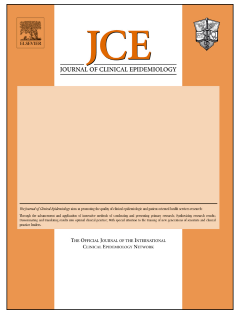
Accepted Manuscript



The methodological and reporting quality of systematic reviews from China and the USA are similar

Jinhui Tian, PhD, Jun Zhang, MD, Long Ge, MD, Kehu Yang, MD, Fujian Song, PhD

PII: S0895-4356(16)30816-2

DOI: 10.1016/j.jclinepi.2016.12.004

Reference: JCE 9290

To appear in: *Journal of Clinical Epidemiology*

Received Date: 30 June 2016

Revised Date: 29 November 2016

Accepted Date: 9 December 2016

Please cite this article as: Tian J, Zhang J, Ge L, Yang K, Song F, The methodological and reporting quality of systematic reviews from China and the USA are similar, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2016.12.004.

The methodological and reporting quality of systematic reviews from China and the USA are similar

Jinhui Tian[1,2], PhD; Jun Zhang[3], MD; Long Ge[1,2], MD; Kehu Yang[1,2*], MD; Fujian Song[4*], PhD.

1. Evidence-based Medicine Center of Lanzhou University, Lanzhou, China
2. Key Laboratory of Evidence-based Medicine and Knowledge Translation of Gansu Province, Lanzhou, China
3. School of Nursing, Gansu University of Chinese Medicine, Lanzhou, China.
4. Norwich Medical School, University of East Anglia, Norwich, U.K.

*** Correspondence to:**

Fujian Song, Norwich Medical School, University of East Anglia, Norwich, NR4 7TJ, U.K. Email: Fujian.song@uea.ac.uk

Kehu Yang, Evidence-Based Medicine Center, Lanzhou University, 199 Donggang West Road, 730000 Lanzhou, Gansu, P.R. China. E-mail：yangkh@lzu.edu.cn

Word count for text: 3728

Word count for abstract: 197

Abstract

Objective: To compare the methodological and reporting quality of systematic reviews by authors from China and those from the United States (the USA).

Study Design: From systematic reviews of randomised trials published in 2014 in English, we randomly selected 100 from China and 100 from the USA. The methodological quality was assessed using the AMSTAR tool, and reporting quality assessed using the PRISMA tool.

Results: Compared with systematic reviews from the USA, those from China were more likely to be a meta-analysis, published in low impact journals, and a non-Cochrane review. The mean summary AMSTAR score was 6.7 (95% confidence interval: 6.5 to 7.0) for reviews from China and 6.6 (6.1 to 7.1) for reviews from the USA, and the mean summary PRISMA score was 21.2 (20.7 to 21.6) for reviews from China and 20.6 (19.9 to 21.3) for reviews from the USA. The differences in summary quality scores between China and the USA were statistically non-significant after adjusting for multiple review factors.

Conclusions: The overall methodological and reporting quality of systematic reviews by authors from China are similar to those from the USA, although the quality of systematic reviews from both countries could be further improved.

Keywords: systematic review; methodological quality; reporting quality; risk of bias; validity; evidence based medicine

What is new?

- **Key findings**: The overall methodological and reporting quality of systematic reviews of randomised trials by authors from China were similar to those from the USA. The differences and similarities in specific quality items between China and the USA were identified.
- **What this adds to what is known**: This is the first study to compare the reporting and methodological quality of systematic reviews of randomised trials by authors from China (a developing country) and the USA (a developed country).
- **What is the implication, what should change now**: Considering the usefulness of systematic reviews in evidence based practice and the development of primary research, the systematic reviewing capacity should be strengthened in China. Identified shortcomings in methodological and reporting quality of published systematic reviews should be considered in further training of authors of systematic reviews in the relevant countries.

Background

Well conducted systematic reviews and meta-analyses of randomized controlled trials (RCTs) provide the most valid research evidence on effects of healthcare interventions [1,2]. Systematic review methods (with or without meta-analysis) have been used in medicine and health research since later 1980s in developed countries [3]. The Cochrane Collaboration and other evidence-based health programmes have promoted the use of systematic reviewing methods globally [4], including China [5].

It has been anticipated that systematic reviews would help address challenges due to rapid increase in clinical literature [6,7]. However, the successful production of systematic reviews during past decades has raised concerns about whether the exponential increase in published systematic reviews might have actually exacerbated information overload [7-11]. Particularly, the increased production of systematic reviews by authors from China has been considered at least partly responsible for the rapid increase in systematic reviews globally [12,13]. For example, a search of PubMed on January 8th 2016 (see Supplementary file-1 for the search strategy) found that the number of published systematic reviews by authors from China was increased exponentially from only 19 in 2005 to 1073 in 2014. During the same time period, the production of systematic reviews by authors from the United States (USA) was only moderately increased from 500 in 2005 to 796 in 2014.

With the rapid increase in the number of systematic reviews by authors from China, their reporting and methodological quality have been scrutinized in previous studies [12,14-18]. These studies usually suggested that the reporting and methodological quality of systematic reviews from China were poor and needed to be much improved. However, it is unclear about the quality of systematic reviews by authors from China relative to those from other countries. There was only one previous study that compared meta-analyses of genetic associations by authors from China and those from the USA [10,12]. According to our knowledge, there were no published studies that systematically compared quality of systematic reviews of RCTS of healthcare interventions by authors from China and those by authors from other countries.

Identification of differences in methodological and reporting quality of systematic reviews by authors from China and developed countries may help appropriately interpret findings from systematic reviews, and set priorities in training of systematic reviewers. Specifically, we consider it appropriate to compare systematic reviews by authors from China and those from the USA for the following reasons: Authors from the USA, along with authors from other high-income nations, have been traditionally the main producer of systematic reviews, and a previous study had compared genetic association meta-analyses by authors from China and the USA [12]. Therefore, the aim of the current study is to compare the main characteristics, methodological and reporting quality of systematic reviews of healthcare interventions between China and the USA. Although the reporting quality was assessed, the focus of the current study was on the methodological quality regarding the validity in the process and results of a systematic review.

Methods:

Identification and selection of systematic reviews

One reviewer (FS) searched PubMed on January 8th 2016 to identify relevant systematic reviews (see Supplementary file-1 for the search strategy). Citations of all identified systematic reviews were downloaded to an EndNote database, and then exported to a Microsoft Excel spreadsheet. Each of the originally identified records by country was assigned a random number from 0 to 1 (generated by Excel). Then the records were ordered from the smallest to the largest by assigned random numbers, and the first 100 eligible systematic reviews from each country were selected. If a selected systematic review was not eligible, a successive record was used to replace it until the total number of included systematic reviews was 100 for each country. Included systematic reviews met the following criteria: (1) It was a review article and explicitly stated as a systematic review or meta-analysis, with a formal (comprehensive or not) literature search, (2) was fully published in English in 2014, (3) included only RCTs, and (4) had a corresponding author with an affiliation in mainland China or in the USA. We did not formally calculate the number of systematic reviews required, because of no information on what would be clinically meaningful differences in quality of systematic reviews between countries.

Quality assessment and data extraction

All authors involved in this study had previous experience of assessing quality of published systematic reviews. Using a data extraction sheet (Supplementary file-1), one reviewer (IZ, LG, or JHT) extracted and a second reviewer (JHT or FS) checked data on the main characteristics from included systematic reviews. Any disagreements were resolved by discussion. Data extracted from systematic reviews included: the journal in which a systematic review was published, type of systematic reviews (narrative or meta-analysis), the number of authors, countries which co-authors came from, whether the review protocol was registered, diseases of interest, interventions evaluated, primary outcome measures, the number of RCTs included, the number of total participants, and conclusions of the systematic reviews. Impact factors of journals in which systematic reviews were published were retrieved by searching 2014 Journal Citation Reports® (Thomson Reuters, 2016) in Web of Science.

The methodological quality of a systematic review reflects risk of bias or validity in its process and results. Previous studies found that the AMSTAR (Assessing the Methodological Quality of Systematic Reviews) tool is reliable and valid [19-21]. Therefore, we used AMSTAR tool to assess the methodological quality of the included systematic reviews. The reporting quality of a systematic review refers to the clarity and transparency of its reporting, and poor reporting reduces the value and usefulness of systematic reviews [22]. We used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) checklist to assess the reporting quality of the included systematic reviews [23]. The methodological quality assessment using the AMSTAR tool was conducted by two independent reviewers (JHT and FS), and the reporting quality assessment using the PRISMA tool was conducted by one reviewer (JZ or LG) and checked by a second reviewer (JHT or FS). Any disagreements between reviewers were resolved by discussion or the involvement of a third reviewer.

To examine the agreement between the two independent reviewers in the assessment of methodological quality of systematic reviews, we calculated the agreement proportion and Cohen's kappa value for each of the 11 AMSTAR items.

Data analysis methods

The main characteristics and quality of systematic reviews from China versus those from the USA were tabulated. We compared the quality of systematic reviews between China and the USA by individual items of the AMSTAR and PRISMA instrument. We calculated a summary AMSTAR score for each systematic review according to the method used by Shea and colleagues [20]: For each of the 11 items of the AMSTAR checklist, it was scored '1' if the answer was 'Yes', and '0' if the answer was 'No' or 'Can't tell'. Some items may be relevant mainly to meta-analysis, such as the use of funnel plot and related statistical test for assessing publication bias. It was scored '1' if a narrative discussion of risk of publication bias was available in systematic reviews when the use of funnel plot was impossible or inappropriate. The summary AMSTAR score for a systematic review was calculated by counting the number of 'Yes' answers, with a possible maximum score of 11. For the assessment of reporting quality of systematic reviews, each of the 27 PRISMA items was scored '1' for full compliance, '0.5' for partial compliance, and '0' for non-compliance [15]. The summary PRISMA score for a systematic review was calculated by adding up scores assigned to each item, with a maximum score of 27.

Chi-squared test was used for differences in proportions (or Fisher's exact test if a contingency table contained a cell with 5 or fewer events). Two sample Wilcoxon rank-sum test or Student t test were used for differences in continuously distributed variables. The summary AMSTAR scores and the summary PRISMA scores were ranked into three groups: low (up to the 25th percentile), moderate (the inter-quartile range), and high (the 75th percentile and above). The association between the summary quality scores and country was calculated in either bivariate or multiple variable linear regression analyses after adjusting for factors with imbalanced distribution between China and the USA. Analyses were conducted by using data from all included systematic reviews and using data from only non-Cochrane systematic reviews. Statistical significance was defined as two sided $P<0.05$. We used Stata/IC® version 14.1 for statistical analyses.

Results

The search of PubMed on January 8th 2016 identified 1073 records of systematic reviews published in 2014 from China and 796 from the USA. The references and main characteristics of the randomly selected systematic reviews are available in Supplementary file-2.

Characteristics of included systematic reviews

The main characteristics of the included systematic reviews are summarised in Table 1. Compared with systematic reviews from the USA, those from China were more likely to contain a quantitative meta-analysis, less likely to be a Cochrane systematic review, and tended to be published in journals with lower impact factors. Systematic reviews from China were more likely to have four or more co-authors, but much less likely to include co-authors from other countries. The proportion of systematic reviews with a registered review protocol was lower for systematic reviews from China, although the difference was no longer significant after excluding Cochrane systematic reviews.

>> **insert Table 1. The main characteristics of the included systematic reviews**<<

There were statistically significant differences in disease conditions investigated. Compared with reviews from the USA, reviews from China were more likely to investigate cancer or tumour diseases, and less likely to study mental or behavioural disorders. In terms of interventions evaluated, systematic reviews from China focused less on pharmacological interventions, and more on surgical interventions and alternative medicine. Systematic reviews from China provided somewhat more significant or positive conclusions and less uncertain conclusions, although the overall difference was statistically non-significant.

Results of methodological quality

Two reviewers independently assessed the methodological quality and proportions of agreement on the initial AMSTAR assessment were greater than 65% for ten of the 11 AMSTAR items (see supplementary file-3). The low agreement (28.5%) on the score for conflict of interests was mainly due to different understanding of definition for this item. All disagreements were resolved by discussion, and results of the assessment of the methodological quality are shown in Table 2.

Differences between systematic reviews from China and the USA were statistically significant for seven of the 11 AMSTAR items. The methodological quality of systematic reviews from one country was not consistently lower or higher than another country for all AMSTAR items. Compared with systematic reviews from the USA, the methodological quality of reviews from China was relatively poor in terms of a priori design, comprehensive literature search, listing both included and excluded studies, and stating sources of support in both the review and included primary studies. On the contrary, systematic reviews from China showed better quality in terms of duplicate study selection, duplicate data extraction, the assessment of scientific quality, and using scientific quality in formulating conclusions. After excluding Cochrane systematic reviews, differences in the methodological quality remained unchanged in general (Table 2).

>> insert Table 2. The methodological quality of systematic reviews by country – results of the AMSTAR checklist assessment<<

The difference in the mean summary AMSTAR score between China and the USA was statistically non-significant using data from all included reviews, but statistically significant after excluding Cochrane reviews (Table 3). The included reviews were ranked into three groups according to summary AMSTAR scores (low, moderate or high), and the proportions of systematic reviews belonging to each of the groups are shown in Figure 1. The difference in proportions between the two countries was statistically significant. Compared with systematic reviews from the USA, those from China were less likely to have a low summary AMSTAR score and more likely to have a moderate summary AMSTAR score. Reviews from China were less likely to have a high AMSTAR score by using data from all systematic reviews, although it was no longer the case after excluding Cochrane reviews (Figure 1).

>>insert Table 3. The summary PRISMA and AMSTAR scores by country <<

>> insert Figure 1. The summary AMSTAR score by country <<

Results of reporting quality

Table 4 shows proportions of systematic reviews with total compliance for each of the 27 PRISMA items. Considering all the included systematic reviews, the differences between China and the USA were statistically significant for 11 of the 27 PRISMA items. Compared with systematic reviews from the USA, those from China had a lower total compliance in reporting of review protocols, study selection methods, additional analysis methods, and funding sources. On the contrary, systematic reviews from China had a higher proportion of total compliance in reporting of titles, eligibility criteria, methods for assessing risk of bias within studies, results of risk of bias within studies, results of individual studies, results of evidence synthesis, and discussion of conclusions. After excluding Cochrane reviews, differences in reporting of titles and protocol registration between countries were no longer statistically significant, while differences in structured abstract and reporting of risk of bias across studies became statistically significant (Table 4).

>>insert Table 4. The reporting quality of systematic reviews by country – results of the PRISMA checklist assessment<<

The difference in the mean summary PRISMA score was non-significant using data from all included reviews, and was statistically significant after excluding Cochrane reviews (Table 3). Figure 2 shows the proportions of systematic reviews stratified into three groups according to summary PRISMA scores (low, moderate or high). Using data from all included systematic reviews, the difference in the proportion between the two countries was statistically non-significant. After excluding Cochrane reviews, the overall difference was statistically significant, indicating that systematic reviews from China were less likely to have a low PRISMA score, and more likely to have a high PRISMA score, compared with those form the USA (Figure 2).

>> insert **Figure 2. The summary PRISMA score by country<<**

Results of regression analyses

Using data from all included systematic reviews, the summary AMSTAR scores were not statistically significantly associated with country in either bivariate or multiple variable linear regression analyses after adjusting for factors with imbalanced distribution between China and the USA (Supplementary file-4). After excluding Cochrane reviews, the summary AMSTAR scores were statistically significantly associated with country in bivariate analysis, although it was no longer significant after adjusting for other factors. Similarly, the summary PRISMA scores were not significantly associated with country when all systematic reviews were included in regression analyses. For non-Cochrane reviews, the association between the summary PRISMA score and country was statistically significant in bivariate analysis, and the association became statistically non-significant in multiple variable analysis (Supplementary file-4).

Discussion

There were significant differences in characteristics of systematic reviews between the two countries, regarding the use of meta-analysis, being a Cochrane review, impact factor of journals in which they were published, and co-authors from multiple countries. The overall differences in the methodological and reporting quality of systematic reviews between China and the USA were not statistically significant after adjusting for multiple review characteristics.

Of the included systematic reviews, eight from China and 26 from the USA were Cochrane systematic reviews. The quality of Cochrane systematic reviews was better than that of non-Cochrane reviews in the current study, which is consistent with findings from previous studies [13,24]. Cochrane systematic reviews do not use ‘systematic review’ or meta-analysis’ terms in titles, and all are required to register their protocols. After excluding Cochrane reviews, differences between the two countries in the reporting of titles and pre-defined protocols were no long statistically significant.

Using the AMSTAR and PRISMA criteria as the "gold standard", systematic reviews from either China or the USA need to be further improved, as systematic reviews from any other countries. For example, systematic reviews from different countries published in 2014 often failed to provide important aspects of review methods, did not search for unpublished studies, and used inappropriate statistical methods [13]. The current study found that systematic reviews from China had poor methodological quality in terms of a priori design, listing of excluded studies, and stating sources of support in both the review and included primary studies. Only 5.4% of non-Cochrane systematic reviews from China (versus 14.9% from the USA) adequately assessed the conflict of interests in primary studies included.

Compared with systematic reviews from the USA, those from China had relatively better methodological quality in duplicate study selection and in duplicate data extraction. The number of authors of systematic reviews from China was on average larger than that from the USA, and the sufficient manpower is necessary to carry out duplicate study selection and data extraction. In addition, systematic reviews from China had better quality in terms of the assessment of scientific quality of studies, use of quality assessment in formulating conclusions, and assessment of publication bias. However, there are still considerable rooms for further improvement by authors from China in these items. For example, 20.6% of the included systematic reviews from China did not appropriately incorporate the scientific quality of the included studies in formulating conclusions, and 37.0% did not assess the risk of publication bias. Even there were no significant differences between the two countries, further improvement is also required. For example, literature search was not sufficiently comprehensive in 23.9%, and the status of publication (such as grey literature) was not used as an inclusion criterion in as high as 79.3% of systematic reviews from China.

Therefore, appropriate training should be provided to authors of systematic reviews in China to avoid or reduce the methodological shortcomings identified in this study. It should be emphasized that the improvement in methodological quality is also relevant to authors from the USA, and likely to be relevant to systematic reviews by authors from any countries [13].

The validity and quality of findings from primary research conducted in China have been assessed in some previous studies. For example, controlled trials of acupuncture in China reported more positive results than those from England, possibly due to publication bias [25]. Another study found that the reporting quality and validity of RCTs in China was low, compared with “gold standard” trials reported in European and North American journals [26]. More recently, Yao and colleagues reported that quality of evidence included in meta-analyses published in Chinese language was lower than that in Cochrane systematic reviews [18]. However, it is important to distinguish the conceptual difference between the quality of primary research and the quality of systematic reviews. Irrespective of quality of primary research studies, high quality systematic reviews can be conducted to correctly indicate the credibility of the available evidence.

Primary research in China, as in other low and middle income countries (LMICs), has been rather limited in quantity and quality. For example, 78% of RCTs of interventions for major NCDs recruited patients in high-income countries, and risk of bias was higher in RCTs from LMICs [27]. Clinical and public health practice in China (as in other LMICs) will currently have to be based on research evidence mainly from high income countries. Evidence based health policy and clinical guidelines in China need sufficient capacity of systematic reviewing to borrow research evidence from other countries. In addition, the improved capacity in conducting systematic reviews may also facilitate more relevant and valid primary research in China [28]. Therefore, we should celebrate the success of international Cochrane Collaboration and other evidence-based medicine efforts to increase the number and to improve the quality of systematic reviews globally during the past two decades. The concern recently raised about the redundant publication of systematic reviews [11,29,30] should be resolved by rigorous peer reviewing and editorial process [8,31].

Strengths and limitations

According to our knowledge, this is the first study to compare the methodological and reporting quality of systematic reviews from China and the USA. Recent systematic reviews of RCTs on health care interventions from the two countries were randomly selected without restriction about medical field or type of interventions, so that the results would be widely generalizable and reflecting the present circumstances. Consequently, the included systematic reviews were diverse in terms of disease conditions, interventions evaluated, and other review characteristics. Studies in future may consider to compare the quality of reviews from different countries on the same topic in terms of patients and interventions evaluated. We used regression analyses to adjust for multiple review characteristics in the comparison of the quality of systematic reviews between the two countries. The results of multiple variable analyses should be interpreted with caution because of the possible multi-collinearity between independent variables.

Only the assessment of methodological quality using the AMSTAR checklist was conducted by two independent reviewers in this study, and the assessment of PRISMA reporting quality was conducted by one reviewer and checked by a second reviewer. In addition, the assessment of methodological quality of systematic reviews was based on what was reported by authors, and the actual conduct might be different. We reported results of the methodological and reporting quality of systematic reviews by checklist items, and as the summary quality scores. Although the use of the summary AMSTAR score for assessing the methodological quality of systematic reviews was validated in previous studies [20], the PRISMA checklist was not originally designed as a scored instrument [23], and further studies are required to assess the validity of the summary PRISMA score for the reporting quality of published systematic reviews. As in a previous study [20], we calculated and presented the mean AMSTAR and PRISMA summary scores in the current study. However, further studies are required to explore the appropriate statistical methods for estimating an average value of the quality scores of multiple systematic reviews.

The current study included only systematic reviews from China and the USA, and assessed only systematic reviews of RCTs and published in English. Further studies are required to compare the quality of systematic reviews between other countries, published in different languages, and included observational studies. Another limitation of the current study is that the representativeness of the randomly selected systematic reviews was not assessed. The number of the included systematic reviews was based on the available time and other resources, and sample size required was not formally calculated because of no information on the meaningful difference in reporting or methodological quality of systematic reviews between countries. It may be interesting to note that the current study included a total of 100 systematic reviews from each of the two countries, twice more than the number of meta-analyses (n=50) from each of the two countries included in a previous study [12].

Conclusions

The overall methodological and reporting quality of systematic reviews by authors from China were similar to those from the USA, although the quality of systematic reviews from both countries could be further improved. Identified shortcomings in methodological and reporting quality of published systematic reviews should be taken into consideration in further training of authors of systematic reviews in the relevant countries.

Acknowledgement

Authors' contribution: KHY and FS conceived initial concept. FS and JHT developed review protocol. FS conducted literature and study selection. JHT, JZ, LG, and FS extracted and checked data from the included systematic reviews, and assessed reporting and methodological quality of systematic reviews. FS performed data analysis and drafted the manuscript. All authors commented on the manuscript.

Funding: There was no specific funding for this study.

Conflict of interest: None to declare.

References

1 Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997; 126:376-380.

2 Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA* 2014; 312:171-179.

3 Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. *Eval Health Prof* 2002; 25:12-37.

4 Summerskill W. Cochrane Collaboration and the evolution of evidence. *Lancet* 2005; 366:1760.

5 Zhang M, Li Y. Cochrane Collaboration in China 1996-2013. *J Evid Based Med* 2014; 7:22-25.

6 Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions? *ACP J Club* 2003; 139:A11.

7 Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med* 2010; 7:e1000326.

8 Fuhr U, Hellmich M. Channeling the flood of meta-analyses. *Eur J Clin Pharmacol* 2015; 71:645-647.

9 Tebala GD. What is the future of biomedical research? *Med Hypotheses* 2015; 85:488-490.

10 Ioannidis JP. Can too many systematic reviews and meta-analyses do harm? Plenary II: Information overload: Are we part of the problem or part of the solution? . *The 24th Cochrane Colloquium*. Vienna, 5th Oct 2015

11 Riaz IB, Khan MS, Riaz H, Goldberg RJ. Disorganized Systematic Reviews and Meta-analyses: Time to Systematize the Conduct and Publication of These Study Overviews? *Am J Med* 2016; 129:339 e311-338.

12 Ioannidis JP, Chang CQ, Lam TK, Schully SD, Khoury MJ. The geometric increase in meta-analyses from China in the genomic era. *PLoS One* 2013; 8:e65602.

13 Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, et al. Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Med* 2016; 13:e1002028.

14 Jin YH, Ma ET, Gao WJ, Hua W, Dou HY. Reporting and methodological quality of systematic reviews or meta-analyses in nursing field in China. *Int J Nurs Pract* 2014; 20:70-78.

15 Li JL, Ge L, Ma JC, Zeng QL, Yao L, et al. Quality of reporting of systematic reviews published in "evidence-based" Chinese journals. *Syst Rev* 2014; 3:58.

16 Ma B, Guo J, Qi G, Li H, Peng J, et al. Epidemiology, quality and reporting characteristics of systematic reviews of traditional Chinese medicine interventions published in Chinese journals. *PLoS One* 2011; 6:e20185.

17 Ma B, Qi GQ, Lin XT, Wang T, Chen ZM, et al. Epidemiology, quality, and reporting characteristics of systematic reviews of acupuncture interventions published in Chinese journals. *J Altern Complement Med* 2012; 18:813-817.

18 Yao L, Sun R, Chen YL, Wang Q, Wei D, et al. The quality of evidence in Chinese meta-analyses needs to be improved. *J Clin Epidemiol* 2016; 74:73-79.

19 Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007; 7:10.

20 Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009; 62:1013-1020.

21 Pieper D, Buechter RB, Li L, Prediger B, Eikermann M. Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties. *J Clin Epidemiol* 2015; 68:574-583.

22 Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med* 2007; 4:e78.

23 Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol* 2009; 62:e1-34.

24 Petticrew M, Wilson P, Wright K, Song F. Quality of Cochrane reviews. Quality of Cochrane reviews is better than that of non-Cochrane reviews. *BMJ* 2002; 324:545.

25 Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Control Clin Trials* 1998; 19:159-166.

26 Zhang D, Freemantle N, Cheng KK. Are randomized trials conducted in China or India biased? A comparative empirical analysis. *J Clin Epidemiol* 2011; 64:90-95.

27 Fan H, Song F. An assessment of randomized controlled trials (RCTs) for non-communicable diseases (NCDs): more and higher quality research is required in less developed countries. *Sci Rep* 2015; 5:13221.

28 Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, et al. How to increase value and reduce waste when research priorities are set. *Lancet* 2014; 383:156-165.

29 Bolland MJ, Grey A. A case study of discordant overlapping meta-analyses: vitamin d supplements and fracture. *PLoS One* 2014; 9:e115934.

30 Helfer B, Prosser A, Samara MT, Geddes JR, Cipriani A, et al. Recent meta-analyses neglect previous systematic reviews and meta-analyses about the same topic: a systematic examination. *BMC Med* 2015; 13:82.

31 D'Agostino RB, Jr. Editorial Commentary: Meta-analyses Are on the Rise, but Are We Learning More? *Arthroscopy* 2016; 32:538-539.

Supplementary material

1: Outline protocol

2: Table -Included systematic reviews and main characteristics

3: Agreement between reviewers in the AMSTAR assessment

4: Results of linear regression analyses

Table 1. Dependence of the statistical significance on sample size (n_1, n_2) with a constant effect difference of 15-10=5 score points.

Verum VAS Difference			Placebo VAS Difference			Borenstein's	SMD			t-test
n_1	m_1	s_1	n_2	m_2	s_2	J	SMD	95%	CI	p
10	15	10	10	10	10	0.9577	0.479	-0.434	1.392	0.285
20	15	10	20	10	10	0.9801	0.490	-0.147	1.127	0.128
30	15	10	30	10	10	0.9870	0.494	-0.024	1.011	0.061
32	15	10	32	10	10	0.9879	0.494	-0.007	0.995	0.053
33	15	10	33	10	10	0.9882	0.494	0.001	0.987	<0.050
34	15	10	34	10	10	0.9886	0.494	0.008	0.980	0.046
50	15	10	50	10	10	0.9923	0.496	0.096	0.896	0.016
100	15	10	100	10	10	0.9962	0.498	0.216	0.780	0.001

Legend: VAS: visual analogue scale of pain (scale: 0-100 score points), difference between baseline and follow-up, n=sample size, m=mean, s=standard deviation, J=correction factor by Borenstein, SMD=standardized mean difference, 95% CI: 95% confidence interval, p=statistical significance (type I error) that SMD is different from zero.

Table 2. Data from the evaluation study (11): knee osteoarthritis before inpatient rehabilitation (baseline) and 3 months later (follow-up) on the WOMAC pain scale

Transition item		**WOMAC pain: difference baseline to follow-up**	
Pain at follow-up was:	**n**	**m**	**s**
much better	19	31.58	17.63
slightly better	49	13.51	21.58
about the same	62	4.77	15.62
slightly worse	44	-1.32	20.81
much worse	16	-3.50	17.03
All: score difference	190	7.60	21.13
All: baseline score	190	50.93	21.48

Legend: n=number of subjects, m=mean, s=standard deviation, both for the score differences (baseline to follow-up). WOMAC pain scaling: 0=maximal pain, 100=no pain. A positive difference reflects pain relief and vice versa; m=mean, s=standard deviation.

Table 3: MCID for improvement on the WOMAC pain scale: absolute, relative, and effect sizes

Method	**Numerator**	**Denominator**	**MCID**	**95% CI**		**p**	**Comment**
Jaeschke (8)	13.51	-	**13.51**	7.25	19.77	<0.001	Score difference of the "slightly better" group
Mean change meth. Redelmeier (9)	13.51-4.77	-	**8.74**	1.73	15.74	0.015	Score difference of the "slightly better" group minus that of the "about the same" group
% baseline score	8.74	50.93	**17.15%**	6.86%	27.39%	0.015	Mean change in % of the baseline score
% total score	8.74	100.00	**8.74%**	3.49%	13.95%	0.015	Mean change in % of the maximal score
ES, Kazis (17)	8.74	21.48	**0.407**	0.024	0.789	0.038	Mean change divided by the standard deviation of the group's baseline score
SRM, Liang (18)	8.74	21.13	**0.413**	0.031	0.796	0.035	Mean change divided by the standard deviation of the group's score differences
SMD, Borenstein (7)	8.74*0.993	18.48	**0.469**	0.092	0.847	0.016	Mean change*J divided by the pooled standard deviation of the two transition group's score differences

Legend: MCID: minimal clinically important difference scaled in score points (scale: 0-100), ES=effect size (according to Kazis), SRM=standardized response mean (according to Liang), SMD=standardized mean difference (according to Borenstein); ES, SRM, SMD: dimensionless (scaled by number of standard deviations). 95% CI: 95% confidence interval, p=type I error of the test that the MCID is different from zero.

Table 4: MCID for improvement on the WOMAC pain scale: ROC and regression methods

Method	Dependent variable	Independent variables	MCID	95%	CI	p	Comment
ROC	transition	Δ pain	**15.00**	8.74	21.26	<0.001	Area under ROC: 0.637 (95% CI: 0.528-0.747), sensitivity=0.531, specificity=0.871
Linear regression: bivariate	Δ pain	transition	**8.74**	1.73	15.74	0.015	same result as by the mean change method
Linear regression: mulitvariate	Δ pain	transition, sex, age, WOMAC pain baseline score	**7.09**	0.93	13.25	0.024	adjusted for the added potential confounders (independent variables)
Logistic regression: bivariate	transition	Δ pain	**OR: 1.026** (beta: 0.0261, se=0.0110)	1.004	1.049	0.018	Odds ratio: probability of being “slightly better” for 1.00 point pain relief
Logistic regression:	transition	Δ pain, sex, age, WOMAC pain	**OR: 1.029** (beta: 0.0286,	1.004	1.055	0.025	Odds ratio, adjusted for the added potential confounders (independent

multivariate	baseline score	se=0.0128)	variables)

Legend: MCID: minimal clinically important difference scaled in score points (scale: 0-100), 95% CI: 95% confidence interval, p=type I error of the test that the MCID is larger than zero. ROC: receiver operation characteristic curve. Δ pain: WOMAC pain score difference baseline to follow-up. Transition item response: 0=about the same, 1=slightly better. Logistic regression: OR=odds ratio, beta=regression coefficient for Δ pain, se=standard error.

Table 5. Application of an a priori evaluated MCID (11) to an RCT (19)

WOMAC pain		**Baseline**		**Follow-up**		**Difference**	**pooled**
	n	m	s	m	s	m	s
Intervention	36	59.6	15.8	71.4	15.8	11.8	15.8
Placebo	35	60.2	17.0	60.4	21.6	0.2	19.4
total	71	59.9	16.4	66.0	18.9	6.1	17.7
		Δ	**s**	**Parameter**	**95%**	**CI**	**p**
Empiric SMD		(11.8-0.2)*0.989	17.7	**0.649**	0.168	1.129	0.008
MCID as SMD		8.74*0.989	17.7	**0.489**	0.013	0.964	0.044
MCID as SRM		8.74	17.7	**0.494**	0.013	0.975	0.044
MCID as ES		8.74	16.4	**0.533**	0.051	1.015	0.031
Logistic regr.		exp(0.0261*11.6)	(se=0.0110)	**OR=1.354**	1.049	1.746	0.021

Legend: WOMAC pain: 0=maximal pain, 100=no pain. n=number of patients, m=mean, s=standard deviation, Δ: relevant difference of score differences, MCID: minimal clinically important difference (positively scaled to reflect improvement), ES: effect size according to

Kazis, SRM: standardized response mean according to Liang, SMD: standardized mean difference according to Borenstein, 95% CI: 95% confidence interval, p: type I error of the test that the effect size is different from zero. OR=odds ratio, se=standard error.

Figure 1. The summary AMSTAR score by country

Note to Figure 1: The difference in proportions between China and the USA was statistically significant (p=0.016 for all systematic reviews, and p=0.007 for non-Cochrane systematic reviews)

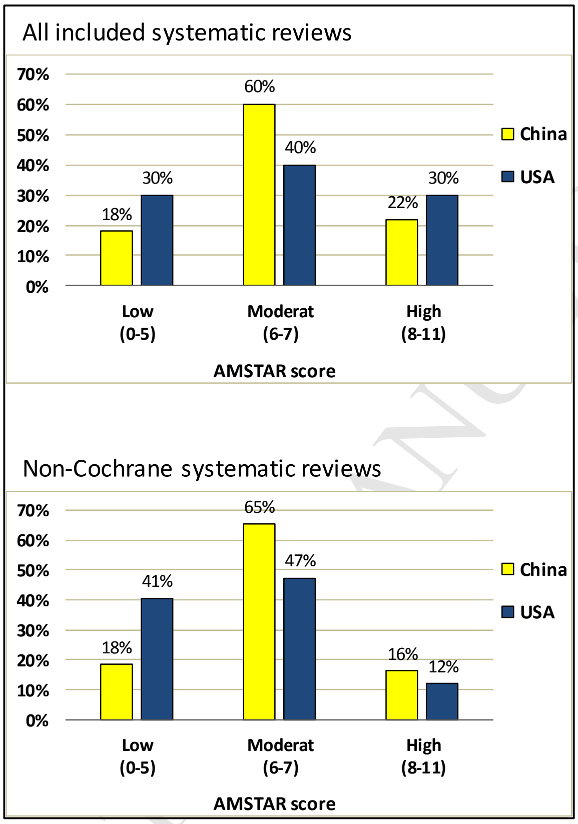
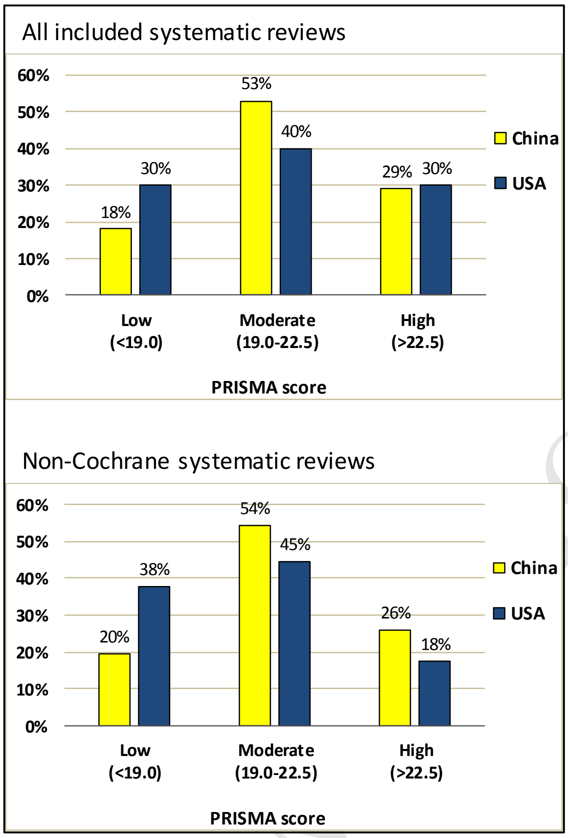


Figure 2. The summary PRISMA score by country

Note to Figure 2: The difference between China and the USA was statistically non-significant for all systematic reviews (p=0.089) and statistically significant for non-Cochrane reviews (p=0.029).



All included systematic reviews
60%
50%
40%
30%
20%
10%
0%
18%
30%
53%
40%
29%
30%
China
USA
Low
(<19.0)
Moderate
(19.0-22.5)
High
(>22.5)
PRISMA score
Non-Cochrane systematic reviews
60%
50%
40%
30%
20%
10%
0%
20%
38%
54%
45%
26%
18%
China
USA
Low
(<19.0)
Moderate
(19.0-22.5)
High
(>22.5)
PRISMA score