Used in Physical Therapy							
Jadad Scale Jadad et al[120] (1996)	Pain research	3 items selected related directly to the control of bias	Preliminary items were produced by each rater in 2 weeks based on previous instruments and own judgment. Items with low and high frequency of endorsement (below 15% and above 85%) were excluded.	Face, content, and construct validity. Concurrent validity with Delphi List was found to be good (Spearman r=.63[79] and .71[151]). Concurrent validity with Maastricht Scale was found to be good as well (Spearman r=.78).[9] Concurrent validity with Detsky, Imperiale, Reisch, and van Tulder scales was .78, .61, .64, and .67, respectively.[24]	Interrater reliability: ICC ranged from .48 to 1.00[24,27,28,60–62,64,71,78,99,118] Test retest reliability: ICC=.98[64] Kappa: interrater agreement ranged from .37 to .89[15,16,19,24,72,144]	10 min	Yes
Maastricht Scale de Vet et al[126] (1997)	Physical therapy	16 based on 3 aspects of the study: (1) internal validity, (2) precision of the study, and (3) intervention, effect parameters	Not reported	Face validity. Concurrent validity with Delphi List and Jadad Scale was found to be good (Spearman r=.87 and .78 respectively).[151]	Interrater reliability: ICC=.85[62]	Not reported	Not reported
Verhagen et al[3] (1998)	Physical therapy	15 main items divided into 5 domains with a total of 47 subitems	Not reported		Interrater agreement: ICC=.77 (ICC range=.64–.89)	Not reported	Provided by the author
Delphi List Verhagen et al[23] (1998)	Randomized controlled trials. No area specified	9 items	Items were taken from already known scales (Maastricht and Chalmers). These items were exposed to a group of experts through a Delphi method: recruitment of a Delphi panel and selection of the final items by the panel of experts (3 rounds). The items finally included were obtained through consensus.	Face and content validity. Concurrent validity with Jadad Scale was found to be good. Spearman r=.63,[79] .71.[151] Concurrent validity with Maastricht List was found to be good as well, Spearman r=.87.[151]	Interobserver reliability: kappa=.67,[109] .613,[83] .73,[110] .76,[99] .84,[92] .54,[81] .69,[82] .70,[107] .85.[108] ICC=.88,[28] .88[90]	Not reported	Yes
PEDro Scale Sherrington et al[130] (2000)	Physical therapy	11 items	The 11-item PEDro Scale is based on the 9-item Delphi List developed by Verhagen et al[23] (1998).	Validity only mentioned, type of validity not reported	Kappa ranged from κ=−.611[140] to .88[74,77,130] Interrater reliability: ICC ranged from .39 to .91[13,130,131]	Not reported	Yes

(Continued)

Table 2.
Continued

Study (Authors, Year)	Area	No. of Items	How Items Were Selected for Inclusion	Validity	Reliability	Time to Complete	Use of Guideline
Scales Used in Physical Therapy							
Maastricht-Amsterdam List (MAL) van Tulder et al[143] (1997)	CCBP: back pain	19 items	The MAL is based on the 9-item Delphi List developed by Verhagen et al[23] (1998).	Face and content validity	Interrater reliability: κ=.29,[89] ICC=.82,[94] r=.76,[113] κ range= −.40 to 1.00,[95] κ=.74,[99] κ=.80,[93] κ=.64,[100] κ=.72,[96] κ=.62,[85] κ=.62[101]	Not reported	Not reported
Maastricht-Amsterdam List Modified[29,97,104,106,112]	CCBP: back pain			Face and content validity	Interrater reliability κ=.45–.51,[104] κ=.75,[97] κ=.81,[112] κ=.63,[29] κ=.82,[106] κ=.63[105]	Not reported	Not reported
van Tulder Scale van Tulder et al[142] (2003)	CCBP: back pain	11 items	The van Tulder Scale is based on the MAL developed by van Tulder et al[143] and the 9-item Delphi List developed by Verhagen et al[23] (1998).	Face and content validity. Concurrent validity with Detsky, Imperiale, Jadad, and Reisch scales was .89, .75, .67, and .77, respectively.[24]	Interrater reliability: κ=.66,[86] κ=.29,[89] κ=.42, κ=1.00 in the quality designation, κ=.84,[114] κ=.76,[87] κ=.29,[102] κ=.69,[103] κ=.67,[91] κ=.88,[98] κ=.65,[111] κ=.74[24] ICC=.71,[84] .80[24]	Not reported	Not reported
Bizzini Scale Bizzini et al[26] (2003)	Patellofemoral pain syndrome	4 main criteria with 14 specific criteria	Based on the *Cochrane Collaboration Handbook*, items that are known to add bias or incorporate other factors that would affect the external validity of the paper. Items created by 6 experienced physical therapists.	Face and content validity	ICC for interrater reliability ranged from .64 to .99 within the 4 main criteria. For the total scale score, the ICC was .97.	Not reported	Not reported
Scales for Other Areas of Health Care Research							
Chalmers Scale Chalmers et al[125] (1981)	Pharmacological studies	32 items: Form 1—basic descriptive material (9 items); Form 2—the study protocol (14 items); Form 3—statistical analysis (9 items); Form 4—presentation of results (4 items)	Based on the author's accumulated experience in the analysis of trials (arbitrary)	Face and content validity. Authors state that this scale requires further validation. Concurrent validity with the European Lung Cancer Working Party Scale was excellent (r=.89, P<.001).[70]	Interrater ICCs ranged from .66 to .92[14,16] Test-retest reliability: ICC=.81[14] Intrarater reliability: κ=.66[139]	Not reported	Yes
Reisch Scale Reisch et al[122] (1989)	Therapeutic studies	34 items in 13 domains	Not reported	Face validity. Concurrent validity with Detsky, Imperiale, Jadad, and van Tulder scales was .81, .78, .64, and .77, respectively.[24]	Interrater reliability: ICC=.51[24]	30 min	Not reported
Tyson et al[123] (1983)	Therapeutic studies	29 items	Accepted standards for research design and performance	Face validity	Interrater reliability: r=.99 for the objective score, and r=.71 for the overall quality	Not reported	Not reported
Andrew Scale Andrew[116] (1984)	Diagnostic contrast media	11 items	Not reported	Face validity	Interrater reliability: r=.95	Not reported	Not reported

(Continued)

Table 2.
Continued

Study (Authors, Year)	Area	No. of Items	How Items Were Selected for Inclusion	Validity	Reliability	Time to Complete	Use of Guideline
Scales for Other Areas of Health Care Research							
Andrew Modified Scale Andrew et al[115] (1990)	Clinical trial with x-ray contrast media	11 items	Was modified from another scale from Andrew	Face validity	Interrater reliability: r=.32	Not reported	Not reported
Imperiale Scale Imperiale and McCullough[135] (1990)	Alcoholic hepatitis	5 items	Not reported	Face validity. Concurrent validity with Detsky, Reisch, Jadad, and van Tulder scales was .79, .61, .78, and .75, respectively.[24]	Interrater reliability: ICC=.31 and κ=.43[24]	Not reported	Not reported
Detsky Scale Detsky et al[16] (1992)	18 clinical trials of parenteral nutritional support for patients undergoing major surgery	5 main items	Not reported	Face validity. Concurrent validity with Reisch, Imperiale, Jadad, and van Tulder scales was .79, .78, .81, and .89, respectively.[24]	Interrater reliability: Spearman correlations from .85 to .96 ICC=.92 (95% CI=.81–.98)[16,59] ICC=.80[24]	Not reported	Not reported
Cho and Bero Scale Cho and Bero[119] (1994)	Drug studies	24 items	Based on Spitzer et al Scale (1990).[147] Pretested instrument: 16 items.	Face and concurrent validity (compared with Chalmers Scale)	Interrater reliability: Kendall coefficient (W=.64) and correlation coefficient (r=.89)	30 min	Not reported
Balas Scale Balas et al[133] (1995)	Health services	20 evaluation criteria	Based on recommendations of textbooks and monographs on the methods of clinical trials	Face validity	Interrater kappa agreement=.94	18 min	Not reported
Sindhu Scale Sindhu et al[20] (1997)	Non-pharmacologic nursing interventions	53 items in 15 dimensions	Delphi method. Asking researchers involved in RCTs. Four rounds; only 8 participants in first round and 7 in the last.	Face and content validity and criterion validity compared with Chalmers tool (correlation coefficients were r_p=.94, r_s=.90, and r_i=.35)	No Cronbach alpha reported. Interrater reliability: r_p=.985, r_s=.90, and r_i=.93. Average bias of 10%. Only evaluated with 2 raters.	Shorter than Chalmers Scale, but no time reported	Not reported
Downs and Black Scale Downs and Black[134] (1998)	Public health	5 subscales (reporting, external validity, bias, confounding, and power), divided in 27 items	Based on epidemiological principles, reviews of study designs, and existing checklist for assessment of RCTs	Criterion validity: The quality index score correlated highly with the score of Standards of Reporting Trials Group (SRTG)[152] (r=.90).	Internal consistency of the quality index: Kuder-Richardson 20 (KR20)=0.89 Test-retest reliability for the quality index was r=.88. The interrater reliability of the quality index was r=.75.	20–25 min on average, range of 10–45 min	Yes
Nguyen Scale Nguyen et al[137] (1999)	Dental injuries	100 items divided into internal and external validity	Based on generally accepted methodological criteria	Face validity	Not reported	Not reported	Manual detailing the scoring procedure
Oxford Pain Validity Scale (OPVS) Smith et al[21] (2000)	Acupuncture in chronic neck and back pain	5 main items. Item 5 is subdivided into 4 items	Items are selected for inclusion based on literature (face validity) on pain	Face validity	Not reported	Not reported	It has instructions within the article

(Continued)

Table 2.
Continued

Study (Authors, Year)	Area	No. of Items	How Items Were Selected for Inclusion	Validity	Reliability	Time to Complete	Use of Guideline
Scales for Other Areas of Health Care Research							
Arrivé Scale Arrivé et al[76] (2000)	Clinical studies of radiological examinations	15 items	Items are based on methodological standards that can be applied to any clinical study of radiological examination evaluation, and items generally related to biases commonly observed in radiological studies.	Face validity	Interrater reliability: separate kappa for each item from .74 to .97. For the total score, r=.91	Not reported	Not reported
CONSORT Scale Huwiler-Muntener et al[18] (2002)	General and specialist medical journals	25 items	Based on 1996 CONSORT statement	Face and content validity	94% agreement between reviewers[132]	Not reported	Not reported
Yates Scale Yates et al[75] (2005)	Psychological trials for pain	8 items and 26 subitems	Delphi method: recruitment of a Delphi panel, development of the items by the panel (3 rounds), 3 raters wrote the scale from the items developed by the panel (face-to-face).	Face, content, and discriminative validity	Interrater reliability for experts: ICC=.91 for the full scale, kappa from .07 to .74 for each item separately Novices: ICC=.81	Not reported	Manual detailing the criteria for each item and the coding system

[a] ICC=intraclass correlation coefficient, CCBP= Cochrane Collaboration Back Review Group, PEDro=Physiotherapy Evidence Database, CI=confidence interval, RCT=randomized controlled trial.

rigorous. Only 5 of the 21 scales included in this study were developed using a standardized procedure (Delphi method and a panel of experts).[7] The Jadad,[120] Delphi List,[23] Yates,[75] and Sindhu[20] scales were based on the Delphi method, and Bizzini et al[26] (Bizzini Scale) used a panel of experts along with Cochrane Collaboration Handbook guidelines.[145] Seven scales were developed based on the authors' knowledge or information about methodological quality and standards for research design in the literature (OPVS,[21] Chalmers,[125] Arrivé,[76] Reisch,[122] Downs and Black,[134] Nguyen,[137] and Balas[133] scales). Five of the scales were based on previous checklists. The CONSORT Scale was based on the CONSORT (Consolidated Standards of Reporting Trials) statement[146]; the Cho and Bero Scale[119] was developed based on the Spitzer guidelines[147]; and the PEDro,[130] MAL,[143] and van Tulder[142] scales were based on the Delphi List.[23] Four scales (the Andrew,[116] Detsky,[16] Maastricht,[126] and Imperiale[135] scales) did not report their method of development.

The scales identified as being within the scope of physical therapist practice were the PEDro, Maastricht, Delphi List, MAL, van Tulder, Bizzini, and Jadad scales. The Jadad Scale was the only scale that was not originally developed for physical therapy but has been used in physical therapy reviews.[144,148–150] It is important to point out that 5 of these scales (the Maastricht, Delphi List, MAL, van Tulder, and PEDro scales) are interrelated. The Maastricht Scale was developed in the Department of Epidemiology of the University of Maastricht without using formal scale development techniques. The same group of authors decided to develop a methodological quality scale using formal techniques of scale development; thus, the Delphi List emerged. Since then, the Maastricht Scale seldom has been used.

The Cochrane Collaboration Back Group (CCBG) used the Delphi List as a basis for their analysis, and added some items they found relevant for back pain. This list was then called the "Maastricht-Amsterdam List" (MAL) (19 items) because of the cooperation between the 2 groups. Later, the CCBG updated the MAL and only considered 11 items. It is this list that is considered the "van Tulder List" and has been used by the CCBG and many systematic reviews since 2003. In addition, the PEDro Scale was derived from the Delphi List. Therefore, the Delphi List is the basis for most of the scales used in physical therapy and its items are included in many of the

Table 3.
Summary of the Psychometric Properties of the Analyzed Scales[a]

Scale	Internal Consistency	Face Validity	Content Validity	Criterion Validity	Construct Validity	Reproducibility (Agreement/ Reliability)
Scales for physical therapy area						
Jadad Scale	−	+	+	+*	+	+
Maastricht Scale	−	+	−	+*	−	+
Delphi List	−	+	+	+*	−	+
PEDro Scale	−	+	+	−	−	+
Maastricht-Amsterdam List	−	+	+	−	−	+
van Tulder Scale	−	+	+	+*	−	+
Bizzini Scale	−	+	+	−	−	+
Scales for other areas of health care research						
Chalmers Scale	−	+	+	+*	−	+
Reisch Scale	−	+	−	+*	−	+
Andrew Scale	−	+	−	−	−	+
Imperiale Scale	−	+	−	+*	−	+
Detsky Scale	−	+	−	+*	−	+
Cho and Bero Scale	−	+	+	+*	−	+
Balas Scale	−	+	−	−	−	+
Sindhu Scale	−	+	+	+*	−	+
Downs and Black Scale	+	+	+	+*	−	+
Nguyen Scale	−	+	−	−	−	−
Oxford Pain Validity Scale	−	+	−	−	−	−
Arrivé Scale	−	+	−	−	−	+
CONSORT Scale	−	+	+	−	−	+
Yates Scale	−	+	+	−	+	+

[a] Quality of measurement properties were based on guidelines established by Terwee et al.[12] Asterisk indicates criterion validity established with "no gold standard tools."

scales. However, the scales derived from the Delphi List (the MAL, van Tulder List, and PEDro) have added some items and did not consider further validation of these new scales. Table 4 summarizes the items of the most commonly used scales in physical therapy.

All of the scales that were included in this study had "face validity." The Cho and Bero,[119] Chalmers,[125] and Sindhu[20] scales were tested for content validity. The Jadad,[24,120] Delphi List,[79] Maastricht,[151] Cho and Bero,[119] Sindhu,[20] Detsky,[16,24] Downs and Black,[134] Imperiale,[24] Reisch,[24] Chalmers,[125] and van Tulder[24] scales were tested for criterion (concurrent) validity. However, it is important to remember that criterion (concurrent) validity is used to validate a tool in relation to a gold standard. In this case, the concurrent validity of the above-mentioned scales was tested using the Chalmers, Delphi List, Jadad, CONSORT, Imperiale, Reisch, van Tulder, Standards of Reporting Trials Group (SRTG),[152] and European Lung Cancer Working Party[70] tools, which are not recognized as gold standards in this field. Based on this fact, the concurrent validity of the Jadad, Delphi List, Maastricht, Cho and Bero, Sindhu, Detsky, Downs and Black, Imperiale, Reisch, Chalmers, and van Tulder scales may be inappropriate.[9]

The Jadad[120] and Yates[75] scales were tested for construct validity. Construct validity, as mentioned previously, refers to the extent to which scores of a scale are based on hypothetical

Table 4.
Items Included in the Analyzed Scales Used in Physical Therapy

Items Included in the Scales	Jadad	Maastricht	Delphi	van Tulder	Maastricht-Amsterdam	PEDro	Bizzini	Total	%
Patient selection									
Inclusion and exclusion criteria clearly defined/eligibility criteria specified		X	X		X	X	X	5	71.4
Study described as randomized	X	X				X	X	4	57.1
Randomization method performed			X				X	2	28.6
Method of randomization described and appropriate				X	X			2	28.6
Method of randomization concealed		X	X	X		X		4	57.1
Baseline comparability (group equivalence, homogeneity) regarding the most important prognostic indicators		X	X	X	X	X	X	6	85.7
Blinding									
Study described as double blind	X	X				X		3	42.9
Method of blinding described and appropriate	X							1	14.3
Blinding of investigator/assessor		X	X	X	X	X	X	6	85.7
Blinding of subjects/patients		X	X	X	X	X		5	71.4
Blinding of therapists/care provider		X	X	X	X	X		5	71.4
Blinding of the outcome (results)	X							1	14.3
Interventions									
Treatment protocol adequately described for the treatment and control groups (eg, frequency, intensity)		X			X		X	3	42.9
Control and placebo adequate							X	1	14.3
Co-interventions avoided				X	X		X	3	42.9
Co-interventions reported for each group separately					X			1	14.3
Control for co-interventions in design					X			1	14.3
Testing of subject adherence to treatment protocol		X						1	14.3
Adherence acceptable in all groups				X	X			2	28.6
Description of withdrawals and dropouts	X	X		X		X	X	5	71.4
Withdrawal/dropout rate described and acceptable				X	X	X		3	42.9

(Continued)

Table 4.
Continued

Items Included in the Scales	Jadad	Maastricht	Delphi	van Tulder	Maastricht-Amsterdam	PEDro	Bizzini	Total	%
Reasons for dropouts							X	1	14.3
Patient follow-up details reported		X						1	14.3
Follow-up period adequate							X	1	14.3
Short follow-up measurement performed					X			1	14.3
The timing of the outcome assessment was comparable in all groups				X	X			2	28.6
Outcomes									
Outcome measures described							X	1	14.3
Relevant outcomes were used					X			1	14.3
Validity reported for main outcome measures							X	1	14.3
Responsiveness for main outcome measures							X	1	14.3
Reliability reported for main outcome measures							X	1	14.3
Use of objective outcome measures		X				X		2	28.6
Statistics									
Descriptive measures (point estimates and measures of variability) identified and reported for the primary outcome		X	X		X	X		4	57.1
Appropriate statistical analysis used		X				X	X	3	42.9
Sample size calculation performed prior to initiation of the study		X						1	14.3
Adequate sample size							X	1	14.3
Sample size described for each group					X			1	14.3
Intention-to-treat analysis used		X	X	X	X	X	X	6	85.7

grounds and should be tested by predefined hypotheses (eg, expected correlations between measures, expected behavior of scales in a determined situation).[9] For example, the Jadad and Yates scales were tested to confirm whether they could discriminate between articles that had good or bad methodological quality and were previously judged by a group of experts (construct validity evidence).

Two scales (the OPVS and Nguyen scales) were not tested for reproducibility (ie, agreement and intrarater and interrater reliability) and internal consistency. Values such as intraclass correlation coefficient, kappa, Kendall coefficient, and Pearson correlation coefficient were used to analyze interrater reliability. Only one scale (Downs and Black) has reported internal consistency values.[134] In general, scales had interrater reliability that ranged from .32 (poor) to 1.0 (excellent). Intrarater reliability was only reported for one of the modified Chalmers scales (κ=.66).[139] Test-retest reliability was reported for the Downs and Black (r=.88),[134] Jadad (intraclass correlation coefficient=.98),[64] and Chalmers (intraclass correlation coefficient=.81)[14] scales. The scales that

have been tested for reliability in different areas are the Jadad[16,19,28,61,64,71,78,99,118,144] and Delphi List[28,81–83,90,92,99,107–110] scales (Tab. 2). The scale that presented the worst reliability was the Andrew Scale (r=.32).[115]

The Jadad Scale presented the best validity evidence and has been tested for reliability in different settings. However, the Delphi List and Yates scales have been developed based on high standards as well (Delphi procedure, panel of experts, and tested for validity and reliability).[7,12] The Delphi List also has been used in many areas; however, it has not been as popular (cited 180 times) as the Jadad Scale (cited 1,780 times). The Yates Scale is a recently published scale that was created for use in cognitive behavior therapy for chronic pain. Its use has been limited (only cited once by the same group of authors).

Regarding the scales used specifically in the physical therapy area, the Delphi List, along with the Jadad Scale, has greater validity evidence compared with the other scales (the MAL, van Tulder, PEDro, and Bizzini scales). However, the Delphi List lacks internal consistency and construct validity. These psychometric properties are of importance in a scale because they indicate that the construct, in this case "methodological quality," is fully represented by the items of the scale (internal consistency), and that the scores of a scale are based on hypothetical grounds and should behave based on predefined hypotheses.[7,9] Scales used in the physical therapy area had interrater reliability that ranged from .37 (poor) to 1.0 (excellent). The reliability values found for each scale depended on the setting used, the raters' characteristics, the length of the scale, and also the training given on how to use a specific scale.

Discussion

This systematic review evaluated the content, construction, areas of development, and psychometric properties of scales used to evaluate the quality of the RCTs in health care research. The findings of this study demonstrated that a large number of scales and modified scales are available in the literature to evaluate methodological quality in different health care areas.

The scales analyzed in this review differed in several aspects, such as area of development, complexity, length, type of items, and importance given to the included items. The scale modifications were performed in the majority of the cases to adapt a scale to a specific topic. The primary scales were normally developed with the objective of analyzing the quality of RCTs in a specific area, and the items cover topics that are important to that area. Chalmers and colleagues,[125] for example, developed a scale to analyze clinical trials on the use of aspirin in coronary heart disease studies and included an item about the taste and appearance of the drug in the scale. This item is very important for this type of study because it provides information regarding the true blinding of the patients. However, this item would become completely inappropriate when using this scale to analyze the quality of a nonpharmacological study. Based on this fact, many authors modified an original scale so they could use it in a systematic review on a topic different from the one for which the scale was originally developed. However, if one single item is added to or taken from a scale, if the weighting system is changed, or if any other minor change is performed, the psychometric properties of the original scale may be no longer applicable.

A modified scale developed from a validated and reliable primary scale cannot be considered valid and reliable unless it is tested for validity and reliability itself. According to Streiner and Norman, "modifications of existing scales often require new validity studies."[9(p186)] This means that the psychometric properties of the modified scale have to be assessed to ensure that the new scale can actually identify papers with good or bad methodological quality. Most of the scales used in the physical therapy area are modifications of the Delphi List. These scales, however, did not follow any further validation process and, therefore, cannot be considered to be as valid as the original (the Delphi List). The use of modified scales is a step forward in creating scales specific to each area; however, they should be used with caution because of the lack of information about their construction, applicability, and psychometric properties. Some reports[136,153] did not account for this fact and considered a modified scale to be a new scale. However, this misunderstanding added confusion about the number of existing scales in the literature. Our systematic review grouped all of the original scales and their modifications to highlight the fact that there are few original scales with clear and reported psychometric properties. Nevertheless, many adaptations of these scales have occurred without considering a new validation process. These unvalidated "new" scales have been freely used in health care research, which makes the interpretation and the validity of the results of these scales even more complex and open to question.

Quality assessment instruments have to be developed according to the principles used to create the scales. However, most instruments analyzed in this study have not been developed rigorously. Our results are in agreement with those obtained by Moher et al.[6] It has been suggest-

ed[5,154] that some specific issues must be addressed when developing a scale: definition of the quality construct, definition of the scope and purpose of quality assessment, definition of the population of end-users (background), selection of raters, and trial scoring (open or blind). Validity evidence such as the internal component of validity (internal consistency, relevance of items, and representativeness of items of the scale) as well as the external component of validity (ie, the relationship with other tests) are needed to support the use of scales to measure the methodological quality of RCTs. Intrarater and interrater reliability (ie, repeatability of measurements taken by the same tester at different times and reproducibility of measurements taken by different testers, respectively) also are important considerations when developing a tool.[155]

As with any procedure, assessment of methodological quality is prone to bias. Thus, in order to be consistent and avoid bias, researchers should use robust tools that are able to objectively evaluate methodological quality. However, which issues are relevant to consider for quality assessment tools? According to our results, randomization is one of the most commonly used items across different scales to measure methodological quality. It has been shown that lack of randomization can change the treatment effects.[2] Chalmers et al[156] found that trials that were not randomized had a 58.1% difference in case-fatality rates in the treatment group compared with the control group, whereas trials with randomization had a 24.4% difference and blinded randomized trials had a 8.8% difference. Allocation concealment was considered in 44.4% of the analyzed scales in our systematic review. Inadequate allocation concealment has been shown to produce a 37% exaggeration of treatment effects in clinical trials that reported allocation concealment inadequately compared with trials that reported it adequately.[5,157] According to these findings, therefore, randomization as well as allocation concealment should be evaluated when assessing methodological quality because they eliminate study selection and confounding biases.[158]

Blinding was another item frequently used by the scales studied in this paper and has been found to be an important consideration when evaluating methodological quality. Schulz et al[159] found that trials with no double-blinding procedure increased the treatment effects by 17%. These results are in agreement with those of Colditz and colleagues.[160,161] However, according to the results of Moher et al,[5] double blinding did not significantly affect the estimates of effect.

In addition, sample size calculation was an item frequently used by the scales, and it has been shown to be important for methodological quality. Trials with small sample sizes have more of a risk for a type II error. For example, Freiman et al[162] and Moher et al[163] found that most of the trials with negative results did not have a large enough sample size, leading to wrong conclusions and wasting of resources. Thus, sample size is an important issue when evaluating methodological quality because if trials are not adequate powered, they will probably not show an effect.

Other scale items considered to evaluate methodological quality were items such as appropriate statistical analysis, description of withdrawals and dropouts, baseline comparability, definition of inclusion and exclusion criteria, use of intention-to-treat analysis,[164] and outcomes objectivity. These items empirically affect the quality of the trial; however, no studies supporting this information have been performed. Nevertheless, basic methodological standards[2,43,146] support the inclusion of these items in quality assessment tools. Future studies should evaluate the influence of these issues in estimates of treatment in different areas because most of the information on this topic comes from medicine[5,157–159,165] and may be not applicable to physical therapy and rehabilitation.[166] This research could guide improvement in the methodological quality assessment scales in physical therapy and rehabilitation. Future research evaluating the relationship between design characteristics and treatment effect sizes in physical therapy is urgently needed in order to determine important items in the scales with a scientific basis for their construction. This also will help to guide research planning.

Although this systematic review focused on scales that summarize their results in a final score, the use of a summary score is still debatable. According to Balk et al,[165] it is clear that the assessment of the methodology of studies is useful to better understand the quality of RCTs and meta-analyses; therefore, summary scores could be useful as a tool for clinicians to evaluate the strength of individual studies. This may be important because scales are used not only by researchers but also by clinicians and students who may not have sufficient knowledge to make their own judgments about quality based on individual items. In this case, the use of a summary score may facilitate an understanding of the literature such as is done when using the PEDro database. Most importantly, clinicians have to apply scientific knowledge to practice, and they need to have a simple and interpretable method to be confident about the quality of the research in order to apply this knowledge in clinical practice.

Most of the studies relating a summary quality score and its effect on outcomes have controversial results. Although the use of a summary quality score in meta-analysis has been suggested, its influence on outcome remains unclear and needs further research.[136,153,165,166] The results of this review showed that many scales have not been developed with a systematic rigor and have been validated only for the areas for which they were originally designed. Therefore, the results of the articles that question the usefulness of a summary score are put into question. Herbison et al,[153] for example, contended that the use of a quality score in adjustments of meta-analysis were not sensitive enough to identify high-quality trials from low-quality trials and that each scale came to a different result. However, lack of sensitivity may be due to the low quality of the scales used and not necessarily related to the summary scores. The scales used may not have represented the construct of "methodological quality," perhaps because they did not reach suitable standards for its validation. Therefore, caution is necessary when using summary scores before a valid and systematically developed scale is used to test the usefulness of a scoring system.

Methodological Quality Scales and Physical Therapy

The use of quality scales to assess RCTs in physical therapy is a relevant issue. As mentioned above, 7 scales have been used to assess physical therapy trials (ie, the Jadad, Maastricht, Delphi List, MAL, van Tulder, PEDro, and Bizzini scales). The Jadad Scale, because of its brevity and widespread use, the Delphi List, and PEDro scales have been used most frequently for this purpose.[144,148–150] In addition, the MAL (1997–2003) and van Tulder (2003–present) scales have been used for physical therapy reviews from the CCBG. The Maastricht Scale was developed for physical therapy; however, it has not commonly been used since the appearance of Delphi List and its modifications (MAL, van Tulder, PEDro). Despite the fact that the Jadad Scale is a pain-validated tool that was not created to evaluate specific information related to physical therapy, it has been widely used in physical therapy research,[17,144,149,150,167–171] and its authors suggest that this scale could be used in other areas. However, until now, there has not been any validation on the use of this scale in areas other than pain.

The Jadad Scale focuses only on randomization, blinding, and withdrawals and dropouts to evaluate methodological quality of primary research. Only 2 of these 3 items would always be applicable to physical therapy because the nature of physical therapy interventions (eg, manual therapy, exercises) does not allow for blinding of the therapists and or the patients on some occasions. Proper double blinding, therefore, is unlikely to be accomplished for most physical therapy trials, and, as a result, this item becomes irrelevant and most likely will not contribute to the ability of the scale to determine the quality of RCTs in physical therapy. Furthermore, the Jadad Scale does not include any item about treatment protocol specifications, treatment adherence, or treatment integrity, which are important issues in physical therapy. As such, the Jadad Scale may not provide the most comprehensive measure of methodological quality for physical therapy trials.

Currently, most clinical trials include the Jadad Scale items in their methodology in order to accomplish with good methodological quality. Herbison et al[153] found that, for a large proportion of studies they analyzed, the Jadad Scale did not allow them to divide the studies into "high" and "low" quality and concluded that this scale might not be responsive enough to distinguish between different levels of quality. Therefore, the use of Jadad Scale and its validity should be reassessed, not only for drug trials but also for different areas of research (eg, physical therapy, nonpharmacological trials).

Conversely, the PEDro, Maastricht, Delphi List, MAL, van Tulder, and Bizzini scales appear to be more connected to the physical therapy field. For example, it has been shown that the PEDro Scale, a modification of the Delphi List, offers a more comprehensive measure of methodological quality in stroke rehabilitation literature compared with the Jadad Scale.[148] Furthermore, in addition to the blinding component, the PEDro Scale assesses the methodological quality of a study based on other important criteria, such as concealed allocation, intention-to-treat analysis, and adequacy of follow-up. As such, the PEDro Scale appears to be a more useful tool to assess the methodological quality of physical therapy trials.

Because physical therapy clinical trials are much more complex than a pharmacological RCT, the physical therapy–related scales should carefully take into account not only patient adherence and standardization of the treatment protocol but also the precise performance of the intervention as well as the validity, reliability, and responsiveness of the tests and measurements included in the trial. None of these variables have been considered in the scales regularly used (the Jadad, Delphi List, van Tulder, and PEDro scales) to assess the methodological quality of trials in physical therapy. As it has been noted (Tab. 4), all of these scales have neglected the intervention items as well the validity, reliability, and responsiveness of the outcomes used. The Maastricht and Bizzini scales consider parameters of treatment such as frequency, inten-

sity, duration, and adherence, making these scales more comprehensive; however, these scales lack the higher levels of validity evidence. Therefore, a content analysis of the current scales used in physical therapy clearly identifies a gap around the issue of treatment implementation and outcome measurements, which are often relevant for physical therapy.

Another relevant issue is the different interpretation of the items. For example, the Delphi and Bizzini scales ask only whether randomization was performed, but the MAL and van Tulder scales require the reviewer to determine whether the method of randomization is appropriate. In addition, regarding baseline comparability of the most important prognostic factors, the Delphi List requires the reviewer to determine the comparable item, whereas the MAL specifically requires adequate description of the patients' age and duration of complaints, percentage of patients with pain, and main outcome measures to evaluate similarity. Thus, these items could elicit different responses and scores, depending on the scale. Therefore, precise guidelines with unified criteria should exist in order to provide the same information.

Based on the information given about the scales, the quality of the existing scales (such as validation testing) needs to be improved or a new tool closely related to physical therapy practice needs to be developed and include all items and relevant issues related to rehabilitation and physical therapy. In addition, this new tool must include the concept of quality in its broadest sense and be tested for validity and reliability across different areas of physical therapy practice (eg, orthopedics, neurology, respiratory care) in order to make sure that this tool is relevant and applicable to different areas of physical therapy research.

Conclusion

Based on the findings of this systematic review, many scales are being used to evaluate the methodological quality of RCTs in health care research. Most of the analyzed scales did not follow methodological standards during development[7] and have not been tested for validity and reliability in the areas to which they have been applied. Our findings indicate that no scale that is being used to evaluate the quality of physical therapy research has been subjected to a scientifically rigorous development or to testing for validity and reliability. Therefore, readers should be careful when using a scale to assess methodological quality of primary research articles. Scale limitations should be taken into consideration and the information provided by scales should be interpreted with caution. Future research looking at developing a new scale to evaluate the methodological quality of RCTs in the physical therapy should take into consideration the results of this review regarding the flaws and limitations of the existing scales.

All authors provided concept/idea/research design and writing. Ms Armijo Olivo, Ms Macedo, Ms Gadotti, Mr Fuentes, and Ms Stanton provided data collection and data analysis. Ms Armijo Olivo provided project management.

The authors acknowledge the following agencies and funding: Alberta Provincial CIHR Training Program in Bone and Joint Health, Izaak Walton Killam scholarship from the University of Alberta, Canadian Institutes of Health Research, Government of Chile (MECESUP Program), University Catholic of Maule, Endeavour International Postgraduate Research Scholarships from the University of Sydney, University of Sydney International Research Scholarship, Strathcona Physiotherapy Scholarship, Province of Alberta Graduate Scholarship, and Ann Collins Whitmore Memorial Award from the Physiotherapy Foundation of Canada. The authors also acknowledge Sandra Shores, librarian specializing in health sciences databases at the University of Alberta, for her assistance.

A platform presentation of this study was presented at the 15th International Conference of the World Confederation for Physical Therapy; June 3, 2007; Vancouver, BC, Canada.

This article was submitted May 14, 2007, and was accepted August 27, 2007.

DOI: 10.2522/ptj.20070147

References

1 Verhagen AP, de Vet HC, de Bie RA, et al. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol.* 2001;54:651–654.

2 Khan KS, ter Riet G, Popay J, et al. Stage II: conducting the review, phase 5: study quality assessment. In: Khan KS, ter Riet G, Glanville J, et al, eds. *Undertaking Systematic Reviews of Research Effectiveness: CRD's Guidance for Those Carrying Out or Commissioning Reviews.* 2nd ed. York, United Kingdom: NHS Centre for Reviews and Dissemination, University of York; 2001:1–20. CRD Report 4.

3 Verhagen AP, de Vet HC, de Bie RA, et al. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol.* 1998; 51:335–341.

4 Moher D, Jadad AR, Klassen TP. Guides for reading and interpreting systematic reviews, III: how did the authors synthesize the data and make their conclusions? *Arch Pediatr Adolesc Med.* 1998;152: 915–920.

5 Moher D, Cook DJ, Jadad AR, et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess.* 1999;3(12):i–iv, 1–98.

6 Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials.* 1995;16:62–73.

7 Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. Oxford, United Kingdom: Oxford University Press; 2003.

8 Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ.* 1994;309:1286–1291.

9 Streiner DL, Norman GR. Validity. In: *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. Oxford, United Kingdom: Oxford University Press; 2004:172–193.

10 Streiner DL, Norman GR. Reliability. In: *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. Oxford, United Kingdom: Oxford University Press; 2004:126–152.

11 Streiner D, Norman G. Measuring change. In: *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. Oxford, United Kingdom: Oxford University Press; 2004: 194–212.

12 Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007; 60:34–42.

13 Maher CG, Sherrington C, Herbert RD, et al. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther.* 2003;83:713–721.

14 Berard A, Andreu N, Tetrault J, et al. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Ann Epidemiol.* 2000;10:498–503.

15 Clark HD, Wells GA, Huet C, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. *Control Clin Trials.* 1999;20:448–452.

16 Detsky AS, Naylor CD, O'Rourke K, et al. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol.* 1992;45: 255–265.

17 Gummesson C, Atroshi I, Ekdahl C. The quality of reporting and outcome measures in randomized clinical trials related to upper-extremity disorders. *J Hand Surg Am.* 2004;29:727–734; discussion 735–737.

18 Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA.* 2002;287:2801–2804.

19 Latronico N, Botteri M, Minelli C, et al. Quality of reporting of randomised controlled trials in the intensive care literature: a systematic analysis of papers published in Intensive Care Medicine over 26 years. *Intensive Care Med.* 2002;28: 1316–1323.

20 Sindhu F, Carpenter L, Seers K. Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. *J Adv Nurs.* 1997;25:1262–1268.

21 Smith LA, Oldman AD, McQuay HJ, Moore RA. Teasing apart quality and validity in systematic reviews: an example from acupuncture trials in chronic neck and back pain. *Pain.* 2000;86:119–132.

22 Stalman W, van Essen GA, van der Graaf Y, de Melker RA. Maxillary sinusitis in adults: an evaluation of placebo-controlled double-blind trials. *Fam Pract.* 1997;14:124–129.

23 Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol.* 1998;51:1235–1241.

24 Colle F, Rannou F, Revel M, et al. Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. *Arch Phys Med Rehabil.* 2002;83:1745–1752.

25 Esposito M, Coulthard P, Worthington HV, Jokstad A. Quality assessment of randomized controlled trials of oral implants. *Int J Oral Maxillopac Implants.* 2001;16:783–792.

26 Bizzini M, Childs JD, Piva SR, Delitto A. Systematic review of the quality of randomized controlled trials for patellofemoral pain syndrome. *J Orthop Sports Phys Ther.* 2003;33:4–20.

27 Bhandari M, Richards RR, Sprague S, Schemitsch EH. Quality in the reporting of randomized trials in surgery: is the Jadad scale reliable? *Control Clin Trials.* 2001;22:687–688.

28 Boutron I, Tubach F, Giraudeau B, Ravaud P. Methodological differences in clinical trials evaluating nonpharmacological and pharmacological treatments of hip and knee osteoarthritis. *JAMA.* 2003;290:1062–1070.

29 Staiger TO, Gaster B, Sullivan MD, Deyo RA. Systematic review of antidepressants in the treatment of chronic low back pain. *Spine.* 2003;28:2540–2545.

30 Maher CG, Sherrington C, Elkins M, et al. Challenges for evidence-based physical therapy: accessing and interpreting high-quality evidence on therapy. *Phys Ther.* 2004;84:644–654.

31 Bath FJ, Owen VE, Bath PM. Quality of full and final publications reporting acute stroke trials: a systematic review. *Stroke.* 1998;29:2203–2210.

32 Bell DM, Nappi J. Myocardial infarction in women: a critical appraisal of gender differences in outcomes. *Pharmacotherapy.* 2000;20:1034–1044.

33 Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in *The Journal of Bone and Joint Surgery* from 1988 through 2000. *J Bone Joint Surg Am.* 2002;84:388–396.

34 Burback D, Molnar FJ, St John P, Man-Son-Hing M. Key methodological features of randomized controlled trials of Alzheimer's disease therapy: minimal clinically important difference, sample size and trial duration. *Dement Geriatr Cogn Disord.* 1999;10:534–540.

35 Cook D. Evidence-based critical care medicine: a potential tool for change. *New Horiz.* 1998;6:20–25.

36 Cummins RO, Chamberlain D, Hazinski MF, et al. Recommended guidelines for reviewing, reporting, and conducting research on in-hospital resuscitation: the in-hospital "Utstein style." A statement for health care professionals from the American Heart Association, the European Resuscitation Council, the Heart and Stroke Foundation of Canada, the Australian Resuscitation Council, and the Resuscitation Councils of Southern Africa. *Acad Emerg Med.* 1997;4: 603–627.

37 Doig GS. Interpreting and using clinical trials. *Crit Care Clin.* 1998;14:513–524.

38 Edwards AG, Russell IT, Stott NC. Signal versus noise in the evidence base for medicine: an alternative to hierarchies of evidence? *Fam Pract.* 1998;15:319–322.

39 Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med.* 2001;344:1594–1602. Erratum in *N Engl J Med*. 2001; 345:304.

40 Hrobjartsson A, Gotzsche PC. Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment. *J Intern Med.* 2004;256:91–100.

41 Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry.* 1995;52:867–871.

42 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001; 134:663–694.

43 Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. *Jt Comm J Qual Improv.* 1999;25:470–479.

44 Moher D, Jones A, Lepage L; CONSORT Group. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA.* 2001;285:1992–1995.

45 Moher D, Soeken K, Sampson M, et al. Assessing the quality of reports of randomized trials in pediatric complementary and alternative medicine. *BMC Pediatr.* 2002;2:3.

46 Moseley AM, Herbert RD, Sherrington C, Maher CG. Evidence for physiotherapy practice: a survey of the Physiotherapy Evidence Database (PEDro). *Aust J Physiother.* 2002;48:43–49.

47 Moyer A, Finney JW. Randomized versus nonrandomized studies of alcohol treatment: participants, methodological features and posttreatment functioning. *J Stud Alcohol.* 2002;63:542–550.

48 Rinck GC, van den Bos GA, Kleijnen J, et al. Methodologic issues in effectiveness research on palliative cancer care: a systematic review. *J Clin Oncol.* 1997; 15:1697–1707.

49 Shilliday IR, Sherif M. Calcium channel blockers for preventing acute tubular necrosis in kidney transplant recipients. *Cochrane Database Syst Rev.* 2004;(1): CD003421.

50 Sjogren P, Halling A. Quality of reporting randomised clinical trials in dental and medical research. *Br Dent J.* 2002;192: 100–103.

51 Sonis J, Joines J. The quality of clinical trials published in The Journal of Family Practice, 1974–1991. *J Fam Pract.* 1994; 39:225–235.

52 Speroff T, James BC, Nelson EC, et al. Guidelines for appraisal and publication of PDSA quality improvement. *Qual Manag Health Care.* 2004;13:33–39.

53 Stein DJ, Ipser JC, Balkom AJ. Pharmacotherapy for social phobia. *Cochrane Database Syst Rev.* 2004;(4):CD001206.

54 Stobart K, Iorio A, Wu JK. Clotting factor concentrates given to prevent bleeding and bleeding-related complications in people with hemophilia A or B. *Cochrane Database Syst Rev*. 2005;(2):CD003429.

55 Stones RW, Mountfield J. Interventions for treating chronic pelvic pain in women. *Cochrane Database Syst Rev.* 2000; (4):CD000387.

56 Clark P, Tugwell P, Bennett K, et al. Injectable gold for rheumatoid arthritis. *Cochrane Database Syst Rev*. 2000; (2):CD000520.

57 Bouter LM. Prevalence of methodologic errors in rehabilitation research. *J Rehabil Sci.* 1994;7(suppl):60–62.

58 McNeely ML, Torrance G, Magee DJ. A systematic review of physiotherapy for spondylolysis and spondylolisthesis. *Man Ther.* 2003;8:80–91.

59 Morrison LJ, Brooks S, Sawadsky B, et al. Prehospital 12-lead electrocardiography impact on acute myocardial infarction treatment times and mortality: a systematic review. *Acad Emerg Med.* 2006; 13:84–89.

60 Baron G, Boutron I, Giraudeau B, Ravaud P. Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. *Arthritis Rheum.* 2005;52:1858–1865.

61 Bender JS, Halpern SH, Thangaroopan M, et al. Quality and retrieval of obstetrical anaesthesia randomized controlled trials. *Can J Anaesth.* 1997;44:14–18.

62 Brockow T, Hausner T, Dillner A, Resch KL. Clinical evidence of subcutaneous CO2 insufflations: a systematic review. *J Altern Complement Med.* 2000;6: 391–403.

63 Cambach W, Wagenaar RC, Koelman TW, et al. The long-term effects of pulmonary rehabilitation in patients with asthma and chronic obstructive pulmonary disease: a research synthesis. *Arch Phys Med Rehabil.* 1999;80:103–111.

64 Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med.* 2001;135:982–989.

65 McCusker J, Cole M, Keller E, et al. Effectiveness of treatments of depression in older ambulatory patients. *Arch Intern Med.* 1998;158:705–712.

66 Moberg-Mogren E, Nelson DL. Evaluating the quality of reporting occupational therapy randomized controlled trials by expanding the CONSORT criteria. *Am J Occup Ther.* 2006;60:226–235.

67 Rintelen B, Neumann K, Leeb BF. A meta-analysis of controlled clinical studies with diacerein in the treatment of osteoarthritis. *Arch Intern Med.* 2006;166: 1899–1906.

68 Salerno SM, Browning R, Jackson JL. The effect of antidepressant treatment on chronic back pain: a meta-analysis. *Arch Intern Med.* 2002;162:19–24.

69 Salpeter SR, Greyber E, Pasternak GA, Salpeter EE. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus: systematic review and meta-analysis. *Arch Intern Med.* 2003; 163:2594–2602.

70 Sculier JP, Berghmans T, Castaigne C, et al. Maintenance chemotherapy for small cell lung cancer: a critical review of the literature. *Lung Cancer.* 1998;19: 141–151.

71 Shakespeare TP, Thiagarajan A, Gebski V. Evaluation of the quality of radiotherapy randomized trials for painful bone metastases: implications for future research design and reporting. *Cancer.* 2005;103: 1976–1981.

72 Sutherland SE, Browman GP. Prophylaxis of oral mucositis in irradiated head-and-neck cancer patients: a proposed classification scheme of interventions and meta-analysis of randomized controlled trials. *Int J Radiat Oncol Biol Phys.* 2001;49:917–930.

73 Tiruvoipati R, Balasubramanian SP, Atturu G, et al. Improving the quality of reporting randomized controlled trials in cardiothoracic surgery: the way forward. *J Thorac Cardiovasc Surg.* 2006;132: 233–240.

74 Van Peppen RP, Kwakkel G, Wood-Dauphinée S, et al. The impact of physical therapy on functional outcomes after stroke: what's the evidence? *Clin Rehabil.* 2004;18:833–862.

75 Yates SL, Morley S, Eccleston C, de C Williams A. A scale for rating the quality of psychological trials for pain. *Pain.* 2005;117:314–325.

76 Arrivé L, Renard R, Carrat F, et al. A scale of methodological quality for clinical studies of radiologic examinations. *Radiology.* 2000;217:69–74.

77 Van Peppen RPS, Kortsmit M, Lindeman E, Kwakkel G. Effects of visual feedback therapy on postural control in bilateral standing after stroke: a systematic review. *J Rehabil Med.* 2006;38:3–9.

78 Linde K, Clausius N, Ramirez G, et al. Are the clinical effects of homoeopathy placebo effects? A meta-analysis of placebo-controlled trials. *Lancet.* 1997;350: 834–843.

79 Wahlbeck K, Tuunainen A, Gilbody S, Adams CE. Influence of methodology on outcomes of randomised clozapine trials. *Pharmacopsychiatry.* 2000;33:54–59.

80 Silva Filho CR, Saconato H, Conterno LO, et al. Assessment of clinical trial quality and its impact on meta-analyses [in Portuguese]. *Rev Saude Publica.* 2005;39: 865–873.

81 Brouwer RW, Jakma TS, Bierma-Zeinstra SM, et al. Osteotomy for treating knee osteoarthritis. *Cochrane Database Syst Rev.* 2005;(1):CD004019.

82 Brouwer RW, Jakma TS, Verhagen AP, et al. Braces and orthoses for treating osteoarthritis of the knee. *Cochrane Database Syst Rev.* 2005;(1):CD004020.

83 Cignacco E, Hamers JP, Stoffel L, et al. The efficacy of non-pharmacological interventions in the management of procedural pain in preterm and term neonates: a systematic literature review. *Eur J Pain.* 2007;11:139–152.

84 Dekker A, Bulley S, Beyene J, et al. Meta-analysis of randomized controlled trials of prophylactic granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor after autologous and allogeneic stem cell transplantation. *J Clin Oncol.* 2006;24: 5207–5215.

85 Furlan AD, Brosseau L, Imamura M, Irvin E. Massage for low-back pain: a systematic review within the framework of the Cochrane Collaboration Back Review Group. *Spine.* 2002;27:1896–1910.

86 Gagnier JJ, van Tulder MW, Berman B, Bombardier C. Herbal medicine for low back pain: a Cochrane review. *Spine.* 2007;32:82–92.

87 Hayden JA, van Tulder MW, Malmivaara A, Koes BW. Exercise therapy for treatment of non-specific low back pain. *Cochrane Database Syst Rev.* 2005; (3):CD000335.

88 Henrotin YE, Cedraschi C, Duplan B, et al. Information and low back pain management: a systematic review. *Spine.* 2006;31:E326–E334.

89 Jacobs WC, Clement DJ, Wymenga AB. Retention versus removal of the posterior cruciate ligament in total knee replacement: a systematic literature review within the Cochrane framework. *Acta Orthop.* 2005;76:757–768.

90 Knols R, Aaronson NK, Uebelhart D, et al. Physical exercise in cancer patients during and after medical treatment: a systematic review of randomized and controlled clinical trials. *J Clin Oncol.* 2005; 23:3830–3842.

91 Kuijer W, Groothoff JW, Brouwer S, et al. Prediction of sickness absence in patients with chronic low back pain: a systematic review. *J Occup Rehabil.* 2006; 16:439–467.

92 Lenssinck MLB, Damen L, Verhagen AP, et al. The effectiveness of physiotherapy and manipulation in patients with tension-type headache: a systematic review. *Pain.* 2004;112:381–388.

93 Lievense AM, Bierma-Zeinstra SMA, Verhagen AP, et al. Influence of hip dysplasia on the development of osteoarthritis of the hip. *Ann Rheum Dis.* 2004;63: 621–626.

94 Markes M, Brockow T, Resch KL. Exercise for women receiving adjuvant therapy for breast cancer. *Cochrane Database Syst Rev*. 2006;(4):CD005001.

95 Pengel HM, Maher CG, Refshauge KM. Systematic review of conservative interventions for subacute low back pain. *Clin Rehabil.* 2002;16:811–820.

96 Proper KI, Koning M, van der Beek AJ, et al. The effectiveness of worksite physical activity programs on physical activity, physical fitness, and health. *Clin J Sport Med.* 2003;13:106–117.

97 Rehn B, Lidstrom J, Skoglund J, Lindstrom B. Effects on leg muscular performance from whole-body vibration exercise: a systematic review. *Scand J Med Sci Sports.* 2007;17:2–11.

98 Rietberg MB, Brooks D, Uitdehaag BM, Kwakkel G. Exercise therapy for multiple sclerosis. *Cochrane Database Syst Rev.* 2005;(1):CD003980.

99 Seferiadis A, Rosenfeld M, Gunnarsson R. A review of treatment interventions in whiplash-associated disorders. *Eur Spine J.* 2004;13:387–397.

100 Smidt N, Assendelft WJ, Arola H, et al. Effectiveness of physiotherapy for lateral epicondylitis: a systematic review. *Ann Med.* 2003;35:51–62.

101 Smidt N, Assendelft WJ, van der Windt DA, et al. Corticosteroid injections for lateral epicondylitis: a systematic review. *Pain.* 2002;96:23–40.

102 Struijs PAA, Smidt N, Arola H, et al. Orthotic devices for tennis elbow: a systematic review. *Br J Gen Pract.* 2001;51:924–929.

103 Sung L, Nathan PC, Lange B, et al. Prophylactic granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor decrease febrile neutropenia after chemotherapy in children with cancer: a meta-analysis of randomized controlled trials. *J Clin Oncol.* 2004;22:3350–3356.

104 Thomas KC, Bailey CS, Dvorak MF, et al. Comparison of operative and nonoperative treatment for thoracolumbar burst fractures in patients without neurological deficit: a systematic review. *J Neurosurg Spine.* 2006;4:351–358.

105 van der Wurff P, Meyne W, Hagmeijer RH. Clinical tests of the sacroiliac joint: a systematic methodological review, part 2: validity. *Man Ther.* 2000;5:89–96.

106 van Der Wurff P, Hagmeijer RH, Meyne W. Clinical tests of the sacroiliac joint: a systematic methodological review, part 1: reliability. *Man Ther.* 2000;5:30–36.

107 Verhagen AP, Damen L, Berger MY, et al. Conservative treatments of children with episodic tension-type headache: a systematic review. *J Neurol.* 2005;252:1147–1154.

108 Verhagen AP, Damen L, Berger MY, et al. Is any one analgesic superior for episodic tension-type headache? *J Fam Pract.* 2006;55:1064–1072.

109 Verhagen AP, Karels C, Bierma-Zeinstra SM, et al. Exercise proves effective in a systematic review of work-related complaints of the arm, neck, or shoulder. *J Clin Epidemiol.* 2007;60:110–117.

110 Verhagen AP, Scholten-Peeters GG, van Wijngaarden S, et al. Conservative treatments for whiplash. *Cochrane Database Syst Rev.* 2007;(2):CD003338.

111 Weevers HJ, van der Beek AJ, Anema JR, et al. Work-related disease in general practice: a systematic review. *Fam Pract.* 2005;22:197–204.

112 Martou G, Veltri K, Thoma A. Surgical treatment of osteoarthritis of the carpometacarpal joint of the thumb: a systematic review. *Plast Reconstr Surg.* 2004;114:421–432.

113 Lussier JP, Heil SH, Mongeon JA, et al. A meta-analysis of voucher-based reinforcement therapy for substance use disorders. *Addiction.* 2006;101:192–203.

114 Lim I, van Wegen E, de Goede C, et al. Effects of external rhythmical cueing on gait in patients with Parkinson's disease: a systematic review. *Clin Rehabil.* 2005;19:695–713.

115 Andrew E, Eide H, Fuglerud P, et al. Publications on clinical trials with X-ray contrast media: differences in quality between journals and decades. *Eur J Radiol.* 1990;10:92–97.

116 Andrew E. Method for assessment of the reporting standard of clinical trials with roentgen contrast media. *Acta Radiol Diagn (Stockh).* 1984;25:55–58.

117 Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research, I: methods. *J Periodontal Res.* 1986;21:305–314.

118 Boutron I, Moher D, Tugwell P, et al. A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. *J Clin Epidemiol.* 2005;58:1233–1240.

119 Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA.* 1994;272:101–104.

120 Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials.* 1996;17:1–12.

121 Oremus M, Wolfson C, Perrault A, et al. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. *Dement Geriatr Cogn Disord.* 2001;12:232–236.

122 Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. *Pediatrics.* 1989;84:815–827.

123 Tyson JE, Furzan JA, Reisch JS, Mize SG. An evaluation of the quality of therapeutic studies in perinatal medicine. *J Pediatr.* 1983;102:10–13.

124 Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research, II: results, periodontal research. *J Periodontal Res.* 1986;21:315–321.

125 Chalmers TC, Smith H Jr, Blackburn B. A method for assessing the quality of a randomized control trial. *Control Clin Trials.* 1981;2:31–49.

126 de Vet HCW, de Bie RA, van der Heijden GJ, et al. Systematic reviews on the basis of methodological criteria. *Physiotherapy.* 1997;83:284–289.

127 Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *Br J Surg.* 1985;72:256–260.

128 Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol.* 1986;4:942–951.

129 Morley JA, Finney JW, Monahan SC, Floyd AS. Alcoholism treatment outcome studies, 1980–1992: methodological characteristics and quality. *Addict Behav.* 1996;21:429–443.

130 Sherrington C, Herbert RD, Maher CG, Moseley AM. PEDro: a database of randomized trials and systematic reviews in physiotherapy. *Man Ther.* 2000;5:223–226.

131 Foley NC, Bhogal SK, Teasell RW, et al. Estimates of quality and reliability with the Physiotherapy Evidence-Based Database scale to assess the methodology of randomized controlled trials of pharmacological and nonpharmacological interventions. *Phys Ther.* 2006;86:817–824.

132 Ah-See KW, Molony NC. A qualitative assessment of randomized controlled trials in otolaryngology. *J Laryngol Otol.* 1998;112:460–463.

133 Balas EA, Austin SM, Ewigman BG, et al. Methods of randomized controlled clinical trials in health services research. *Med Care.* 1995;33:687–699.

134 Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health.* 1998;52:377–384.

135 Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials. *Ann Intern Med.* 1990;113:299–307.

136 Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282:1054–1060.

137 Nguyen QV, Bezemer PD, Habets L, Prahl-Andersen B. A systematic review of the relationship between overjet size and traumatic dental injuries. *Eur J Orthod.* 1999;21:503–515.

138 Nicolucci A, Grilli R, Alexanian AA, et al. Quality, evolution, and clinical implications of randomized, controlled trials on the treatment of lung cancer: a lost opportunity for meta-analysis. *JAMA.* 1989;262:2101–2107.

139 Poynard T. Evaluation of the methodological quality of randomized therapeutic trials [in French]. *Presse Med.* 1988;17:315–318.

140 Brandt C, Sole G, Krause MW, Nel M. An evidence-based review on the validity of the Kaltenborn rule as applied to the glenohumeral joint. *Man Ther.* 2007;12:3–11.

141 Collins N, Bisset L, McPoil T, Vicenzino B. Foot orthoses in lower limb overuse conditions: a systematic review and meta-analysis. *Foot Ankle Int.* 2007;28:396–412.

142 van Tulder M, Furlan A, Bombardier C, Bouter L; Editorial Board of the Cochrane Collaboration Back Review Group. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine.* 2003;28:1290–1299.

143 van Tulder MW, Assendelft WJ, Koes BW, Bouter LM. Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for Spinal Disorders. *Spine.* 1997;22:2323–2330.

144 McNeely ML, Armijo Olivo S, Magee DJ. A systematic review of the effectiveness of physical therapy interventions for temporomandibular disorders. *Phys Ther.* 2006;86:710–720.

145 Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6* [updated September 2006]. In: The Cochrane Library, Issue 4, 2006. Chichester, United Kingdom: John Wiley & Sons Ltd; 2006.

146 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet.* 2001;357:1191–1194.

147 Spitzer WO, Lawrence V, Dales R, et al. Links between passive smoking and disease: a best-evidence synthesis. A report of The Working Group on Passive Smoking. *Clin Invest Med*. 1990;13:17–42.

148 Bhogal SK, Teasell RW, Foley NC, Speechley MR. The PEDro scale provides a more comprehensive measure of methodological quality than the Jadad scale in stroke rehabilitation literature. *J Clin Epidemiol.* 2005;58:668–673.

149 Kroeling P, Gross AR, Goldsmith CH; Cervical Overview Group. A Cochrane review of electrotherapy for mechanical neck disorders. *Spine.* 2005;30:E641–E648.

150 Gross AR, Hoving JL, Haines TA, et al; Cervical Overview Group. A Cochrane review of manipulation and mobilization for mechanical neck disorders. *Spine.* 2004;29:1541–1548.

151 Verhagen AP, de Bie RA, Lenssen AF, et al. Quality assessment of trials: a comparison of three criteria lists. *Physical Therapy Reviews.* 2000;5:49–58.

152 Andrew E, Anis A, Chalmers T, et al. A proposal for structured reporting of randomized controlled trials. *JAMA*. 1994;272:1926–1931.

153 Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. *J Clin Epidemiol.* 2006;59:1249–1256.

154 Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technol Assess Health Care.* 1996;12:195–208.

155 Gadotti I, Vieira E, Magee D. Importance and clarification of measurements properties in rehabilitation. *Rev Bras Fisioter.* 2006;10(2):137–146.

156 Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358–1361.

157 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* 1998;352:609–613.

158 Schulz KF. Assessing allocation concealment and blinding in randomised controlled trials: why bother? *Evid Based Nurs.* 2001;4:4–6.

159 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995;273:408–412.

160 Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy, II: surgical. *Stat Med.* 1989;8:455–466.

161 Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med.* 1989;8:441–454.

162 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. *N Engl J Med.* 1978;299:690–694.

163 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA.* 1994;272:122–124.

164 Ruiz-Canela M, Martinez-Gonzalez MA, de Irala-Estevez J. Intention to treat analysis is related to methodological quality. *BMJ.* 2000;320:1007–1008.

165 Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA.* 2002;287:2973–2982.

166 Verhagen AP, de Bie RA, Lenssen AF, et al. Impact of quality items on study outcome: treatments in acute lateral ankle sprains. *Int J Technol Assess Health Care.* 2000;16:1136–1146.

167 Nestoriuc Y, Martin A. Efficacy of biofeedback for migraine: a meta-analysis. *Pain.* 2007;128:111–127.

168 Ezzo J, Haraldsson BG, Gross AR, et al; Cervical Overview Group. Massage for mechanical neck disorders: a systematic review. *Spine.* 2007;32:353–362.

169 Trinh K, Graham N, Gross A, et al. Acupuncture for neck disorders. *Spine.* 2007;32:236–243.

170 Graham N, Gross AR, Goldsmith C. Cervical Overview Group. Mechanical traction for mechanical neck disorders: a systematic review. *J Rehabil Med.* 2006;38:145–152.

171 Kay TM, Gross A, Goldsmith C, et al; Cervical Overview Group. Exercises for mechanical neck disorders. *Cochrane Database Syst Rev.* 2005;(3):CD004250.