Journal of Clinical Epidemiology 86 (2017) 176–181

The PEDro scale had acceptably high convergent validity, construct validity, and interrater reliability in evaluating methodological quality of pharmaceutical trials

Tie Parma Yamato[a,*], Chris Maher[a], Bart Koes[b], Anne Moseley[a]

[a]School of Public Health, Sydney Medical School, The University of Sydney, Edward Ford Building (A27), Sydney, NSW 2006, Australia
[b]Department of General Practice, Erasmus MC, Rotterdam 3000 CA, The Netherlands

Accepted 6 March 2017; Published online 11 March 2017

Abstract

Background and Objective: The Physiotherapy Evidence Database (PEDro) scale has been widely used to investigate methodological quality in physiotherapy randomized controlled trials; however, its validity has not been tested for pharmaceutical trials. The aim of this study was to investigate the validity and interrater reliability of the PEDro scale for pharmaceutical trials. The reliability was also examined for the Cochrane Back and Neck (CBN) Group risk of bias tool.

Methods: This is a secondary analysis of data from a previous study. We considered randomized placebo controlled trials evaluating any pain medication for chronic spinal pain or osteoarthritis. Convergent validity was evaluated by correlating the PEDro score with the summary score of the CBN risk of bias tool. The construct validity was tested using a linear regression analysis to determine the degree to which the total PEDro score is associated with treatment effect sizes, journal impact factor, and the summary score for the CBN risk of bias tool. The interrater reliability was estimated using the Prevalence and Bias Adjusted Kappa coefficient and 95% confidence interval (CI) for the PEDro scale and CBN risk of bias tool.

Results: Fifty-three trials were included, with 91 treatment effect sizes included in the analyses. The correlation between PEDro scale and CBN risk of bias tool was 0.83 (95% CI 0.76–0.88) after adjusting for reliability, indicating strong convergence. The PEDro score was inversely associated with effect sizes, significantly associated with the summary score for the CBN risk of bias tool, and not associated with the journal impact factor. The interrater reliability for each item of the PEDro scale and CBN risk of bias tool was at least substantial for most items (>0.60). The intraclass correlation coefficient for the PEDro score was 0.80 (95% CI 0.68–0.88), and for the CBN, risk of bias tool was 0.81 (95% CI 0.69–0.88).

Conclusion: There was evidence for the convergent and construct validity for the PEDro scale when used to evaluate methodological quality of pharmacological trials. Both risk of bias tools have acceptably high interrater reliability.

Keywords: Risk of bias; Drug; Reliability; Validation; Pharmaceutical Research; Methods

1. Introduction

The Physiotherapy Evidence Database (PEDro) scale [1] was developed to measure methodological quality and completeness of statistical reporting of reports of randomized and quasi-randomized controlled trials in physiotherapy. It was originally created to rank the search results in the PEDro evidence resource (www.pedro.org.au), so users could quickly identify trial reports that are more likely to be valid and have sufficient information to make their results interpretable. The PEDro scale has become one of the most used and useful tools to quantify methodological quality in physiotherapy trials [2–4]. The scale comprises 11 items: (1) inclusion criteria and source; (2) random allocation; (3) allocation concealment; (4) baseline comparability; (5) blinding of subjects; (6) blinding of therapists; (7) blinding of assessors; (8) over 85% follow-up; (9) intention-to-treat analysis; (10) between-group comparison; and (11)

Funding: T.P.Y. is supported by CAPES (Coordination for the Improvement of Higher Education Personnel), Ministry of Education, Brazil. C.M. is supported by a Principal Research Fellowship from the National Health and Medical Research Council, Australia.

* Corresponding author. School of Public Health, Sydney Medical School, The University of Sydney, Edward Ford Building (A27), Sydney, NSW 2006, Australia. Tel.: +61 2 9036 9262; fax: +61 2 9036 9019.
E-mail address: tyamato@georgeinstitute.org.au (T.P. Yamato).

http://dx.doi.org/10.1016/j.jclinepi.2017.03.002
0895-4356/

What is new?

Key findings

- The Physiotherapy Evidence Database (PEDro) scale summary score and individual items have acceptable measurement properties to evaluate methodological quality of pharmacological trials.

What this adds to what was known?

- The PEDro scale and Cochrane Back and Neck risk of bias tool have similar clinimetric properties, and scores are strongly correlated.

What is the implication and what should change now?

- There is no empirical basis to favor one tool over the other.

point estimates and variability. The total PEDro score is calculated by counting the number of "yes" responses for items 2–11 (item 1 is not used for calculation of the total PEDro score because it is more related to external validity) and ranges from 0 to 10 points.

The clinimetric properties of the PEDro scale have been tested, and it has been shown to have acceptable validity, with evidence for convergent and construct validity for 8 of 11 individual items [5]. Previous studies reported acceptably high reliability for the total PEDro score (intraclass correlation coefficient [ICC] = 0.58–0.91) [1,6], as well as the reliability of the individual scale items (kappa = 0.50–0.88) [1,7–9]. In addition, a Rasch analysis revealed that the PEDro scale can be used as a continuous scale [10].

Other instruments have been developed to measure the methodological quality of trials evaluating health interventions [2]; however, none are considered to be the gold standard measure. The Jadad scale [11], for example, is widely used by the health care community [2] but considers only three components (randomization, blinding, and withdrawals or dropouts), all of which are included in the 11-item PEDro scale. Furthermore, the Jadad scale may not be responsive enough to allow discrimination between different levels of quality [12]. This is an important issue as evaluation of methodological quality influences how the results of clinical trials are incorporated into systematic reviews [1,13,14]. The Cochrane risk of bias tool [15] is used to assess methodological quality in Cochrane systematic reviews. This tool addresses seven domains (sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reporting, and "other issues"), all of which are included in the PEDro scale except for the domain "other issues." Different versions of the Cochrane risk of bias tool are in use; for example, the Cochrane Back and Neck (CBN) Group use a 12-item version which includes intention-to-treat analysis, group similarity at baseline, cointerventions, compliance, and timing of outcome assessments in addition to the seven domains [16]. In terms of clinimetric properties, the Cochrane risk of bias tool appears to have only been evaluated for reliability. The interrater reliability ranges from poor to substantial for individual domains (kappa = 0.13–0.74) [17,18], and the reliability between pairs of reviewers is considered "fair" for most domains (kappa = 0.24–0.37) [19].

Although the PEDro scale has acceptable validity and high reliability, one criticism is that it can only be used for physiotherapy trials [2,17]. This is despite the fact that the PEDro scale contains no physiotherapy-specific items and was based on a Delphi list of trial characteristics judged by clinical trial experts to be related to trial quality for all health care interventions [14]. In addition, the PEDro scale has been used in systematic reviews of many different health care interventions, such as exercise [16], psychological or behavioral interventions [20,21], and medical or pharmacological treatments [22,23]. Although this background suggests that it may be reasonable to use the PEDro scale for rating nonphysiotherapy trials, no study has directly evaluated the measurement properties of the PEDro scale when used in this manner. Therefore, the aim of this study was to investigate the convergent validity, the construct validity, and interrater reliability of the PEDro scale when used to evaluate the methodological quality of randomized controlled trials of analgesic medicines for back pain or osteoarthritis. The interrater reliability was also examined for the CBN Group risk of bias tool.

2. Methods

This is a secondary analysis of data from a previous study which evaluated the impact of an enriched enrollment design on the estimates of treatment effects of randomized placebo controlled trials evaluating any pain medication for chronic spinal pain (back or neck) or osteoarthritis (under review). Full details of the methods are available in the review protocol (PROSPERO 2014: CRD42014009988). Briefly, the previous study included 53 pharmaceutical trials that evaluated pain medications (i.e., opioid, nonsteroidal anti-inflammatory, paracetamol, alternative medicine, or combinations of these medicines) for spinal pain or osteoarthritis for short-term pain outcome. The trials were scored using the PEDro scale by two raters, with a third rater arbitrating any disagreements between the raters. Effect sizes for pain intensity for each trial were also extracted by two raters, with any disagreements resolved by consensus discussions and, if necessary, arbitration by a third rater.

To evaluate convergent validity of the PEDro scale for pharmaceutical trials, we compared the total PEDro score with summary score calculated for the risk of bias tool of the CBN Group [16]. We used the PEDro ratings from a previously published study evaluating pharmacological trials and the ratings for the CBN Group risk of bias tool

were determined by two independent reviewers (same reviewers from the PEDro ratings). Any disagreements were resolved by discussion or arbitration of a third reviewer. Although the Cochrane Handbook recommends against using a summary score for reporting quality in systematic reviews [15], we calculated a summary score for the Cochrane risk of bias tool to allow comparison with the total PEDro score. The number of items classified as ''low risk of bias'' was counted to give a score out of 12, ranging from 0 to 12.

We set three constructs to evaluate the construct validity of the PEDro scale. The first was that the total PEDro score would be inversely associated with treatment effect size. This is based on research showing that higher quality trials are more likely to report smaller treatment effect sizes [24]. The second was that the total PEDro score would be positively correlated with journal impact factor; this is based on the view that higher impact journals are more likely to publish trials of higher quality. The third was that the summary score of the Cochrane risk of bias tool would be associated with the PEDro scale, based on the idea that both scales are sensible to measure methodological quality in pharmaceutical trials.

Effect sizes were extracted from the included studies for the primary analysis and expressed as mean differences and 95% confidence intervals. Effect sizes were extracted for short-term (less than 3 months from randomization), intermediate-term (at least 3–12 months from randomization), and long-term (more than 12 months from randomization) follow-ups, but only the short-term follow-ups were used in this secondary analysis as this provided the highest number of included trials. Pain outcomes were all converted to a 0 (no pain) to 100 (worst possible pain) scale. We extracted the journal impact factor for each study from the SCImago database for the year that the article was published.

To evaluate interrater reliability of the PEDro scale for pharmaceutical trials, we used the ratings generated by the two independent raters (i.e., before arbitration of disagreements by a third rater). The ratings for individual items and summary scores for both the PEDro scale and CBN Group risk of bias tool were used for the analysis.

2.1. Analysis

Convergent validity was evaluated by correlating the total PEDro score and the summary score for the Cochrane risk of bias tool using Spearman correlation coefficient (and 95% confidence interval [CI]). A correlation coefficient of 0.70 or higher was considered as strong convergence, 0.50–0.69 as moderate convergence, 0.20–0.40 as moderate divergence, and less than 0.20 as strong divergence [25,26]. Because neither tool has been reported with perfect reliability, we corrected for an attenuation due to imperfect reliability (calculated in our study) using the Spearman Brown Prophecy formula [5].

The construct validity was tested using a linear regression analysis to determine the degree to which the total PEDro score is associated with treatment effect sizes, journal impact factor, and the summary score for the Cochrane risk of bias tool. We conducted this regression analysis with each of the three variables separately. P-values <0.05 were considered to be significant.

The interrater reliability for the PEDro scale and CBN Group risk of bias tool were estimated by comparing the first and second assessor's ratings. The Prevalence and Bias Adjusted Kappa (PABAK) coefficient and 95% CI were used to quantify the reliability for each item. Interpretation of the PABAK coefficient has been described as follows: <0 = poor, 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, 0.81–1.00 = almost perfect [27]. Percentage of exact agreement was also calculated for each item. Intraclass correlation coefficients (ICCs) and 95% CIs were calculated for the total PEDro score and summary score for the Cochrane risk of bias tool.

All statistical analyses were performed with SAS Enterprise Guide 5.1 (Statistical Analysis System, Cary, NC, USA), IBM SPSS Statistics, version 20.0 (IBM Corp., Armonk, NY, USA), and the Diagnostic and Agreement Statistics calculator [28].

3. Results

Of the 53 trials included in the analysis (n = 21,183 participants), 17 evaluated opioid analgesics (29 subgroups of comparisons) and 36 evaluated other types of analgesics (62 subgroups of comparisons). Some trials evaluated both opioids and other analgesics in different groups and were included in both comparisons. Ninety-one comparisons from the 53 included trials were used in the regression analyses. The characteristics of the included studies are described in Table 1, and the full description of the included trials is presented in Appendix 1 at www.jclinepi.com.

The convergent validity was tested by a correlation between the total PEDro score and the summary score for the Cochrane risk of bias tool. The correlation between these two instruments was 0.61 (95% CI 0.46–0.72) representing a moderate convergence. When the analysis was

Table 1. Characteristics of included studies (n = 53)

Sample size, median (IQR)	268 (83–491)
Mean age, median (IQR), yrs	58.3 (50.6–61.8)
Total PEDro score, median (IQR)	7.0 (7.0–8.0)
Cochrane risk of bias tool summary score, median (IQR)	7.0 (5.0–8.0)
Condition, n (%)	
Spinal pain	15 (28)
Osteoarthritis (hip or knee)	38 (72)

Abbreviations: IQR, interquartile range; PEDro, Physiotherapy Evidence Database.

adjusted for reliability, the corrected correlation was 0.83 (95% CI 0.76–0.88) indicating a strong convergence.

The construct validity of the PEDro scale was tested using a linear regression of the total PEDro score with treatment effect size, journal impact factor, and the summary score of the Cochrane risk of bias tool (Table 2). The total PEDro score was significantly associated with the effect sizes, where a 1-point higher total PEDro score was associated with a decrease in effect size of 0.07 (on a 100-point visual analog scale). The total PEDro score was also significantly associated with the summary score for the Cochrane risk of bias tool, where 1 point in the total PEDro score was associated with an increase of 0.51 points in the summary score of the Cochrane risk of bias tool. The total PEDro score was not associated with the journal impact factor in the final model (regression coefficient = 0.31).

The PABAK coefficient was used to quantify the interrater reliability for each item of the PEDro scale for the total sample of 53 studies (Table 3). PABAK values exceeded 0.60 (i.e., classified as at least substantial agreement—italicized in the table) for all but two PEDro scale items. Random allocation, between-group comparison, over 85% follow-up, intention-to-treat analysis, baseline comparability, and point estimates and variability were classified as almost perfect agreement. Blinding of assessors, allocation concealment, and inclusion criteria and source were classified as substantial agreement. Blinding of therapists and blinding of subjects were classified as having moderate agreement. The ICC for the total PEDro score was 0.80 (95% CI 0.68–0.88), indicating substantial agreement.

The PABAK coefficient was also used to quantify the interrater reliability for each item of the Cochrane risk of bias tool for the total sample of 53 studies (Table 4). PABAK values exceeded 0.60 (i.e., classified as at least substantial agreement—italicized in the table) for all but three Cochrane tool items. Sequence generation, incomplete outcome data, selective outcome reporting, group similarities at baseline, and timing of outcome assessment were classified as almost perfect agreement, whereas allocation concealment, blinding of personnel, cointerventions, and compliance were classified as substantial agreement, and the blinding of participants, blinding of assessors, and intention-to-treat analysis were classified as moderate agreement. The ICC for the Cochrane risk of bias tool summary score was 0.81 (95% CI 0.69–0.88), indicating an almost perfect agreement.

Table 2. Construct validity—linear regression on the association of the total PEDro score, effect sizes, journal impact factor, and summary score for the Cochrane risk of bias tool

	Linear regression	
Comparison	**Regression coefficient (95% CI)**	***P*-value**
Total PEDro score vs. effect size (mean difference)	0.07 (0.03, 0.10)	0.0007***
Total PEDro score vs. journal impact factor	0.31 (−0.15, 0.78)	0.19[N.S.]
Total PEDro score vs. summary score for Cochrane risk of bias tool	0.51 (0.40, 0.62)	<.0001***

Abbreviations: PEDro, Physiotherapy Evidence Database; CI, confidence interval; N.S., not significant.
***$P < 0.001$.

Table 3. Interrater reliability—Prevalence and Bias Adjusted Kappa (PABAK) and percent exact agreement for each PEDro scale item

PEDro scale items	**PABAK coefficient**	**% Agreement**
Item 5, blinding of subjects	0.51	75.5
Item 6, blinding of therapists	0.59	79.3
Item 1, inclusion criteria and source	*0.62*	81.1
Item 3, allocation concealment	*0.70*	84.9
Item 7, blinding of assessors	*0.77*	88.7
Item 4, baseline comparability	*0.85*	92.5
Item 11, point estimates and variability	*0.85*	92.5
Item 9, intention-to-treat analysis	*0.89*	94.3
Item 8, over 85% follow-up	*0.93*	96.2
Item 2, random allocation	*1.00*	100.0
Item 10, between-group comparison	*1.00*	100.0

Abbreviation: PEDro, Physiotherapy Evidence Database.
PABAK values exceeding 0.60 are italicized.

4. Discussion

There was evidence for the convergent validity of the PEDro scale when used to assess pharmacological trials as evidenced by a strong corrected correlation ($r = 0.83$) between the total PEDro score and Cochrane risk of bias tool summary score. There was also evidence for construct validity with the total PEDro score being inversely associated with the effect sizes and positively associated with the summary score for the Cochrane risk of bias tool (but the total PEDro score was not associated with journal impact factor). The interrater reliability for each item of the PEDro scale and Cochrane risk of bias tool was substantial for most items. The ICCs for both the total PEDro score and summary score Cochrane risk of bias tool indicated

Table 4. Interrater reliability—Prevalence and Bias Adjusted Kappa (PABAK) and percent exact agreement for each Cochrane tool item

Cochrane tool items	**PABAK coefficient**	**% Agreement**
Item 3, blinding of participants	0.59	77.4
Item 5, blinding of outcome assessment	0.59	75.5
Item 7, intention-to-treat analysis	0.59	66.0
Item 10, cointervention	*0.66*	81.1
Item 4, Blinding of personnel	*0.74*	77.4
Item 2, Allocation concealment	*0.77*	84.9
Item 11, Compliance	*0.77*	86.8
Item 6, Incomplete outcome data	*0.81*	86.8
Item 8, Selective outcome reporting	*0.85*	92.5
Item 9, Group similarity at baseline	*0.89*	92.5
Item 1, Sequence generation	*0.96*	98.1
Item 12, Timing of outcome assessment	*1.00*	100.0

PABAK values exceeding 0.60 are italicized.

substantial agreement and almost perfect agreement, respectively.

This is the first study evaluating the measurement properties of the PEDro scale in pharmaceutical trials, including comparing the PEDro scale and the Cochrane risk of bias tool. The PEDro scale was shown to be both valid and reliable for assessing methodological quality of trials evaluating pharmacological interventions. Although the lack of a gold standard for assessing the methodological quality of clinical trials prevented the evaluation of criterion validity, both convergent validity and construct validity were evaluated. Our analysis of construct validity was thorough, including effect sizes, journal impact factor, and the Cochrane risk of bias tool.

Our results indicating that the PEDro scale is a valid and reliable tool for assessing risk of bias in trials evaluating pharmaceutical interventions concur with previous studies that have tested the measurement properties of the PEDro scale when used to assess trials evaluating physiotherapy interventions [2,5,17]. Based on the comparison between the PEDro scale and Cochrane risk of bias tool, we can conclude that there is no reason to favor the Cochrane risk of bias tool over the PEDro scale to assess methodological quality of trials or summarize the methodological quality in systematic reviews as the scales are strongly correlated. Our data repudiate two previous comparisons, which suggested that the PEDro scale produced different ratings of methodological quality when compared with the Cochrane risk of bias tool [29,30]. In both cases, arbitrary cut points were applied to the total PEDro score and these were compared to an abbreviated Cochrane risk of bias tool (three items only—sequence generation, allocation concealment, blinding of outcome assessment). Importantly, we have shown that when the full version of each instrument is used (and analyses are corrected for the imperfect reliability of the instruments), the PEDro scale and Cochrane risk of bias tool are highly correlated [31]. This concordance is to be expected as the two scales contain quite similar items.

According to our findings, either the PEDro scale or the Cochrane risk of bias tool can be used to measure methodological quality in pharmaceutical trials as they are highly correlated. In our study, the PEDro total score and the summary score for the Cochrane risk of bias tool had a positive association. The inverse association between the total PEDro score and effect sizes confirms our hypothesis that poor methodological quality is related with higher effect estimates [24]. However, we did not find any association with journal impact factor. Macedo et al. [5] found a weak but significant association with journal impact factor but they did not investigate pharmaceutical trials. Moreover, we cannot discard this association in future observations.

Regarding the interrater reliability, nearly all of the PEDro scale and Cochrane risk of bias tool items had almost perfect agreement. Items with the lowest reliability were blinding of subjects and blinding of therapists for the PEDro scale and blinding of participants, blinding of outcome assessment, and intention-to-treat analysis for the Cochrane risk of bias tool. This low reliability of the blinding items may have more to do with the reporting of trials than the scale items per se [32]. The use of terms like single-blind, double-blind, and triple-blind to describe blinding in trials is confusing because health care professionals and textbooks have used different definitions for these terms (e.g., in a survey of physicians and textbooks, there were 15 and 7 different definitions of triple-blind, respectively) [32]. If trial authors do not provide further details on who was blinded (i.e., participant, assessor, care provider), it makes it difficult for the reviewers to score blinding on risk of bias instruments.

We have provided strong evidence that the PEDro scale can be used to evaluate risk of bias in trials evaluating health care interventions beyond physiotherapy trials. The PEDro scale was valid and reliable for assessing methodological quality in pharmaceutical trials. The Cochrane risk of bias tool presented similar reliability and was strongly correlated with the PEDro scale. Future studies could explore the agreement between individual items in common to both instruments as well as explore the question of whether methodological quality can be quantified by a continuous score or needs to be evaluated by a series of individual items.

5. Conclusion

We have provided strong evidence that the PEDro scale can be used to evaluate methodological quality in trials evaluating health care interventions beyond physiotherapy trials. The PEDro scale was valid and reliable for assessing methodological quality in pharmaceutical trials. The Cochrane risk of bias tool presented similar reliability and was strongly correlated with the PEDro scale. Future studies could explore the agreement between individual items in common to both instruments as well as explore the question of whether methodological quality can be quantified by a continuous score or needs to be evaluated by a series of individual items.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2017.03.002.

References

[1] Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther 2003;83:713−21.

[2] Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008;88:156−75.

[3] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of

intervention efficacy reported in meta-analyses? Lancet 1998;352:609−13.
[4] Moseley AM, Elkins MR, Herbert RD, Maher CG, Sherrington C. Cochrane reviews used more rigorous methods than non-Cochrane reviews: survey of systematic reviews in physiotherapy. J Clin Epidemiol 2009;62:1021−30.
[5] Macedo LG, Elkins MR, Maher CG, Moseley AM, Herbert RD, Sherrington C. There was evidence of convergent and construct validity of Physiotherapy Evidence Database quality scale for physiotherapy trials. J Clin Epidemiol 2010;63:920−5.
[6] Foley NC, Bhogal SK, Teasell RW, Bureau Y, Speechley MR. 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;86:817−24.
[7] 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(1):3−11.
[8] Van Peppen RP, 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(1):3−9.
[9] Van Peppen RP, Kwakkel G, Wood-Dauphinee S, Hendriks HJ, Van der Wees PJ, Dekker J. The impact of physical therapy on functional outcomes after stroke: what's the evidence? Clin Rehabil 2004;18(8):833−62.
[10] de Morton NA. The PEDro scale is a valid measure of the methodological quality of clinical trials: a demographic study. Aust J Physiother 2009;55(2):129−33.
[11] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1−12.
[12] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249−56.
[13] Colle F, Rannou F, Revel M, Fermanian J, Poiraudeau S. 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;83:1745−52.
[14] Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, 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;51:1235−41.
[15] Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. Available at. www.cochrane-handbook.org. Accessed on August 2016.
[16] Furlan AD, Malmivaara A, Chou R, Maher CG, Deyo RA, Schoene M, et al. 2015 Updated Method Guideline for Systematic Reviews in the Cochrane Back and Neck Group. Spine (Phila Pa 1976) 2015;40(21):1660−73.
[17] Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs Seida J, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ 2009;339:b4012.
[18] Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, Fuentes J, et al. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. PLoS One 2014;9:e96920.
[19] Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;66:973−81.
[20] Steel JL, Bress K, Popichak L, Evans JS, Savkova A, Biala M, et al. A systematic review of randomized controlled trials testing the efficacy of psychosocial interventions for gastrointestinal cancers. J Gastrointest Cancer 2014;45:181−9.
[21] Sheinfeld Gorin S, Krebs P, Badr H, Janke EA, Jim HS, Spring B, et al. Meta-analysis of psychosocial interventions to reduce pain in patients with cancer. J Clin Oncol 2012;30:539−47.
[22] Machado GC, Ferreira PH, Harris IA, Pinheiro MB, Koes BW, van Tulder M, et al. Effectiveness of surgery for lumbar spinal stenosis: a systematic review and meta-analysis. PLoS One 2015;10:e0122800.
[23] Pinto RZ, Maher CG, Ferreira ML, Ferreira PH, Hancock M, Oliveira VC, et al. Drugs for relief of pain in patients with sciatica: systematic review and meta-analysis. BMJ 2012;344:e497.
[24] 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;273:408−12.
[25] Fowler J, Jarvis P, Chevannes M. Practical statistics for nursing and health care. 1st ed. Hoboken, New Jersey: Wiley; 2002.
[26] Streiner D, Norman G. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford, UK: Oxford University Press; 2008.
[27] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[28] Mackinnon A. A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. Comput Biol Med 2000;30(3):127−34.
[29] da Costa BR, Hilfiker R, Egger M. PEDro's bias: summary quality scores should not be used in meta-analysis. J Clin Epidemiol 2013;66:75−7.
[30] Armijo-Olivo S, da Costa BR, Cummings GG, Ha C, Fuentes J, Saltaji H, et al. PEDro or Cochrane to assess the quality of clinical trials? A meta-epidemiological study. PLoS One 2015;10:e0132634.
[31] Costa LO, Maher CG, Moseley AM, Elkins MR, Shiwa SR, Herbert RD, et al. da Costa and colleagues' criticism of PEDro scores is not supported by the data. J Clin Epidemiol 2013;66:1192−3.
[32] Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001;285:2000−3.