ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 934–939

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement

Damian Hoy[a,*], Peter Brooks[b], Anthony Woolf[c], Fiona Blyth[d], Lyn March[d], Chris Bain[a], Peter Baker[a], Emma Smith[d], Rachelle Buchbinder[e,*]

[a]*University of Queensland, Herston Road, Herston, Brisbane, QLD 4006, Australia*
[b]*Australian Health Workforce Institute, 766 Elizabeth Street, Melbourne 3010, Australia*
[c]*Peninsula College of Medicine and Dentistry, Truro TR1 3LJ, United Kingdom*
[d]*University of Sydney, Royal North Shore Hospital, St Leonards, Sydney 2065, Australia*
[e]*Cabrini Hospital and Monash University, Cabrini Medical Centre 183 Wattletree Rd, Malvern, Melbourne 3144, Australia*

Accepted 22 November 2011; Published online 27 June 2012

Abstract

Objective: In the course of performing systematic reviews on the prevalence of low back and neck pain, we required a tool to assess the risk of study bias. Our objectives were to (1) modify an existing checklist and (2) test the final tool for interrater agreement.

Study Design and Setting: The final tool consists of 10 items addressing four domains of bias plus a summary risk of bias assessment. Two researchers tested the interrater agreement of the tool by independently assessing 54 randomly selected studies. Interrater agreement overall and for each individual item was assessed using the proportion of agreement and Kappa statistic.

Results: Raters found the tool easy to use, and there was high interrater agreement: overall agreement was 91% and the Kappa statistic was 0.82 (95% confidence interval: 0.76, 0.86). Agreement was almost perfect for the individual items on the tool and moderate for the summary assessment.

Conclusion: We have addressed a research gap by modifying and testing a tool to assess risk of study bias. Further research may be useful for assessing the applicability of the tool across different conditions.

Keywords: Assessment; Bias; Prevalence; Instrument; Quality; Review

1. Introduction and background

Systematic reviews are often used by academic institutions, governments, and other agencies to synthesize information on a particular scientific question [1,2]. To minimize the potential for biases such as sampling or measurement bias in undertaking these reviews, rigorous methods should be used to locate, select, and aggregate the results of individual studies. Incorporating an assessment of risk of bias of these studies is essential in interpreting their results and may help to avoid under- or overestimating the parameter of interest [3].

Many tools have been developed to assess study quality, although most have been developed for experimental studies [4], and those relevant to observational studies are generally considered to be unsatisfactory and not widely used [5]. A recent systematic review of tools used to assess the quality of observational studies performed by Shamliyan et al. [6] identified only five tools relevant to prevalence or incidence studies. In general, the reviewed tools provided only limited information about development, reliability, and validation; did not discriminate poor reporting from quality of studies; and did not separate internal from external validity. In addition, there was no consensus around individual criteria of validity or ranking of overall quality and the authors highlighted the need for collaborative efforts to develop transparent quality assessment for observational research. Importantly, they also acknowledged the paucity of empirical knowledge about how potential sources of bias may influence the results of observational research.

Some expert groups, most notably The Cochrane Collaboration, have drawn a distinction between assessment of methodological quality and assessment of risk of bias, which refers to whether a study has answered the research question in a manner that is free from bias [3]. This change in emphasis for considering the internal validity of a study reflects recognition that studies may be performed to the highest possible standards, yet still have an important risk of bias, and that some current markers of study quality

* Corresponding author.
E-mail address: damehoy@yahoo.com.au (D. Hoy) or rachelle.buchbinder@med.monash.edu.au (R. Buchbinder).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.11.014

What is new?

Key findings

- We developed a risk of bias tool for prevalence studies based on a review of the literature, expert consensus, pilot testing of draft items, and refinement of the tool.
- The tool consists of 10 items addressing four domains of bias plus a summary risk of bias assessment.
- Raters found the tool easy to use, and we demonstrated high interrater agreement of the tool in assessing risk of bias of prevalence studies of low back and neck pain (overall agreement: 91%; Kappa statistic: 0.82).

What this adds to what was known?

- Many tools have been developed to assess study quality, but of those relevant to prevalence studies, most are considered unsatisfactory.
- The tool developed in this study makes a distinction between assessing whether the research was conducted to the highest possible standards (methodological quality) and the extent to which the results can be believed (risk of bias).
- Assessment of risk of bias provides invaluable information that can be incorporated into the analysis and/or used to interpret the findings of systematic reviews of disease prevalence.

What is the implication and what should change now?

- Further research is needed to assess the applicability of the tool across different conditions and in different settings.

(e.g., obtaining ethical approval and performing a sample size calculation) are unlikely to have direct implications for risk of bias [3]. In addition, an emphasis on risk of bias helps overcome ambiguity between the quality of reporting and the quality of the underlying research. Although these issues have primarily been considered for systematic reviews of randomized controlled trials, they are also pertinent for systematic reviews of other study types.

In the course of performing systematic reviews on the prevalence of low back and neck pain as part of the Global Burden of Disease 2010 Study (GBD 2010) [7], we were unable to find a tool specifically designed to assess risk of bias in prevalence studies that provided detailed information on how it was developed, had been tested for interrater agreement, and had clear guidelines for use. We did identify a checklist developed by Leboeuf-Yde and Lauritsen [8] in the early 1990s to assess study quality in a review of the prevalence of low back pain in Nordic countries. It included 11 criteria related to representativeness, data quality, and case definition. As well as a lack of provision of instructions for its use and empirical evidence of its reliability, it used an arbitrary system to numerically score study quality (if 75% or more of the criteria were fulfilled, the authors considered the study to be methodologically acceptable). The authors acknowledged the likelihood of reviewer bias with the nominal cutoff point for determining study acceptability and emphasized the importance of having stringent, systematic criteria in future such reviews [8]. Walker [9], Dionne et al. [10], and Louw et al. [11] subsequently adapted the tool by respectively adding an item on whether a proxy respondent was used in the data collection process, modifying the response rate item, and altering the threshold for study acceptability.

The objectives of our study were to (1) develop a new risk of bias tool for prevalence studies based on modifying the Leboeuf-Yde and Lauritsen tool from recommendations from previous users, a literature search, and an expert consensus exercise and (2) test the final tool for interrater agreement.

2. Methods

The process we followed is shown in Fig. 1. To establish face validity of the tool, we (1) examined checklists that were regarded as good quality in previous reviews [12–14]; (2) listed key recommendations from previous reviews of study assessment tools [4,6,12,15]; (3) searched the peer-reviewed literature for tools for assessing risk of study bias, particularly those relevant to low back and neck pain prevalence studies [3,8–11,13–17]; (4) collated a list of relevant criteria for assessing risk of bias in prevalence studies; and (5) undertook a series of consensus exercises with the Global Burden of Disease GBD 2010 Low Back and Neck Pain expert group (comprising six musculoskeletal epidemiologists), to generate a more refined list of core criteria. The exercise involved rounds of discussion and ranking of the list of core criteria. Each round comprised commenting on and ranking items via e-mail, followed by a teleconference to further discuss and refine the list. The aim was not to make an exhaustive list but to reduce the list to the most relevant and practical items to apply. In line with recent recommendations, items related to study reporting, as opposed to risk of bias, were not included [6]. After five rounds of discussion and ranking, we agreed on a draft tool.

The pretesting of the draft tool was conducted in three stages. In stage 1, one author (D.H.) used the tool on a random sample of 30 of the 240 studies included in the low back and neck pain prevalence systematic reviews, and areas of ambiguity within the tool were clarified. In stage 2, a further 12 studies were randomly selected and these were

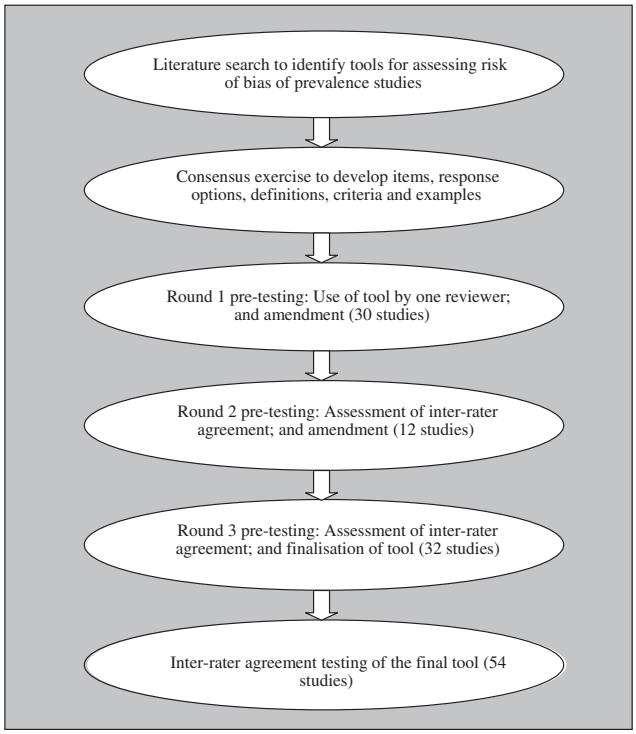

Fig. 1. The process for developing the risk of bias tool and testing interrater agreement.

assessed by three authors (D.H., P.Br., and R.B.). Agreement was examined and further refinements made. For the third stage, six researchers (R.B., P.Br., L.M., A.W., F.B., and a researcher from Thailand) assessed a further four to six randomly selected studies each. D.H. also assessed these studies, and these results were compared with those of the six reviewers. Based on these comparisons, the tool was further amended to clarify areas of uncertainty that appeared to limit agreement.

Using the final tool, two researchers (D.H. and a research assistant) independently assessed the risk of bias of 54 studies chosen at random from the list of the neck pain prevalence articles. One of the reviewers (D.H.) was involved in the development of the tool and had significant experience in critical appraisal. The research assistant had a master's degree in Public Health, had some experience in critical appraisal, and was not involved in the tool development. She had 30 minutes of training in use of the tool that primarily included demonstration of how to check articles for data errors resulting from typographical errors or miscalculations (item 10). There was no attempt to blind reviewers to study authors, institutions, or journal.

Agreement between the two raters was assessed overall and for each item in the tool using proportion of agreement (p_0) and the Kappa statistic. The p_0 is the raw proportion of the number of times the two raters agree. A disadvantage of p_0 is that it does not take into account chance agreement, how agreement is distributed, or ordering of agreement in the case of ordinal data [18]. A benefit of using the Kappa statistic is that it does take chance agreement into account. Kappa values range from −1 to +1, with 0 and less considered poor agreement, 0.01 to 0.20 slight, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and 0.81 to 0.99 almost perfect agreement [19]. For assessment of the interrater agreement for the summary item, we weighted the Kappa statistic to take ordering of response options into account [20]. Raw agreement, Kappa values, and their 95% confidence intervals (CIs) were calculated using Stata Version 10 (StataCorp, College Station, TX, USA) [21].

3. Results

The final risk of bias tool comprises 10 items plus a summary assessment (see Appendix on the journal's Web site at www.jclinepi.com). Items 1 to 4 assess the external validity of the study (domains are selection and nonresponse bias), and items 5 to 10 assess the internal validity (items 5 to 9 assess the domain of measurement bias, and item 10 assesses bias related to the analysis). We included the item assessing whether the results of the study were representative of the national population (item 1) as this is an important issue for GBD 2010 and indeed for any attempts to describe disease patterns in a country. The summary assessment evaluates the overall risk of study bias and is based on the rater's subjective judgment given responses to the preceding 10 items. This is based on the Grades of Recommendation, Assessment, Development and Evaluation (GRADE) and Cochrane approaches [3,22].

The main differences from the tool originally developed by Leboeuf-Yde and Lauritsen [8] and adapted by others [9–11] are that (1) items on representativeness are separated into representativeness of the target population and representativeness of the sample frame and determination of whether random methods were used to select the sample or whether a census was undertaken; (2) the item on whether the study was primarily performed to collect data on low back pain prevalence was excluded as we argue that if measures for minimizing study bias are performed, such as provision of a clear case definition, representativeness, and validation or other testing of the survey instrument, then bias should not be affected by whether the survey was designed specifically for a particular condition, in this case, low back pain; (3) for each item, the emphasis is risk of bias rather than quality of reporting; (4) we added instructions for use and examples to our tool; and (5) we undertook rigorous pretesting to ensure that the instructions and criteria for assessing risk of bias are clear. Where appropriate, we aligned the tool with the risk of bias tool used by The Cochrane Collaboration for assessing risk of bias of experimental studies [3].

Response options for individual items were either low or high risk of bias. If there was insufficient information in the article to permit a judgment for a particular item, then the item was deemed to be at high risk of bias. We elected not to include an unclear response option because in most

instances methods that minimize risk of bias are likely to be reported in the article, and, in practice, most empirical studies have combined the high and unclear categories into a single category in any case [3]. We initially also included a moderate risk of bias response option but found that this response option was used to negate having to make a decision between high and low risk of bias. When this response option was removed, agreement improved considerably, although this may have also been influenced by other amendments to the tool. Response options for the summary assessment were low, moderate, or high risk of bias.

Detailed criteria and examples are given for each item to help guide reviewers. Raters reported that they found the tool easy to use and the examples and additional notes very helpful. We had initially been concerned that reviewers may find the amount of text in the tool overwhelming and the process too time consuming; however, reviewers reported that once the first two or three studies had been assessed, the process became substantially quicker.

Overall interrater agreement, which was calculated by considering each of the 10 items plus the summary assessment for the 54 studies, was 91% (539/594 items) with a Kappa statistic of 0.82 (95% CI: 0.76, 0.86), indicating an almost perfect level of agreement (Table 1). Raw agreement between the two reviewers for individual items ranged from 83% to 100% with Kappa values indicating substantial or almost perfect agreement for nine of the 10 items. Moderate agreement between the raters was observed for the item assessing whether the study's sampling frame was a true or close representation of the target population (item 2).

For the summary assessment of a study's risk of bias, agreement between the raters was 72% (weighted agreement, 85%) with a weighted Kappa value of 0.48 (95% CI: 0.31, 0.64), indicating a moderate level of agreement. For this item, both assessors rated two studies as low risk of bias, 27 studies as moderate risk of bias, and 25 studies as high risk of bias. There was perfect agreement for 39 of the 54 studies, moderate disagreement for 14 studies (i.e., where one assessor rated a study as moderate risk of bias and the other rated it low or high risk of bias), and substantial disagreement (i.e., where one assessor rated a study as high risk of bias and the other rated it low risk of bias) for one of the 54 studies. When we limited our analysis to the 10 individual items on the tool, agreement between the raters was 93% with a Kappa value of 0.83 (95% CI: 0.78, 0.88).

4. Discussion

Assessing the risk of study bias is an important step in performing and interpreting systematic reviews of the literature. Risk of bias tables are now routinely included in Cochrane systematic reviews that assess the efficacy and safety of treatment interventions [3]. Reviews have found that tools used for assessing quality of nonrandomized studies are poorly developed or have been specifically designed to assess randomized controlled trials and thus fail to include key criteria important for nonrandomized studies [12].

We developed a new risk of bias tool because we were unable to identify any existing risk of bias tools for prevalence studies. Although checklists developed by

Table 1. Results of interrater agreement testing between two raters in using the risk of bias tool

Item	% Complete agreement	Kappa value	95% Confidence intervals	
			Lower limit	Upper limit
External validity				
1. Was the study's target population a close representation of the national population in relation to relevant variables?	94	0.84	0.67	1.00
2. Was the sampling frame a true or close representation of the target population?	83	0.43	0.12	0.74
3. Was some form of random selection used to select the sample, OR was a census undertaken?	96	0.85	0.66	1.00
4. Was the likelihood of nonresponse bias minimal?	89	0.78	0.61	0.94
Internal validity				
5. Were data collected directly from the subjects (as opposed to a proxy)?	98	0.79	0.39	1.00
6. Was an acceptable case definition used in the study?	94	0.82	0.63	1.00
7. Was the study instrument that measured the parameter of interest shown to have validity and reliability?	94	0.89	0.77	1.00
8. Was the same mode of data collection used for all subjects?	100	1.00	1.00	1.00
9. Was the length of the shortest prevalence period for the parameter of interest appropriate?	87	0.61	0.35	0.87
10. Were the numerator(s) and denominator(s) for the parameter of interest appropriate?	89	0.61	0.34	0.88
11. Summary item on the overall risk of study bias	72	0.48[a]	0.31[a]	0.64[a]
Overall agreement for the 11 items	91	0.82	0.76	0.86

[a] Weighted.

Strengthening the Reporting of Observational Studies in Epidemiology statement [23] and Tooth et al. [24] are useful for assessing quality of reporting, their intention is not to specifically measure risk of study bias. As outlined previously, bias can occur in well-conducted studies and not all methodological flaws introduce bias. The only tool that we found that approached our criteria was a methodological checklist that had been used in previous systematic reviews of the prevalence of low back pain [8−11,25]. However, there was no published information about its development and no guidelines for use and it used an arbitrary system to numerically score study quality and had no data on reliability. Furthermore, several authors had already modified the checklist based on observed deficiencies in the content and scoring of the tool [9−11].

The strengths of our risk of bias tool are that it was developed based on an exhaustive literature review to identify potentially relevant items followed by an expert consensus exercise. Although it is not possible to directly measure the presence or absence of bias, our tool focuses on identifying whether studies had attempted to minimize bias. We deliberately did not include an overall numeric rating of risk of study bias but, instead, made a judgment of the overall risk of study bias based on assessment of risk of bias of 10 individual items. This approach is consistent with the Cochrane and GRADE systems [3,22]. In addition, we have provided empirical evidence of its interrater agreement.

Agreement was generally higher for individual items on the tool than for the summary assessment of overall risk of bias. This was not surprising, given that criteria for the overall assessment were less prescriptive than for individual items, and there were three response options rather than two increasing the opportunity for disagreement. Most summary ratings were for moderate or high risk of bias (104/108), and raters reported difficulty in drawing a distinction between the two. The inclusion of a summary assessment of risk of bias was discussed at length in the development of the tool. Some authors wanted to avoid a global study rating as they felt it could be potentially misleading as some biases may be more important than others, a position also held by a number of other experts [26,27], whereas other authors argued that a summary rating would be important in the analysis phase of GBD 2010 Study. The resulting consensus was to include the item to be consistent with approaches used by The Cochrane Collaboration and GRADE [3,22]. In addition, having an overall judgment of risk of bias does not preclude the performance of sensitivity analyses to explore individual items of bias in systematic reviews.

An important consideration for being able to use the tool in other settings is whether a minimal level of critical appraisal expertise and training is required to ensure the concepts within the tool are well understood. For the reliability exercise, we purposely included a research assistant with no prior knowledge of the tool. We found that 30 minutes of training was sufficient for her to understand and use the tool, and both raters found the tool easy to use. However, this needs further evaluation in different settings. A limitation of our study was that we did not record the time taken to assess the risk of bias of each study. However, both raters estimated that the average time to review an article was between 30 and 60 minutes.

Coincident with the development of our tool, Shamliyan et al. [28] recently published a pilot study reporting on the development and testing of two checklists developed to assess the quality of observational studies of incidence, prevalence, or risk factors of diseases. They undertook a similar approach in developing key criteria for assessing study quality including literature searches for existing tools for assessing the quality of prevalence studies, generation of a list of items for use, and refinement of items by incorporating the views of experts. While items considered important for inclusion were largely consistent with our tool, including sampling bias, nonresponse bias, and measurement bias, their objective was to assess the methodological *and* reporting quality of observational studies. In contrast to our reliability results, the interrater agreement of their checklist tested by seven experts on 10 articles was found to be poor and further refinement was deemed necessary to improve interrater agreement. One possible explanation for the poor agreement may be the lack of specific examples to assist users, a strength of our tool. They reported that the completion of the study assessment was time consuming, particularly for those studies with major flaws.

5. Conclusions

We have developed a risk of bias tool, which builds on previous work, to assess risk of bias of studies measuring disease prevalence. This tool was found to be easy to apply and demonstrated high interrater agreement. We have now applied our risk of bias tool for assessing prevalence studies of gout, osteoarthritis, and rheumatoid arthritis for GBD 2010. Further research is needed to assess the reliability of the tool for assessing prevalence studies of other conditions.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.jclinepi.2011.11.014.

References

[1] Fox DM. Evidence of evidence-based health policy: the politics of systematic reviews in coverage decisions. Health Aff (Millwood) 2005;24:114−22.

[2] Lavis J, Davies H, Oxman A, Denis JL, Golden-Biddle K, Ferlie E. Towards systematic reviews that inform health care management and policy-making. J Health Serv Res Policy 2005;10:35−48.

[3] Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0.: The Cochrane Collaboration, 2011. Available at www.cochrane-handbook.org; Accessed August 24, 2011.

[4] Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004;4:1−11.
[5] Mallen C, Peat G, Croft P. Quality assessment of observational studies is not commonplace in systematic reviews. J Clin Epidemiol 2006;59:765−9.
[6] Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol 2010;63:1061−70.
[7] Hoy D, March L, Brooks P, Woolf A, Blyth F, Vos T, et al. Measuring the global burden of low back pain. Best Pract Res Clin Rheumatol 2010;24(2):155−65.
[8] Leboeuf-Yde C, Lauritsen JM. The prevalence of low back pain in the literature. A structured review of 26 Nordic studies from 1954 to 1993. Spine 1995;20(19):2112−8.
[9] Walker BF. The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. J Spinal Disord 2000;13(3):205−17.
[10] Dionne CE, Dunn KM, Croft PR. Does back pain prevalence really decrease with increasing age? A systematic review. Age Ageing 2006;35:229−34.
[11] Louw QA, Morris LD, Grimmer-Somers K. The prevalence of low back pain in Africa: a systematic review. BMC Musculoskelet Disord 2007;8:1−14.
[12] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:1−173.
[13] 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;52:377−84.
[14] Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses. Ottawa, Canada: University of Ottawa; 2008.
[15] Khan K, ter Riet G, Glanville J, Sowden A, Kleijnen J. Systematic reviews: CRD's guidance for undertaking reviews in health care. York, UK: Centre for reviews and dissemination; 2001.
[16] Loney PL, Stratford PW. The prevalence of low back pain in adults: a methodological review of the literature. Phys Ther 1999;79:384−96.
[17] Effective Public Health Practice Project. Quality assessment tool for quantitative studies. Hamilton, Canada: McMaster University; 2007.
[18] Feinstein A. Principles of medical statistics. Boca Raton, FL: Chapman and Hall/CRC; 2002.
[19] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159−74.
[20] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213−20.
[21] Statacorp. STATA 10.1. TX.2009. http://www.stata.com/.
[22] Terracciano L, Brozek J, Compalati E, Schunemann H. GRADE system: new paradigm. Curr Opin Allergy Clin Immunol 2010; 10(4):377−83.
[23] von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147: 573−7.
[24] Tooth L, Ware R, Bain C, Purdie DM, Dobson A. Quality of reporting of observational longitudinal research. Am J Epidemiol 2005;161: 280−8.
[25] Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: a systematic critical review of the literature. Eur Spine J 2006;15(6):834−48.
[26] Maxwell L, Santesso N, Tugwell PS, Wells GA, Judd M, Buchbinder R. Method guidelines for Cochrane Musculoskeletal Group systematic reviews. J Rheumatol 2006;33:2304−11.
[27] Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283: 2008−12.
[28] Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, et al. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. J Clin Epidemiol 2011;64: 637−57.