THE JOURNAL OF ALTERNATIVE AND COMPLEMENTARY MEDICINE
Volume 17, Number 5, 2011, pp. 389–395

DOI: 10.1089/acm.2010.0175

Opposing Systematic Reviews: The Effects of Two Quality Rating Instruments on Evidence Regarding *T'ai Chi* and Bone Mineral Density in Postmenopausal Women

Sunny Y. Alperson, PhD, CRNP,[1] and Vance W. Berger, PhD[2]

Abstract

Purpose: This article compares and contrasts two systematic reviews of *t'ai chi* (TC) interventions on bone mineral density in postmenopausal women. The aim is to examine how chosen quality rating instruments can impact systematic reviews of TC literature.
Methods: The rating instruments in the reviews, the three-item scale of Jadad et al. and the *ad hoc* checklist of Wayne et al., were analyzed using Oxman's evaluation criteria for systematic reviews regarding inclusion of articles, interpretation of results, and overall implications for the efficacy of TC on bone mineral density.
Results: According to Oxman's criteria, the Jadad scale did not address advances in statistical methods and was not comprehensive enough to adapt to the clinical context or topic. In contrast, the checklist by Wayne et al. was comprehensive, adaptable to clinical context and topical relevance, and compatible with recent developments in statistics and experimental design. These quality rating instruments were critical in the inclusion of studies, analyses, and overall conclusions summarizing the TC literature. The conclusions from the two systematic reviews were starkly opposing; Lee et al. found no convincing evidence, dismissing TC studies as low quality, while Wayne et al. stated that TC may be an effective, safe, and practical intervention.
Conclusions: Readers must exercise caution concerning high or low ratings from systematic reviews of TC studies because the choice of quality rating tool can dramatically influence the summary and conclusions of the reviews. There is no consensus on quality rating standards at this time. Of the two, the Jadad scale was not only inadequate but also inappropriate for reviewing TC studies, potentially misleading researchers, clinicians and policymakers. Future systematic reviews of TC should utilize instruments that are updated to current scientific standards, comprehensive, adaptable to clinical context, and relevant to the research topic.

Introduction

Systematic reviews have become a critical component of current evidence-based practice. The purpose of systematic reviews is to provide clinicians, educators, and other health care decision makers with updated information upon which to base patient care and to make policy decisions.[1–5] Contrasting with traditional expert narrative reviews, systematic reviews synthesize evidence using a structured reproducible approach based on relevant studies that are sufficiently high in quality determined by quality rating instruments. Despite their widespread use, systematic reviews remain controversial, with issues involving poor construct validity of quality ratings from some of these instruments. Several analyses have demonstrated inconsistent quality ratings of individual studies, as well as high-quality ratings for studies with serious deficiencies in design.[6–8]

The specific aim of this article is to critically evaluate the quality rating instruments utilized in two starkly opposing systematic reviews on the efficacy of *t'ai chi* (TC) interventions for bone mineral density in postmenopausal women: the 2008 review by Lee et al. and the 2007 review by Wayne et al.[9,10] It will examine the impact of these scales on the inclusion of articles for reviews, the interpretation of results, and the overall conclusions regarding the efficacy of TC for bone mineral density in postmenopausal women. First, the quality rating instruments employed in the two reviews were compared. Then, Oxman's systematic review criteria were applied to evaluate these two instruments as used in the systematic reviews of TC for bone mineral density.[3] Using three different well-known quality rating scales, one study selected by Wayne et al. and Lee at al. were re-analyzed to elucidate their divergent ratings. Finally, suggestions were offered for future systematic reviews on the efficacy of TC interventions.

[1]National Institute of Nursing Research, and NIH Clinical Center, National Institutes of Health, Bethesda, MD.
[2]National Cancer Institute, National Institutes of Health, Bethesda, MD.

Quality Rating Instruments Used in Two Systematic Reviews of TC

Both reviews searched databases of articles in English and Chinese, with Lee et al. adding Korean.[9,10] Lee and colleagues then selected a scale by Jadad et al. as a quality-rating instrument. This scale was originally developed to evaluate studies on the efficacy of pharmacological interventions for pain relief.[11] With only three items, this brief scale gives little reviewer burden, and it remains the most popular scale employed in systematic reviews across all interventions.[12,13]

As seen in Table 1, the Jadad scale measured study quality by answering yes or no to three major components of methodological quality: randomization, double blinding, and description of participant dropouts, with possible total scores of 0–5. Randomization was scored as a 2 when unbiased randomization was done with a proper method such as with a computer algorithm; 1 when randomization was potentially biased, as with an alternation method; and 0 if no randomization. Two (2) was awarded for double blinding; 1 if outcome assessor blinding only; and 0 if no blinding. Finally, a score of 1 was given for a description of dropouts and 0 if none. Thus, the Jadad scoring system provided Lee et al. with a simple scoring system for evaluating study quality for their systematic review, primarily focusing on blinding and randomization (4 of 5, or 80%), with the remaining 1 point for participant dropouts and withdrawal.

The Wayne et al. scale consisted of 9 items.[9,10] Items 1, 2, 4, and 5 were comparable to those in Jadad, focusing on blinding, randomization, and participant dropouts and withdrawals. However, Wayne et al. only checked for outcome assessor blinding, with no provision for double blinding. They included items for inclusion and exclusion criteria, statistical power, and appropriate statistical inferential analyses. Last, they added items specific to TC context: whether there were clear descriptions of the TC intervention and the qualifications of the TC instructors.

Table 1. Scoring Criteria from the Jadad Scale and the Wayne Checklist

Three-item Jadad scale used by Lee et al.[9]	*Nine-item checklist used by Wayne* et al.[10]
1. Randomized? *2 if a proper method;* *1 if no description or an improper method;* *0 otherwise.*	1. Randomized? 2. Proper randomization methods? 3. Clear inclusion and exclusion criteria?
2. Blinding? *2 if double blinding;* *1 if outcome assessor blinding only; 0 if no blinding.*	4. Outcome assessors blinded? 5. Description of patient withdrawals and dropouts?
3. Description of patient withdrawals and dropouts? *1 if a description is given;* *0 otherwise.*	6. Sample size justified and estimated for power? 7. Appropriate data analysis? 8. *T'ai chi* intervention described? 9. Qualifications of *t'ai chi* instructors? *For all items 1 if yes;* *0 otherwise.*
Possible scores: 0–5	Possible scores: 0–9

Methods and Results

Oxman's criteria—a listing of standards for systematic reviews—has been used as a steering guide by the Cochrane Collaborative for over 4000 systematic reviews on health care.[14,15] The criteria consist of 11 items related to the quality of the systematic review in problem formulation, data collection, data synthesis, and interpretation of results.[3] Table 2 presents these criteria, along with our evaluations of the standards as applied to the systematic reviews of Lee et al. (2008) and Wayne et al. (2007) on the efficacy of TC for bone mineral density.

Because six of the criteria from Table 2 differ between the reviews, they are addressed in more detail below (Criteria 3, 4, 6, 8, 9, and 10). The remaining five are comparable and are not discussed further.

Criterion 3: Are the inclusion criteria appropriate?

The selection of articles for inclusion from the computer database searches is determined by the scoring criteria. These scoring systems have critical implications not only for the selection of studies but also for the conclusions. In their discussion of systematic reviews in complementary medicine, Linde and Willich posit that discrepancies in conclusions can stem from subtle differences in inclusion criteria of the reviews.[16]

Using the Jadad scoring system, Lee et al. selected three randomized controlled trials (RCTs) and one controlled clinical trial (CCT) for postmenopausal women, while for "elderly," they selected two RCTs and one CCT. Wayne et al. included six controlled studies for postmenopausal women, with two RCTs, two cross-sectional studies, and two prospective parallel cohort studies. Three (3) postmenopausal studies were selected by both reviews.[17–19]

Under the Jadad rating, TC studies can never achieve a perfect score of 5. The 5th point cannot be awarded for double blinding because of the visibility of TC movements; at best, TC studies can create a partial single blind for the outcome assessor. In Lee et al., studies that were deemed "higher quality" were necessarily RCTs or CCTs. Consequently, no cross-sectional or prospective cohort studies were included. Though the omission of cross-sectional and prospective cohort studies by Lee et al. was consistent with the Jadad criteria, it effectively limited Lee et al.'s final pool of studies to only short-term TC studies. In their suggestions for future research, Lee et al. stated that current TC research needs to examine long-term benefits of TC to better understand its potential, and that prolonging TC interventions for longer than a year "might give a different picture of the effects of Tai Chi." These longer term cross-sectional and prospective cohort studies may have value, particularly for hypothesis generation and future research directions.[20] However, the Jadad criteria did not allow them to be included for review.

In contrast, Wayne and colleagues included two cross-sectional studies and one prospective parallel cohort study.[21–23] These three studies by Wayne et al. included subjects with long-term experience with TC, comparing them to control subjects matched for age and sex. The Wayne et al. checklist also examined more details in methodological qualities of the primary studies than Jadad. Their checklist verified clear inclusion and exclusion criteria, sample size calculations,

Table 2. Oxman et al. Criteria and *T'ai Chi* (TC) Bone Mineral Density Systematic Reviews

Criterion	*Lee* et al. *(2008)*[9]	*Wayne* et al. *(2007)*[10]
1. Is the primary question clearly focused?	Yes	Yes
2. Is the search for relevant studies thorough?	Yes; comprehensive without language barriers (English, Chinese, and Korean databases included along with unpublished abstracts)	Yes; comprehensive without language barriers (English, Chinese database included; no unpublished studies)
3. Are the inclusion criteria appropriate?	No; omit relevant studies. Double-blind criterion. Inappropriate for TC context.	Yes; included relevant cross-sectional and cohort studies. Criteria appropriate to TC context
4. Is the validity of included studies adequately assessed?	Focuses on internal validity.	Considers internal and external validity.
5. Data collection: Is missing information obtained from investigators?	No	No
6. How sensitive are the results to changes in the way the review is done?	"Sensitivity analyses" needed but not conducted.	Discusses how different bone mineral density outcome measures affect conclusions.
7. Do the conclusions flow from the evidence that is reviewed?	Yes; overall conclusions are consistent with those studies satisfying the chosen scale.	Yes; overall conclusions are consistent with those studies satisfying the chosen scale.
8. Are recommendations linked to the strength of the evidence?	No. Recommendations emphasize weaknesses of the studies. Few strengths presented.	Recommendations tied to the selected studies and their strengths.
9. Are judgments about preferences (values) explicit?	Preferred a quality-rating scale that met the "accepted standards of trial methodology."	Preferred a quality-rating scale that considered the nature of the intervention and the relevance to the clinical meaningfulness.
10. If there is "no evidence of effect," is caution taken not to interpret this as "evidence of no effect"?	No. Stated TC does not affect bone mineral density, and offered reason: TC was not a weight-bearing exercise.	Yes. Caution is taken to say "no strong evidence."
11. Are subgroup analyses interpreted cautiously?	Subgroup analyses not conducted.	Subgroup analyses not conducted.

and appropriate descriptive and inferential statistics. They included two criteria specific to TC: how TC was implemented in the intervention, and the qualifications of the TC instructor. The frequency, intensity, and duration of TC intervention and the way it was implemented can make a difference in the outcome and should be reported in studies. For instance, one can expect a lower "dosage" of TC consisting of 30 minutes per week to have less impact than a study with 5 intensive hours of practice per week. Furthermore, highly knowledgeable and skilled instructors may provide a stronger role model and may increase compliance and effects due to more skillful home practice.[24] Inclusion of these TC-specific criteria can enable more consistent replication of the intervention and can improve the quality of future systematic reviews.

Criterion 4: Is the validity of included studies adequately assessed?

Both teams of reviewers evaluated their databases of studies using the standards proposed by their quality rating tools. This prescription of method is no guarantee of internal validity, even for double-blind RCT designs.[25] The nature of the control condition (active, passive, waiting list) is essential for internal validity considerations and the interpretation of the active components in the intervention; it needs to be addressed. Neither review supplied this information. In addition, inter-rater reliability is missing from both reviews, limiting statistical measurements of validity.[26] Both reviews mentioned that inconsistencies leading to low inter-rater reliability were resolved by discussion. Although this forced agreement is expedient, it does not allow the independent verification that is central to inter-rater reliability. Other authors have stated that the Jadad scale has low reliability and face validity.[27] For example, a study could have received a perfect rating of 5 in Jadad yet be fraught with errors that distorted data and conclusions.[6]

In these two reviews of TC studies on bone mineral density, there were starkly conflicting evaluations of one article used in both systematic reviews, a study by K. Chan et al. titled "A randomized, prospective study of the effects of Tai Chi Chun exercise on bone mineral density in postmenopausal women".[17] In this RCT, the study authors compared bone mineral density levels following assignment to a TC exercise group versus assignment to a sedentary control group. Using the Jadad scale, Lee and colleagues awarded 2 of 5 total points to the study.[9] Although the details of their rating are not given, one may surmise that it received 1 point for its randomized design and 1 point for description of patient withdrawals and dropouts. Lee et al. referred to the study very briefly, dismissing it as low quality. In contrast, Wayne et al. rated the Chan et al. study 7 of 9 quality checks, the highest rating of all studies reviewed.[10] It lost a total of 2

checks: one for the absence of outcome assessor blinding and the other point for not mentioning the qualifications of the TC instructor. Wayne and colleagues[10] described the Chan et al. study results in greater detail and referred to it repeatedly when they formed their overall conclusions recommending TC for widespread dissemination.[10,p 675–678]

To better understand these diverging ratings of the Chan et al. study, the present authors conducted our own investigation regarding its construct validity using three other well-regarded quality rating instruments: scales by W. Chan and Bartlett, Cho and Bero, and Downs and Black.[28–30] The choice of these quality-rating instruments was purposely broad, in order to sample different approaches to systematic review. The Chan and Bartlett approach was similar to Wayne et al. in that it was an *ad hoc* method developed for reviewing TC studies. The Cho and Bero approach was based on guidelines for systematic reviews of drug studies, and thus, ostensibly had the same goal as Jadad et al. Finally, the Downs and Black approach represented methodological and quality considerations for systematic reviews as applied to epidemiological and public health concerns; accordingly, their scale included standards for both randomized and nonrandomized studies. The raters were two PhDs (a postdoctoral fellow of integrative medicine and a professor of behavioral statistics and psychology) and an MD (a postdoctoral clinical research student in medicine).

The results, as summarized in Table 3, show that the checklist of Wayne et al. received a 78% rating, in good agreement with the three other scales (75%, 77%, and 86%). The Jadad scale shows a 40% rating. Among other aspects, the number of criteria included in the scale may be one of the factors explaining this isolated divergence; the Jadad scale had only 3 criteria to evaluate, while the remaining rating scales were more comprehensive, with 9–27 criteria.

Reviewing the original study by K. Chan et al. in more depth, it was found to address many issues related to internal and external validity. Power calculations were performed initially to gauge sample size appropriately. Reliability measurements and validity studies were supplied on outcome measurements, helping establish their precision and suitability for inclusion. The TC intervention was described in sufficient detail for replication. Confounding factors that might complicate the interpretation were briefly described, along with measurements of relevant anthropometric, hormonal, and dependent variables. Detailed specifications of baseline and follow-up values were presented by means, standard deviation, and percentage differences across six dependent variables for the TC and control groups. Appropriate inferential statistics were used and the results of the tests were properly reported. Annual changes in bone mineral density at different anatomic sites were presented, along with adjustments for expected rates of loss due to aging. Related dependent variables such as fall rates and fractures in the TC and control groups were documented and discussed. Weaknesses of the study included lack of outcome assessor blinding and external validity of both the particular TC intervention and the sample, which limited generalizability to other TC interventions and populations.

Table 3. Comparison of Five Quality Rating Scales: Mean Ratings of Chan et al. Article on Effects of *T'ai Chi* on Bone Mineral Density[17]

Rating scale	*Number of criteria*	*Mean quality rating*[a]	*Percentage*
Jadad et al.[10]	3	2 of 5	40%
Wayne et al.[9]	9	7 of 9	78%
Chan and Bartlett[28]	18	42 of 49	86%
Cho and Bero[29]	24	48 of 62	77%
Downs and Black[30]	27	24 of 32	75%

[a]Quality ratings were averaged from 3 independent raters.

Criterion 6: How sensitive are the results to change in the way the review is done?

Oxman states that systematic reviews need to check against "sensitivity analysis" using various means. One basic way is to examine how the results change when inclusion criteria were modified. The discordance between these two reviews cited above demonstrates how different inclusion criteria in the quality-rating tools dramatically affect the ratings of individual studies and the formation of the pool of quality studies.

The inclusion criteria also ultimately affect the conclusions regarding the effectiveness of TC. Lee et al. stated that the "results for post-menopausal women failed to show specific effects of Tai Chi for bone mineral density. . . .Overall our findings provide no convincing evidence that Tai Chi is beneficial for preventing or treating osteoporosis."[9, p. 141] On the other hand, Wayne et al. offered strikingly different conclusions: "Tai Chi may be an effective, safe and practical intervention for maintaining bone mineral density in postmenopausal women."[10, p. 673] Citing their selected studies, other systematic reviews on balance and fractures, and economics of the intervention, they conclude "Tai Chi may be a logical and practical response to the Surgeon General's recent call for novel exercise programs for women with low bone density."[10, p. 677–678.]

Wayne et al. emphasize another factor that can dramatically change results: the way that bone mineral density is operationalized. There are several procedures to gauge bone mineral density, including dual-energy x-ray absorptiometry (DXA), quantitative computerized tomography (QCT), and broadband ultrasound attenuation; and there are multiple body sites for assessing bone mineral density. Though the Chan et al. study did not find significant change using DXA at a spinal site, they did find significant improvements at other sites using different measures. Wayne et al. argued that "QCT has the advantage of being able to quantify true volumetric density as well as partition the 2 types of bone, trabecular and cortical, which may respond differently to exercise. Moreover, it has the potential to have higher precision."[10, p. 677]

Criterion 8: Are recommendations linked to the strength of the evidence?

Lee et al. stated that "the evidence is not convincing for Tai Chi in preventing or treating osteoporosis."[9] Reporting that TC studies were methodologically weak, their recommendations focused on ways to improve the design of future studies, such as RCT designs, larger patient samples, longer

assessment durations, appropriate inferential tests, and more complete write-ups. They also recommended additional outcome measures, such as balance and fractures related to falls.

While agreeing with the Lee et al. contention that many TC bone mineral density studies were poor methodologically, Wayne et al. focused on the strength of six studies that formed their evidentiary pool. They also utilized conclusions from other systematic reviews and individual studies to form conclusions and recommendations, going beyond the evidence base from their selected studies to reflect a more encompassing segment of the current TC literature.

Criterion 9: Are judgments about preferences (values) explicit?

The issue that Oxman raises in this criterion is more complex than whether validity was adequately assessed according to quality-scale criteria. Tool choice may reflect the authors' attitudes toward particular methodologies. Furthermore, following a given set of evaluative steps does not guarantee adequate ratings of study quality. It is problematic in the translation of the proper evidence, precisely because the chosen quality-rating instruments affect the analysis, interpretation, and conclusions about individual studies, as well as the overall summary of evidence. This question can be rephrased as, "Are the reviewers aware of their own value judgments or preference of the topic or their own agenda?" In their comparison of systematic reviews on complementary medicine, Linde and Willich note that unless the outcome measure is very clear with obvious differences, authors of different reviews often interject their own hypotheses, which reflect their philosophies.[16]

Lee et al. focused on older traditional standards of experimental design embodied in the Jadad scale, and they did not extend any conclusions beyond the results of their screened sample. While they did not explicitly state why they found the Jadad criteria appropriate for TC systematic reviews, they encouraged future TC researchers to "utilize accepted standards of trial methodology."[9] Though Lee et al. did not explicitly explain what "accepted standards" were, one can assume that Lee et al. were referring to the Jadad scale as the "gold standard" for research studies. In contrast, Wayne et al. explicitly rejected the Jadad scale as a standard for evaluation criteria "because RCTs employing TC interventions are not amenable to double-blinding."[10] Instead, they offered their *ad hoc* standards for assessing methodology that reflected the context of TC. These included the checks on description of the TC intervention and the qualifications of the TC instructor. Although their qualification of TC instructors needs to be operationalized and their checklist needs validation on other TC topics, their focus is clearly more inclusive than the Lee et al. "accepted standards."

Criterion 10: Is "no evidence of effect" interpreted as "evidence of no effect"?

Both systematic reviews generally avoided this logical error, which confuses insufficient evidence of change with no change. However, when Lee et al. discussed their findings, they stated that TC had evidence of no effect on bone mineral density. They state: "One should also note that Tai Chi is not the type of exercise that provides for much loading on weight bearing joints, which is a precondition for an effect on bone metabolism."[9] This statement is unwarranted on at least two grounds. This statement presumes that TC has no effect, rather than the evidence that the effect is weak, which can be related to other factors such as poorly validated measures and inadequate sample size. Furthermore, kinesthetic studies indicate that TC imposes substantial loads on weight-bearing joints.[23,31] A more accurate statement of the Lee et al. review is a restricted claim in line with Oxman's caution, that there is no evidence of effect according to the studies cited in their article.

Essential Aspects of Ratings Instruments for TC

Attention is now turned to a broader consideration of the elements in the quality-rating tool, that is, what should be included in quality rating instruments for TC studies? Optimal choice of a rating instrument must include updated criteria, comprehensiveness, and some relevance to context of the TC topic for its clinical meaningfulness.

Updated criteria

Methodologies for systematic review across health care literature have evolved since 1996 when Jadad et al. first published their scale. The three items in Jadad, while contributing to a low reviewer burden, do not address whether the studies utilized updated standards for high-quality methodology.[6,32] For example, power considerations have gradually been adopted as an essential element of study methodology, following the pioneering work of Cohen.[33] When randomized and blinded groups are formed with samples that are too small, within-group variance can easily overweigh between-group variance, obscuring small real effects. The omission of power calculations can lead to type 2 errors, where the researcher misses significant findings. In this context, RCT studies with inadequate samples might be selected for inclusion, indicating nonsignificance for TC and bone mineral density, when it is entirely possible that adequately sized samples with the exact same intervention would be significant. Thus, there could easily be a selection bias toward reporting summary conclusions of no effect in the overall synthesis of evidence in the literature. The Jadad scale did not stipulate power calculations, in contrast with Wayne et al.

Recent improvements related to experimental design need to be incorporated in the criteria. The method of implementing randomization in RCTs demands careful attention, in order to better match control and experimental groups and to minimize bias.[13,34] Testing after randomization should include covariate adjustments as necessary to ensure baseline comparability. In addition to demographic and anthropometric similarities, comparable prior exposures or nonexposures to TC at baseline should be demonstrated, as in the Chan et al. study. Neither Wayne nor Jadad scales set any standards for baseline comparability. To minimize bias, assignment to group should be concealed using a procedure such as permutated blocks of varying sizes. This allocation method reduces chances of predicting future allocations.[32]

Comprehensiveness

A major shift toward more comprehensive quality-rating scales has occurred during the past 10 years. A leading organization of evidence practice in the United States, the Agency for Healthcare Research and Quality, examined research quality and measuring systems. West et al. developed a grid containing 10 basic required domains that should be evaluated for the quality of any randomized controlled trials.[13] The Jadad scale covered only 2.5 of the 10 domains for RCT quality evaluation. In comparison, other evaluation scales such as the Downs and Black checklist addressed 8 domains,[30] while Cho and Bero addressed 7.[29] The Wayne et al. checklist covered 8 domains.

Relevance of the topic and clinical context

As different facets of any health care research study, methodological rigor and clinical relevance are both essential components to be considered.[13] Quality-rating instruments must rate studies to find methodologically strong studies. However, the studies also must relate to the fundamental characteristics of the topic and the concerns of clinicians, lest the conclusions become irrelevant.[12,35–37]

In addition to its shortcomings as a quality-rating tool for study methodology, the Jadad scale omits the clinical context. A TC intervention study on bone mineral density in postmenopausal women is a completely different context than Jadad originally intended: the evaluation of efficacy of drugs for pain relief. The Jadad scale itself gives no guidance for adapting to complementary and alternative medicine movement interventions. In their implementation of Jadad, Lee et al. did not consider the nature of the clinical intervention or topic relevance, and as a consequence, they did not properly summarize the research literature for clinicians interested in TC.

Some rating tools reflect both rigor and clinical topic relevance. Cho and Bero argued that clinical relevance and meaningfulness must be considered as essential elements to be assessed in the evaluation of primary studies, and their quality rating instrument includes a separate section of "clinical relevance." Downs and Black's comprehensive Quality Index (QI) includes 27 items, with many addressing clinical relevance. The Chan and Bartlett scale has 18 items that specifically reflect TC content.[28] All three of these instruments are suited to both randomized and nonrandomized designs. Furthermore, one systematic review of TC interventions combines two instruments: Down and Black's QI and Chan and Bartlett's criteria, in order to increase rigor and relevance.[38]

Conclusions

Readers must exercise caution concerning high or low ratings from existing systematic reviews of TC studies because there are no universally accepted quality standards. Inconsistencies and conflicting interpretations of evidence in the systematic reviews can mislead clinicians and health care policymakers.[7,13,39] The three-item Jadad scale is unable to adequately address technical aspects of general trial quality in current research methodologies. Its criteria are too limited; it fails to reflect criteria for current methodological design standards; and it lacks the adaptability for the topical and clinical relevance to produce meaningful reviews. Although the Jadad scale is known for its popularity for the least respondent burden and for its historical value, it is not appropriate or adequate for evaluating TC research in the current research environment. In contrast, the scale adopted by Wayne et al. is both updated in methodology that evaluates primary studies and comprehensive, particularly for TC studies. However, it needs additional criteria for reliability and validity of the measures. We strongly urge the adoption of quality assessment tools much more comprehensive than Jadad for evaluating TC studies. These tools should meet criteria for psychometric validation and be clinically relevant.

Lastly, a balanced perspective is needed to review overall quality and evidence of studies.[40] Rather than viewing evidence as an absolute produced by a measurable quality-scoring instrument, we agree with Guyatt et al. and Hopayian regarding the importance of clinical context.[36,41] Reviews must be rigorous, but the "rigor" of the evidence must have meaningfulness for application to clinical health contexts, so that imbalanced or improper reviews of the current literature do not impede further clinical applications and policies, research funding, and knowledge development of the phenomenon under study. In the end, the central theme always boils down to such basic and honest questions: How can we better help our patients and where do we go from here?

Acknowledgments

We would like to express our thanks to Karen Lewison, MD and Jay Alperson, PhD for their participation in rating the Chan et al. article with different quality rating scales. This article was supported in part by the Bravewell Collaborative, the National Institutes of Health, Intramural division of Research Program at the National Institute of Nursing Research and the Clinical Center, NIH. Its contents are solely the responsibility of the authors and do not represent views of Bravewell nor the National Institutes of Health.

Disclosure Statement

No competing financial interests exist.

References

1. Brouwers MC, Johnston ME, Charette ML, et al. Evaluating the role of quality assessment of primary studies in systematic reviews of cancer practice guidelines. BMC Medical Res Methodol 2005;5:8.
2. Evans D, Kowanko I, Hodgkinson B. Systematic reviews in nursing research. Austr Nurs J 1998;5:42.
3. Oxman AD, Cook DJ, Guyatt GH. Users guides to the medical literature: 6. How to use an overview. *JAMA* 1994;272:1367–1371.
4. Pai M, McCulloch M, Gorman JD, et al. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. Natl Med J India 2004;17:86–95.
5. White P. Evidence-based medicine for consumers: A role for the Cochrane Collaboration. J Med Library Assoc 2002;90: 218–222.
6. Berger VW. Is the Jadad score the proper evaluation of trials? J Rheumatol 2006;33:1710–1711.
7. 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.

8. 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.
9. Lee MS, Pittler MH, Shin BC, Ernst E. Tai chi for osteoporosis: A systematic review. Osteoporosis Int 2008;19:139–146.
10. Wayne PM, Kiel DP, Krebs DE, et al. The effects of Tai Chi on bone mineral density in postmenopausal women: A systematic review. Arch Phys Med Rehabil 2007;88:673–680.
11. 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.
12. Moja LP, Telaro E, D'Amico R, et al. Assessment of methodological quality of primary studies by systematic reviews: Results of the metaquality study cross sectional study. BMJ 2005;330:1053–1055.
13. West S, King V, Carey TS. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. Rockville, MD: Agency for Healthcare Research and Quality, 2002.
14. Cochrane Collaboration. Working Together to Provide the Best Evidence for Health Care. 2010. Online document at: www.cochrane.org/ Accessed June 8, 2010.
15. Cochrane Collaboration. Cochrane Reviews 2010. Online document at: www.cochrane.org/cochrane-reviews Accessed June 8, 2010.
16. Linde K, Willich SN. How objective are systematic reviews? Differences between reviews on complementary medicine. J R Soc Med 2003;96:17–22.
17. Chan KM, Qin L, Lau MC, et al. A randomized, prospective study of the effects of Tai Chi Chun exercise on bone mineral density in postmenopausal women. Arch Phys Med Rehabil 2004;85:717–722.
18. Xu DH, Lawson D, Kras A. A study on tai ji exercise and traditional Chinese medical modalities in relation to bone structure, bone function and menopausal symptoms. J Chin Med 2004;74:10–14.
19. Zhao JX, Zhang L, Tian Y. Effect of 6 months of Tai Chi Chuan and calcium supplementation on bone health in females aged 50–59 years. J Exercise Sci Fitness 2007;5:88–94.
20. Saunders LD, Soomro GM, Buckingham J, et al. Assessing the methodological quality of nonrandomized intervention studies. West J Nurs Res Mar 2003;25:223–237.
21. Gong M, Zhang SZ, Wang B, Wang DH. Effects of longterm shadowboxing exercise on bone mineral density in the aged [in Chinese]. Chin J Clin Rehabil 2003;7:2238–2239.
22. Qin L, Au S, Choy W, et al. Regular Tai Chi Chuan exercise may retard bone loss in postmenopausal women: A case–control study. Archives of Physical Med Rehabil 2002;83: 1355–1359.
23. Qin L, Choy W, Leung K, et al. Beneficial effects of regular Tai Chi exercise on musculoskeletal system. J Bone Miner Metab 2005;23:186–190.
24. Stead M, Wimbush E, Eadie D, Teer P. A qualitative study of older people's perceptions of ageing and exercise: The implications for health promotion. Health Educ J 1997;56:3–16.
25. Berger VW. Selection Bias and Covariate Imbalances in Randomized Clinical Trials. Hoboken, NJ: Wiley, 2005.
26. Pearson MJT, Lindop FA, Mockett SP, Saunders L. Validity and inter-rater reliability of the Lindop Parkinson's Disease Mobility Assessment: A preliminary study. Physiotherapy 2009;95:126–133.
27. 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.
28. Chan WW, Bartlett DJ. Effectiveness of Tai Chi as a therapeutic exercise in improving balance and postural control. Phys Occup Ther Geriatr 2000;17:1–22.
29. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;272:101–104.
30. 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 Commun Health 1998;52:377–384.
31. Lai YM, Qin L, Hung VWY, et al. Trabecular bone status in ultradistal tibia under habitual gait loading: A pQCT study in postmenopausal women. J Clin Densitometry 2006;9:175–183.
32. Berger VW, Alperson SY. A general framework for the evaluation of clinical trial quality. Rev Recent Clin Trials 2009;4:79–88.
33. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 1988.
34. Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. J Am Podiatric Med Assoc 2001;91:437–442.
35. Gartlehner G, Hansen R, Nissman D, et al. Criteria for Distinguishing Effectiveness from Efficacy Trials in Systematic Reviews. Technical Review 12: Agency for Healthcare Research and Quality. 2006. Online document at: www.ahrq.gov/downloads/pub/evidence/pdf/efftrials/efftrials.pdf Accessed February 12, 2010.
36. Hopayian K. The need for caution in interpreting high quality systematic reviews. BMJ 2001;323:681–684.
37. Sackett DL, Rosenberg WMC, Gray JAM, et al. Evidence based medicine: What it is and what it isn't. It's about integrating individual clinical expertise and the best external evidence. BMJ 1996;312:71–72.
38. Komagato S, Newton R. The effectiveness of *Tai Chi* on improving balance in older adults; an evidence-based review. J Phys Ther Sci 2003;26:9–16.
39. Khan KS, Riet G, Popay J, et al. Study quality assessment (phase 5): Conducting the review (stage 2). Undertaking systematic reviews of research on effectiveness. CRD Rep 2001;4:1–20.
40. Alperson SY. Tai chi philosophy and nursing epistemology. Adv Nurs Sci 2008;31:E1–E15.
41. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature, how to use an article about therapy or prevention: B. What were the results and will they help me in caring for my patients? JAMA 1994;271:59–63.

Address correspondence to:
Sunny Y. Alperson, PhD, CRNP
National Institute of Nursing Research
and NIH Clinical Center
National Institutes of Health
10 Center Drive, Room 2B02A MSC 11151
Bethesda, MD 20892

E-mail: alpersonsy@mail.nih.gov