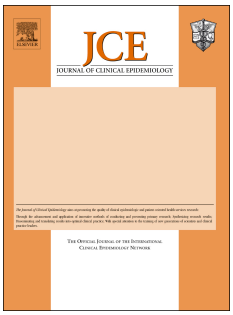
Accepted Manuscript



Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES): development of a new tool for systematic reviews

L. Susan Wieland, Brian M. Berman, Douglas G. Altman, Jürgen Barth, Lex M. Bouter, Christopher R. D'Adamo, Klaus Linde, David Moher, C. Daniel Mullins, Shaun Treweek, Sean Tunis, Danielle A. van der Windt, Merrick Zwarenstein, Claudia Witt

PII: S0895-4356(16)30375-4

DOI: 10.1016/j.jclinepi.2017.01.010

Reference: JCE 9320

To appear in: *Journal of Clinical Epidemiology*

Received Date: 27 August 2016

Revised Date: 20 December 2016

Accepted Date: 21 January 2017

Please cite this article as: Wieland LS, Berman BM, Altman DG, Barth J, Bouter LM, D'Adamo CR, Linde K, Moher D, Mullins CD, Treweek S, Tunis S, van der Windt DA, Zwarenstein M, Witt C, Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES): development of a new tool for systematic reviews, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2017.01.010.

Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES): development of a new tool for systematic reviews

L. Susan Wieland[1]*, Brian M. Berman[1], Douglas G. Altman[2], Jürgen Barth[3], Lex M. Bouter[4,5], Christopher R. D'Adamo[1], Klaus Linde[6], David Moher[7], C. Daniel Mullins[8], Shaun Treweek[9], Sean Tunis[10], Danielle A. van der Windt[11], Merrick Zwarenstein[12], Claudia Witt[1,3,13]

1. Center for Integrative Medicine, University of Maryland School of Medicine, Baltimore MD, USA

2. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK.

3. Institute for Complementary and Integrative Medicine, University Hospital Zurich, University of Zurich, Zurich, Switzerland

4. Department of Epidemiology & Biostatistics, VU University Medical Center, Amsterdam, The Netherlands

5.Department of Philosophy, Faculty of Humanities, Vrije Universiteit, Amsterdam, The Netherlands

6. Institute of General Practice, Technical University Munich, Munich, Germany

7. Clinical Epidemiology Program, Ottawa Hospital Research Institute, School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada

8. Pharmaceutical Health Services Research Department, University of Maryland School of Pharmacy, Baltimore, MD, USA

9. Health Services Research Unit, University of Aberdeen, Aberdeen, UK.

10. Center for Medical Technology Policy, Baltimore MD, USA

11. Arthritis Research UK Primary Care Center, Research Institute for Primary Care & Health Sciences, Keele University, Keele, Staffordshire, UK

12. Centre for Studies in Family Medicine, Schulich School of Medicine & Dentistry, Western University, Western Centre for Public Health and Family Medicine, London, Canada

13. Institute for Social Medicine Epidemiology and Health Economics, Charité Universitätsmedizin Berlin, Berlin, Germany

Abstract

Background: Randomized trials may be designed to provide evidence more strongly related to efficacy or effectiveness of an intervention. When systematic reviews are used to inform clinical or policy decisions, it is important to know the efficacy-effectiveness nature of the included trials.

Objective: To develop a tool to characterize randomized trials included in a systematic review on an efficacy-effectiveness continuum.

Methods: We extracted rating domains and descriptors from existing tools, and used a modified Delphi procedure to condense the domains and develop a new tool. The feasibility and inter-rater reliability of the tool was tested on trials from 4 systematic reviews.

Results: The RITES (Rating of Included Trials on the Efficacy-effectiveness Spectrum) tool rates clinical trials on a 5-point Likert scale in four domains: (1) participant characteristics, (2) trial setting, (3) flexibility of interventions, and (4) clinical relevance of interventions. When RITES was piloted on trials from 3 reviews by unaffiliated raters, ratings were variable (Intraclass Correlation Coefficient 0.25-0.66 for the four domains), but when RITES was used on 1 review by the review authors with expertise on the topic the ratings were consistent (ICCs >0.80.

Conclusion: RITES may help to characterize the efficacy-effectiveness nature of trials included in systematic reviews.

Running title: RITES: A new tool for rating trials in systematic reviews

Keywords: Comparative Effectiveness Research; Systematic reviews; Randomized controlled trials; Pragmatic trial; Explanatory trial; Effectiveness; Efficacy; Applicability

Introduction

Randomized controlled trials (RCTs) are often characterized as designed with either a more explanatory or a more pragmatic approach [1]. RCTs taking an *explanatory* design approach determine whether an intervention produces the expected result under ideal research circumstances and are intended to provide evidence on the *efficacy* of an intervention: Does the treatment work in an optimal setting under standardized conditions? RCTs taking a *pragmatic* design approach measure the degree of beneficial effect under "real world" clinical conditions and are intended to provide evidence on the *effectiveness* of an intervention: Does the treatment work in the usual care setting under realistic conditions? The design of RCTs is generally not either fully explanatory or completely pragmatic but rather placed along a continuum between the two, where this continuum may vary for different aspects of the trial design and conduct. The Pragmatic-Explanatory Continuum Indicatory Summary (PRECIS, later modified to PRECIS-2) is a tool which was developed to help designers of RCTs make decisions regarding 10 trial domains in accordance with explanatory versus pragmatic design goals [2, 3]. Similarly, the evidence provided by a trial may be situated along an efficacy-effectiveness continuum. We use the terms *explanatory* and *pragmatic* when we address the trials and their design, and we use the terms *efficacy* and *effectiveness* when we address the evidence provided by a RCT.

To understand whether a RCT is potentially useful to inform clinical decision making in usual care (i.e., the setting and type of care routinely received by patients with the condition) it is important to know if the study provides evidence about the efficacy or the effectiveness of an intervention. Evidence about efficacy may be obtained from a carefully controlled experimental

comparison (e.g., between an active drug and a placebo, or in a highly selected (homogenous) group of participants). Evidence about effectiveness may be obtained from comparisons between clinically relevant interventions carried out in settings and participants that are representative of usual care. In the first scenario the trials provide evidence about efficacy, which may provide important information on the specific effects of an intervention when deployed under optimally controlled conditions. In the second scenario, the trials may be susceptible to some forms of bias (e.g. information bias, due to difficulty in blinding the comparison between two clinically relevant interventions), but they provide evidence to inform decision-making in usual care. Understanding whether the trials included in a systematic review describe the efficacy or the effectiveness of a treatment will help readers, including clinicians and health policy decision makers, understand whether the review provides information that is more relevant to the specific actions of the intervention under assessment circumstances or information that may be more directly applicable to real-world implementation.

Researchers have previously used PRECIS (or adaptations of PRECIS) to retrospectively characterize ongoing or completed trials along the efficacy-effectiveness continuum and thus describe the nature of the reported evidence [4-9]. However, PRECIS and PRECIS-2 were developed to inform choices during the trial design phase, rather than to assess the characteristics of trial evidence retrospectively from the publication of the trial. They assume detailed familiarity with available design options at the time that the trial is being designed, and this information may not be available in the report of a completed trial. In addition, PRECIS-2 assesses nine trial domains which may limit the practicality for use on the often substantial number of trials included in a systematic review. A tool for use with systematic reviews should

be short and focused on the essential elements of the efficacy-effectiveness spectrum that are likely to be described in a trial report. We are not aware of any short, practical tools that have been systematically designed and validated specifically for characterizing completed trials along an efficacy-effectiveness continuum for retrospective use in systematic reviews [4]. Our aim was therefore to identify all available tools for evaluating the efficacy-effectiveness of trials, to extract the concepts from these tools, and to develop a short and feasible tool that informs decision-makers reading systematic reviews about whether the evidence provided by the included trials is information about the efficacy or the effectiveness of an intervention.

Methods

Searching for existing tools

We began by searching the literature for existing tools used to measure the pragmatic-explanatory or efficacy-effectiveness characteristics of RCTs. One author (LSW) searched PubMed and Web of Science in March 2014 for tools or adaptations of tools used to classify trials along this pragmatic-explanatory or efficacy-effectiveness continuum (Search strategy in Box 1).

Developing the tool

We used a modified Delphi procedure that included 2 rounds working with two expert groups. A core expert group (the14 authors of this publication), consisting of scientists affiliated with our

Center and leading experts in comparative effectiveness research or in methods for conducting or reporting randomized trials and systematic reviews, developed the tool. Our Delphi expert panel was a larger group that included stakeholders in comparative effectiveness research, trial methodology, and reporting methods, who were invited to take part in the Delphi procedure and provide feedback. We collected names of these experts from lists of authors of publications using tools to measure the pragmatic-explanatory or effectiveness-efficacy characteristics of trials, authors of CONSORT (Consolidated Standards of Reporting Trials) statements or extensions, authors of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) statements, and members of the Practihc (Pragmatic Randomized Controlled Trials in HealthCare) group, the Cochrane Applicability and Recommendations Methods Group, and the Cochrane Complementary Medicine Field.

Delphi round 1: We extracted information from tools and adaptations of tools found in the literature and constructed a table of the domains and criteria (Table 1) for which a RCT could be characterized as having a more explanatory-efficacy or a more pragmatic-effectiveness orientation, together with descriptors of maximum explanatory-efficacy or maximum pragmatic-effectiveness characteristics for each criterion. We consulted with the core expert group to ensure that all relevant domains were included in this table. We then e-mailed a group of researchers we had identified as appropriate participants for a Delphi expert panel. We sent them the table of domains and criteria, and invited them to reply to an online survey on the perceived importance (on a scale of 1-10 with 1=not at all important and 10=extremely important) and ease of rating (on a scale of 1-10 with 1=not at all easy and 10=extremely easy) for each domain. We also asked the panel to rate the ease of using (on a scale of 1-10 with 1=not at all easy and

10=extremely easy) a rating scheme based on three categories (efficacy, between efficacy and effectiveness, effectiveness), or on a 5, 10 or 20-point numerical rating scale.

We circulated the feedback from the first Delphi round to our core expert group, and held a teleconference and had written exchanges to decide upon the most relevant domains, revised the domain descriptors, and selected the rating scheme for an initial draft of the tool.

In the second round of the Delphi procedure we repeated the e-mailed invitation and this time presented the initial draft of the tool and asked the panel to rate the clarity of an introductory explanation of the tool concepts on a scale of 1-10 with 1=not at all clear and 10=extremely clear. We also requested respondents to rate how confident they would be in rating each domain from a completed publication on a scale of 1-10 with 1=not at all confident and 10=extremely confident. Finally, we asked whether transforming the summarized results of the 5-point rating scale into percentages would be a clear way to present the results. Respondents were asked to rate the clarity of such a transformation on a scale of 1-10 with 1=not at all clear and 10=extremely clear. We brought the results of the second online survey to our core expert group during a second teleconference and series of written exchanges, and finalized the tool.

Pilot testing

We tested the feasibility and inter-rater reliability of the tool by applying it to four Cochrane reviews. We selected the trials included in a Cochrane review on artichoke leaf extract for treating hypercholesterolaemia (3 trials) [10], a Cochrane review on yoga for epilepsy (2 trials)

[11], and a Cochrane review being prepared for acupuncture in the treatment of hip osteoarthritis (5 trials) (unpublished update of a previous review [12]). Our goal was to test the tool on a range of different populations, interventions, and comparators, and we therefore used multiple small reviews from our list of Cochrane reviews related to complementary medicine. We asked 12 researchers, 7 of whom (JB, LB, CD, ST, DvdW, LSW, CW) were members of the core expert group in the tool development. The remaining five had experience in systematic reviews but were 'naïve' to the tool, to independently carry out ratings of each trial such that 10 raters rated each domain for each trial. We also wanted to test the tool with authors of an ongoing systematic review, and for this purpose we asked the two authors of an update to a Cochrane review on acupuncture for migraine to rate each of 8 trials in the updated review [13]. One of the authors (KL) was involved in the tool development, but the other author was 'naïve'. All raters were requested to give comments on anything that was difficult or unclear about the ratings process. We then observed the range of ratings and calculated intraclass correlation coefficients [14] using SAS version 9.3 (SAS Institute Inc., Cary, North Carolina) to determine inter-rater reliability for each domain of the tool for each trial. The results were discussed in a conference call with the core expert group and followed up by written exchanges to finalize the descriptors of the tool domains.

Developing visual representation of the ratings

We subsequently developed a visual representation of the ratings of effectiveness/efficacy of individual trials within a systematic review. This visual representation was presented to the core experts in a conference call and discussed in written exchanges until consensus was reached.

Results

Results of searches

The searches retrieved 1,880 citations after de-duplication. They were screened by two authors (LSW, CW) for descriptions of tools or revisions of tools for measuring RCTs on a pragmatic-explanatory or efficacy-effectiveness continuum. One new citation describing the planned revision of PRECIS (PRECIS-2) [15] was identified, and several citations mentioning that PRECIS was used to assess ongoing or completed trials. Aside from the citation describing the plans for PRECIS-2, we did not identify any other publications describing the existence of tools or modifications of tools that we had not already identified during the planning for this project. The tools or measures that we had already identified were: Gartlehner et al.'s simple tool to distinguish between efficacy and effectiveness studies [16], PRECIS [2], and four modifications or adapted uses of PRECIS, including Koppenaal et al.'s adaptation of PRECIS for systematic reviews [5], Tosh et al.'s Pragmascope [8], Selby et al.'s enhancement of PRECIS [7], and Glasgow et al.'s application of PRECIS to effectiveness trials [6]. None of these instruments was developed to rate published trials retrospectively. We included these tools together with the description of the development of PRECIS-2 [15] and later the final report for PRECIS-2 [3] as resources for the initial identification of efficacy-effectiveness criteria for the tool.

Results of the Delphi rounds

There were 72 respondents (20.45% response rate) out of 352 persons invited to the first Delphi round. Respondents classified each of the criteria in Table 1 as at least moderately important for rating the efficacy-effectiveness of an RCT, with medians ranging from 7-9 on a scale of 1-10 where higher numbers reflect greater importance. However, the confidence with which individual criteria could be rated was lower for some items, ranging from 5-9 on a scale of 1-10 where higher numbers reflect greater confidence. Respondents also noted that some criteria were already assessed and reported elsewhere in a systematic review (e.g., blinding of participants and personnel), some criteria would likely require special clinical or other expertise to rate, and several criteria could probably be combined. With regard to the preferred rating scheme, the respondents reported that rating of each aspect of trial design from 1 (efficacy) to 5 (effectiveness) would be easier than rating each aspect from 1 to 10, from 1 to 20, or in three categories (efficacy, between efficacy and effectiveness, effectiveness). (See Table 2 for the results of the first Delphi round).

There were 69 respondents (19.62% response rate) out of 352 persons invited to the second Delphi round. The respondents found the proposed introductory text to the draft tool to be clear (mean (sd) of 7.6 (2.3) and median of 8 on a scale of 1-10 where higher numbers reflect greater clarity) and were fairly confident that each of the four items on the draft tool could be rated on the basis of a published report of an the RCT at issue (see the description of each domain of the final RITES tool, below, for details). The display or summarized results using percentages was rated as not very clear (mean (sd) of 6.5 (2.7) and median of 7 on a scale of 1-10 where higher numbers reflect greater clarity). Based on the feedback from the second Delphi round, no modifications were made to the proposed introductory text for the final tool. However, after

discussion, the core experts made some wording changes to the four domains of the draft tool, and added notes defining the meaning of 'usual care' and explaining how participants were to be judged.

Description of the final RITES Tool

The final RITES tool contains 4 domains, which are each rated on a 5-point scale from a strong emphasis on efficacy to a strong emphasis on effectiveness (Table 3). The domains are: participants' characteristics, trial setting, flexibility of intervention(s), and clinical relevance of experimental and comparison intervention(s).

Participants' characteristics

The domain of participants' characteristics corresponds closely to the concept of eligibility in PRECIS and PRECIS-2, although it considers additional factors beyond eligibility criteria, such as whether the participants were actually similar in age, severity of illness, and comorbidities to those participants who would be candidates for the intervention in usual care [2, 3]. In the second Delphi round, the confidence in assessing this domain from a study report was a mean (sd) of 7.3 (1.9) and a median of 8.

Trial setting

The domain of trial setting blends elements of setting and practitioner expertise in PRECIS and organization in PRECIS-2 [2, 3]. In the second survey, the confidence in assessing this domain from a study report was a mean (sd) of 7.4 (2.0) and a median of 8.

Flexibility of intervention(s)

The domain of flexibility of intervention(s) corresponds closely to the combined PRECIS concepts of flexibility in practitioner adherence to study protocol, flexibility in delivery of the intervention, and flexibility in participant adherence to the intervention [2, 15]. In the second survey, the confidence in assessing this domain from a study report was a mean (sd) of 7.7 (1.9) and a median of 8.

Clinical relevance of experimental and comparison intervention(s)

The domain of clinical relevance of experimental and comparison intervention(s) draws from the PRECIS domain of the flexibility of the comparison intervention [2], Gartlehner et al.'s focus on intervention duration [16], and the IOM concept of clinical and policymaker relevance [17]. In the second survey, the confidence in assessing this domain from a study report was a mean (sd) of 7.1 (2.2) and a median of 8.

Rating scale

The evidence from the RCT is rated for each domain along a 5-point Likert Scale ranging from 1= strong emphasis on efficacy to 5 = strong emphasis on effectiveness. When information on this domain is not available, the rating NA may be used. Because the second Delphi round indicated that transforming the ratings into percentages of efficacy or effectiveness was not very clear, the core experts group decided that this transformation was not appropriate.

Results of piloting

The results of RITES piloting are presented in the online Appendix. For five of the ten RCTs from three Cochrane reviews on a range of topics, one or two domains within individual RCTs were rated as reflecting more emphasis on efficacy (e.g., mean (sd) of 1.4 (0.7) and range 1 to 3) or effectiveness (e.g., mean (sd) of 4.3 (0.7) and range 3 to 5). However, most of the ratings were inconsistent, with a large amount of variability between ratings and many cases in which the ratings spanned the entire spectrum of possible ratings on a dimension (i.e., 1 to 5). The intraclass correlation coefficients were low to moderate for each of the four tool dimensions, ranging from a low of 0.23 for trial setting to a high of 0.45 for clinical relevance. When three outlying raters with ratings ± 3 standard deviations from the mean were removed from the dataset, most correlations were substantially higher, ranging from 0.25 for flexibility of interventions to 0.66 for clinical relevance. There was no obvious pattern of differences between the more expert and the 'naïve' raters. In some cases raters stated that it was difficult to come up with a rating on a domain when one aspect of the domain tended to reflect a more efficacy orientation and other aspects of the domain tended to reflect a more effectiveness orientation. However, raters primarily commented on difficulties in ratings due to lack of available information in the study report, or due to lack of clinical expertise in the medical condition or knowledge of the interventions.

In contrast, when two review authors rated trials included in their updated Cochrane review of acupuncture for migraine, the ratings were highly consistent. Although one author was involved in development of this tool (KL) and one author was 'naïve' to the tool, the two raters frequently agreed on ratings and the ratings were never more than one point apart (e.g., 2 and 3, or 4 and 5),

although in some instances one of the raters felt that there was insufficient information in the published report to make a rating (Appendix). Five of the RCTs were rated as showing a greater emphasis on efficacy across all domains, two of the RCTs were rated as showing a greater emphasis on effectiveness across all domains, and one RCT was rated as possessing a stronger emphasis on efficacy in two domains, and as slightly more on the efficacy side in the third domain and slightly more on the effectiveness side in the fourth domain. The intraclass correlation coefficients were 0.8 or higher for each domain.

Results of visual representation

The visual representation (Figure 1) displays the averages of the ratings by the two raters for each domain for each trial in the acupuncture for migraine review. The ratings do not take into account sample size of the trials, or weighting of the RCTs within one or more meta-analyses. Only RCTs for which both authors rated information as available on a domain are included in the visual for that domain. In the case of the this review, the rated studies showed a lot of heterogeneity across the spectrum. The visual clearly shows that 2 of the studies (numbers 7 and 8) tended to be more on the effectiveness side in all domains, while the other studies for which there was available information tended to be more on the efficacy side.

Discussion

We developed the RITES tool to provide authors and readers of systematic reviews with a rating scale to assess whether the included RCTs provide predominantly information about the efficacy

or effectiveness of an intervention. We have taken multiple criteria for effectiveness and efficacy from PRECIS and other sources, and consolidated the criteria into four key domains in which randomized trials may display information that has a greater emphasis on effectiveness or on efficacy. Unlike PRECIS, which was developed to assist in planning trials, RITES was systematically designed to be suitable for assessing completed trials from a study report. Our next steps will include development of processes for incorporating RITES into the conduct and reporting of a systematic review.

The ratings of effectiveness and efficacy are not intended to reflect trial quality. Pragmatic trials (which produce effectiveness information) and explanatory trials (which produce efficacy information) can each be of higher or lower quality and consequently differ in their risk of bias, depending upon how they are designed and carried out. Ratings of risk of bias are expected to be carried out separately for trials within a review, and although these ratings may in some cases be associated with aspects of efficacy or effectiveness such that it is easier for an explanatory study that tests efficacy to earn a low risk of bias score, we do not seek to establish or explore any such relationships here. However, it may be promising to include an examination of this relationship in future studies in order to get an impression of the construct validity of RITES.

A possible limitation of this study was the low response rate (approximately 20%) to our invitations to participate on the Delphi panels. We invited a wide range of participants to the Delphi panel, in order to receive input from stakeholders in comparative effectiveness research, systematic reviews, trial methodology, and research reporting. It is possible that most of the people we invited did not have sufficient interest or expertise in the intersections of effectiveness

research, systematic reviews, and research methods, to consider participating. We did not collect information on the characteristics of the participants in the first Delphi panel. However, we asked respondents to the second Delphi panel for their expertise with systematic reviews and their prior knowledge of PRECIS; 46/69 (67%) responded that they were regularly involved in producing systematic reviews and 39/69 (57%) that they were familiar with PRECIS, of whom 19/39 (49%) had used PRECIS, indicating that we were successful in soliciting input from a range of users and producers of research.

The variability of the results from our pilot suggests that RITES will perform best when it is used as intended, by authors rating the trials included in their systematic review. Our piloting on eight trials of acupuncture for migraine has shown that the tool may be successfully used in systematic reviews of RCTs to identify domains in which information is more about efficacy or more about effectiveness.

A common difficulty in carrying out the pilot ratings of the ten trials on a range of topics was the lack of background knowledge of the raters regarding the patient population, the setting, and the intervention. Several raters commented during piloting that they had insufficient clinical knowledge to carry out the RITES rating with confidence. This problem of uncertainty was not seen in the pilot ratings of the eight acupuncture for migraine trials, probably because review authors are likely to have more familiarity with the clinical population and the relevant treatment options and delivery characteristics. However, we should consider whether there might be a need to develop guidance for review authors who have limited clinical expertise, and incorporate this guidance in training resources developed for raters.

Even if raters are familiar with the clinical condition, they may still encounter difficulties when rating RCTs that were carried out in very different environments, such as older trials or trials carried out in other countries. For example, if the usual care setting for treatment is primary care, but in previous decades or particular regions of the world the usual care setting is instead secondary or tertiary care, the raters must be able to recognize this. Furthermore, it is unclear how to apply the assessments of efficacy or effectiveness from these 'other' environments to the clinical setting for which the review needs to be informative. This situation was not encountered in our pilot ratings, but piloting with a range of trials carried out in different eras and different geographical regions will likely present opportunities to develop methods to deal with these situations.

Finally, work needs to be done with the RITES tool to identify and develop ways in which the ratings may be of practical use to users of the systematic review. A visual representation of the status of individual trials across domains, similar to the one we developed for the acupuncture for migraine ratings, would likely be part of this development. An indication of the importance and interpretation of measurement differences in ratings (e.g., of one unit in the domain rating) would also be important. This tool, which is greatly indebted to previous work elucidating the concepts of the pragmatic-explanatory approaches for clinical trial design [2, 3], is an opportunity to clarify the importance and relevance of effectiveness and efficacy approaches within the context of the systematic review. We welcome suggestions and collaboration in the further refinement of the tool and its application.

Contributions of authors:

Conception and development of project: BB, CD, LSW, CW

Design of tool: DGA, JB, BB, LMB, CD, KL, DM, CDM, STr, STu, DV, LSW, CW, MZ

Piloting of tool: JB, LMB, CD, KL, ST, DV, LSW, CW

Statistical analyses: CD

Drafting of manuscript: LSW, CW

Final approval of manuscript: DGA, JB, BB, LMB, CD, KL, DM, CDM, STr, STu, DV, LSW, CW, MZ

Acknowledgements:

Participation in online Delphi survey(s): Edgardo Abalos; Samuel A. Abariga MD, MS, DTMH, PhD-c; Matthew J. Bair, MD, MS; Christopher Beardall, DC, L.Ac.; Alan Bensoussan; Jesse Berlin; Jane M Blazeby; Jean-Pierre Boissel; Heather Boon; Cynthia Boyd ; Prof Melanie Calvert; Timothy S Carey MD MPH; Jin Chen; Yaolong Chen; Ke Cheng; Cindy Crawford; Stephen Evans ; Yutong Fei; Ronald M. Glick, MD; Carol M. Greco, Ph.D.; François Gueyffier; Dr. Xinfeng Guo; Patrick Hanaway, MD; Maria Hondras; Daren Heyland; Gudula Kirtschig; Jos Kleijnen; Fredi Kronenberg; Miranda Langendam; Dana J. Lawrence, DC, MMedEd, MA; Myeong Soo Lee; Prof. P.C. Leung; George Lewith; Jianping Liu; Kathleen N. Lohr; Carl Lombard; Rainer Lüdtke; Martha Lucas, Ph.D., L.Ac.; Hugh MacPherson; Ram Manohar; Jair Mari; Joerg J Meerpohl; Carmen Moga; Jane Nadel; Arya Nielsen, PhD; Andrew Nunn; Karen Pilkington; Dr Kavita Prasad; P. Ravaud; Marco Romoli; Janet Lynn Roseman, Ph.D.; Lee Hullender Rubin, DAOM, LAc, FABORM; Margaret Sampson; Howard Schachter PhD; Alexander G. Schauss, PhD; Rosa N Schnyer; Peter Selby; Nimrod Sheinman; Xueyong Shen;

Byung-Cheul Shin; João Bosco Guerreiro da Silva; Mark Speechley, Western U, London Canada; Mark Spigt; Frances Stewart, M.D.; Julie Taylor; Jennifer Marie Tetzlaff; Karine Toupin-April; Kiichiro Tsutani; Ruth Taylor-Piliae, PhD, RN, FAHA; Peter Tugwell; Andrew Vickers; Vasiliy Vlassov; Sunita Vohra; Laura Weeks, PhD Ottawa Integrative Cancer Centre; Vivian Welch; Prof Hywel C. Williams; Hans Wohlmuth; Professor Charlie Xue; Hongbo Yuan

Pilot ratings of RITES: Opeyemi Babatunde, Joie Ensor, Jo Jordan, Alain Mayhew, Michael Mehring, Kimberly Yang

Funding: This work was supported by the National Institutes of Health National Center for Complementary and Integrative Health (R24 AT001293)

Box 1. Search strategies

PubMed search strategy:

("Randomized Controlled Trials as Topic" OR "Clinical Trials as Topic") AND (tool[tiab] OR spectrum[tiab] OR continuum[tiab]) AND (pragmatic[tiab] OR explanatory[tiab] OR efficacy[tiab] OR effectiveness[tiab])

Web of Science searches:

All citations of the main publications on PRECIS [2, 18] or the explanatory publication for the CONSORT extension to pragmatic trials [19].

Table 1. Table of domains previously used to characterize trials along the effectiveness-efficacy continuum

Domain	**Criterion***	**Descriptor for trial with maximum *effectiveness* orientation**	**Descriptor for trial with maximum *efficacy* orientation**
Participants	Eligibility [15]	How similar are the participants to those who would receive the intervention if it was part of usual care?	There are many exclusions (e.g. those who don't comply, respond to treatment, or are not at high risk for primary outcome, are children or elderly), or the trial uses many selection tests not used in usual care.
	Recruitment [15]	Recruitment is through usual appointments or clinic	Recruitment is through targeted information letters, advertising in newspapers, radio plus incentives and other routes that would not be used in usual care.
Setting and Interventions	Setting [15]	The setting is identical to usual care.	The setting is only a single center or only specialized trial or academic centers.
	Organization [15]	The resources, provider expertise and the organization of care delivery in the intervention arm are identical to usual care.	The trial increases staff levels, gives additional training, requires more than the usual experience or certification, or certification and increase resources.
	Comparison intervention practitioner expertise [2]	The comparison intervention typically is applied by the full range of practitioners, and in the full range of clinical interest, regardless of their expertise, with only ordinary attention to their training, experience, and performance.	In an explanatory trial, practitioner expertise in applying the comparison intervention(s) is standardized so as to maximize the chances of detecting whatever comparative benefits the experimental intervention might have.
	Flexibility (delivery) [15]	Flexibility in the intervention is identical to that in usual care.	Intervention delivery is less flexible than usual care. There is a strict protocol, monitoring and measures to improve compliance, with specific advice on allowed co-interventions and complications.
	Flexibility of the comparison intervention [2]	In a pragmatic trial, ''Usual practice'' or the best available alternative management strategy, offering practitioners considerable leeway in deciding how to apply it.	In an explanatory trial, restricted flexibility of the comparison intervention.
Other Design Aspects	Comparison relevance to a	The comparison has the objective of directly informing a specific clinical decision from the	The comparison is not relevant to a specific patient or policymaker decision.

	specific patient or policymaker decision [17]	patient perspective or a health policy decision from the population perspective.	
	Clinical relevance of experimental and comparison intervention [17]	Both the experimental and comparison intervention have the potential to be 'best practice'.	One or both of the experimental and comparison interventions is not a clinically relevant treatment (e.g., placebo, sub-clinical doses).
	Study duration [16]	The duration of the study should mimic a minimum length of treatment in a clinical setting to allow the assessment of health outcomes	The duration of the study is shorter than the minimum length of treatment in a clinical setting.
	Blinding of participants and personnel to the intervention [20, 21]	Clinician and patient biases are not necessarily viewed as detrimental but accepted as part of physicians' and patients' responses to treatment and included in the overall assessment.	Participants and investigators blinded where possible to minimize bias
	Follow-up [15]	No more than the follow-up expected in usual care.	Compared to usual care, more frequent or longer visits, unscheduled visits triggered by primary outcome event or intervening event, and more extensive data collection.
	Primary outcome [15]	The primary outcome is of obvious importance to participants.	The primary outcome uses a surrogate, physiological outcome, central adjudication or uses assessment expertise that is not available in usual care, or the outcome is measured at an earlier time than in usual care.
	Flexibility (adherence) [15]	There is no more than usual encouragement to adhere to the intervention.	Exclusion based on adherence and measures to improve adherence if found wanting.
	Practitioner adherence to study protocol [2]	There is unobtrusive (or no) measurement of practitioner adherence and no special strategies to maintain or improve it are used.	There is close monitoring of how well the participating clinicians and centers are adhering to even the minute details in the trial protocol and ‘‘manual of procedures.’’
	Primary analysis [15]	The most pragmatic approach is to use intention to treat with all available data.	The most explanatory analysis is one that excludes ineligible post-randomization participants and includes only those following the treatment protocol.

	Analysis at the population and subgroup level [16, 17]	Subgroup analyses to discern the effects in subjects with common clinical characteristics. Study is adequately powered to detect minimally important differences in important outcomes for important subgroups.	No preplanned subgroup analyses.

*Each criterion is associated with one or more references to a rating tool or other source for the concept.

Table 2 Results from the Delphi survey round 1 (n= 72 responders)

Domain	Criterion	Importance*			Rating confidence*		
		n responses/ n missing	Mean (sd)	Median (range)	n responses/ n missing	Mean (sd)	Median (range)
Participants	Eligibility	70/2	8.8 (1.9)	9 (1-10)	69/3	6.8 (1.9)	7 (3-10)
	Recruitment	69/3	6.4 (2.4)	7 (1-10)	69/3	6.5 (2.4)	7 (1-10)
Setting and Interventions	Setting	69/3	7.4 (2.1)	8 (2-10)	69/3	7.2 (2.0)	8 (2-10)
	Organization	68/4	7.4 (2.2)	8 (2-10)	69/3	6.0 (2.1)	6 (2-10)
	Comparison intervention practitioner expertise	69/3	7.9 (1.8)	8 (2-10)	69/3	6.0 (2.2)	6 (1-10)
	Flexibility (delivery)	69/3	8.3 (1.9)	9 (1-10)	69/3	6.7 (2.3)	7 (1-10)
	Flexibility of the comparison intervention	68/4	7.8 (2.1)	8 (2-10)	68/4	6.2 (2.2)	6 (1-10)
Other Design Aspects	Comparison relevance to a specific patient or policymaker decision	67/5	6.9 (2.3)	7 (2-10)	67/5	6.0 (2.3)	5 (2-10)
	Clinical relevance of experimental and comparison intervention	68/4	8.1 (2.3)	9 (1-10)	68/4	7.5 (2.0)	8 (1-10)
	Study duration	69/3	7.8 (2.2)	8 (1-10)	70/2	7.7 (2.2)	8 (1-10)
	Blinding of participants and personnel to the intervention	70/2	7.6 (2.4)	8 (1-10)	70/2	7.5 (2.4)	8 (1-10)
	Follow-up	70/2	7.3 (2.2)	8 (1-10)	69/3	7.3 (2.2)	8 (1-10)
	Primary outcome	69/3	8.0 (2.5)	9 (1-10)	69/3	8.0 (2.0)	9 (1-10)
	Flexibility (adherence)	70/2	7.9 (2.2)	8 (1-10)	69/3	6.6 (2.1)	7 (1-10)
	Practitioner adherence to study protocol	70/2	7.3 (2.2)	8 (1-10)	70/2	5.5 (2.6)	6 (1-10)
	Primary analysis	70/2	8.1 (2.4)	9 (1-10)	70/2	7.6 (2.2)	8 (1-10)
	Analysis at the population and subgroup level	70/2	6.5 (2.9)	7 (1-10)	69/3	6.6 (2.5)	7 (1-10)

* Ratings made on a scale from 1-10, where 1 is least importance or confidence, and 10 is maximum importance or confidence.

Table 3: RITES (Rating Included Trials on the Efficacy-Effectiveness Spectrum) tool introductory text, criteria, and rating scale

Trials are often characterized as designed with either a more explanatory or a more pragmatic approach. Trials taking an explanatory design approach determine whether an intervention produces the expected result under ideal circumstances and are intended to provide evidence on the efficacy of an intervention. Trials taking a pragmatic design approach measure the degree of beneficial effect under "real world" clinical settings and are intended to provide evidence on the effectiveness of an intervention. A trial design is often not completely on either the explanatory or pragmatic side but rather along a continuum between the two and the placement of the trial along this continuum may vary for different aspects of the trial design and conduct. Similarly the evidence provided by a trial is placed within an efficacy-effectiveness continuum. We use the terms explanatory and pragmatic when we address the trials and their design. We use the terms efficacy and effectiveness when we address the evidence provided by a trial.

Criteria	**Descriptor for *efficacy* orientation**	**Descriptor for maximum *effectiveness* orientation**
1. Participants characteristics *	The participants are a homogeneous population and are markedly different from those seen in usual care. Participants may be deliberately selected to comply with treatment, respond to treatment, or demonstrate the efficacy of the experimental intervention (e.g., be at high risk for the primary outcome). There may be other exclusions that would not be seen in usual care** (e.g. exclusion of participants with comorbidities).	The participants are representative of the population who would receive the experimental intervention if it was part of usual care**. They are similar in age, severity of illness, and comorbidities to those patients who would be candidates for the intervention in a usual care** setting. They reflect diversity along parameters that could impact adherence.
2. Trial setting	The setting is selected to maximize the ability to carry out the trial and identify an intervention effect if there is one. The setting may be more specialized than the setting in which the experimental intervention would be delivered in usual care (e.g. a single	The overall setting of the trial is similar to usual care** and includes diverse sub-settings common in usual care** for this intervention (e.g. primary care, specialized care).

	center, specialized clinics or only specialized trial or academic centers or only providers with high levels of experience).	
3. Flexibility of intervention(s)	Experimental and comparison intervention delivery is less flexible than usual care. There is a strict protocol, monitoring, and measures to improve intervention adherence. There is specific advice on prohibited co-interventions. [15](15)(14)(16)	Flexibility in the experimental and comparison interventions is identical to that in usual care**. Co-interventions may be permitted.
4. Clinical relevance of experimental and comparison intervention(s)	One or both of the experimental and comparison interventions is not a clinically relevant or best current treatment (e.g., placebo, no-treatment control, sub-clinical doses), or study duration is shorter than the minimum length of treatment in usual care**.	Both the experimental and comparison intervention have the potential to be 'best practice'/best current treatment. The duration of the interventions is similar to the minimum length of treatment in usual care**.

*Participants included in the study should be judged according to whether they are representative of the general population of those with the condition of interest who would receive usual care in the geographic area in which the trial is carried out.
**Usual care describes the type of care routinely received by patients in the geographic area in which the trial is carried out. The care provided can vary by patient characteristics (e.g. age, sex, education), the setting (inpatient/outpatient, primary care/specialized care) in which the patient is seen, individual providers and insurance plans.

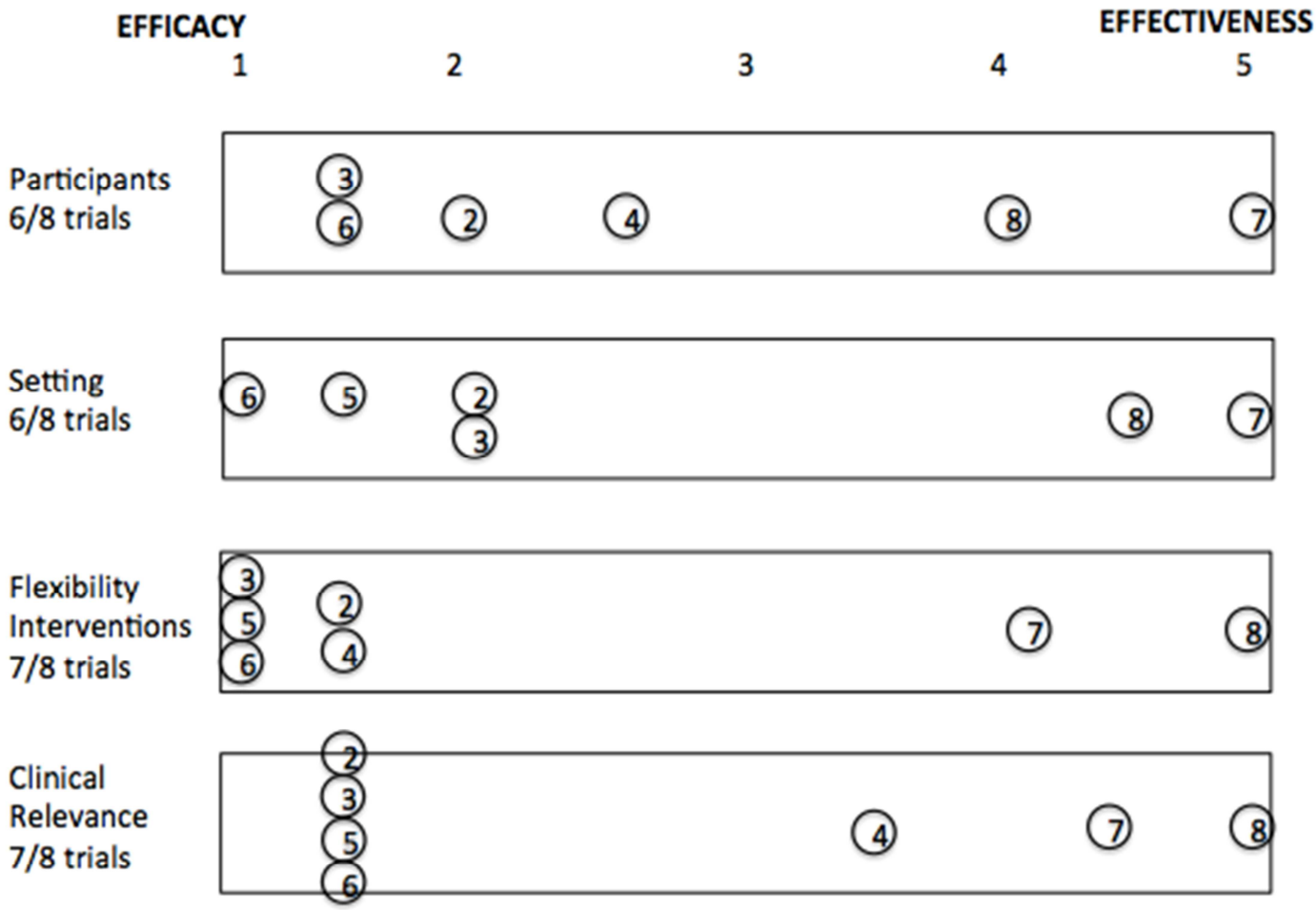
Based upon a 5-point Likert Scale the evidence deriving from each domain should be rated:
1 = strong emphasis on efficacy
2 = rather strong emphasis on efficacy
3 = balanced emphasis on both efficacy and effectiveness
4 = rather strong emphasis on effectiveness
5 = strong emphasis on effectiveness
NA = information not available

References

[1] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. Journal of clinical epidemiology. 2009;62:499-505.
[2] Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. Journal of clinical epidemiology. 2009;62:464-75.
[3] Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. Bmj. 2015;350:h2147.
[4] Witt CM, Manheimer E, Hammerschlag R, Ludtke R, Lao L, Tunis SR, et al. How well do randomized trials inform decision making: systematic review using comparative effectiveness research measures on acupuncture for back pain. PloS one. 2012;7:e32399.
[5] Koppenaal T, Linmans J, Knottnerus JA, Spigt M. Pragmatic vs. explanatory: an adaptation of the PRECIS tool helps to judge the applicability of systematic reviews for daily practice. Journal of clinical epidemiology. 2011;64:1095-101.
[6] Glasgow RE, Gaglio B, Bennett G, Jerome GJ, Yeh HC, Sarwer DB, et al. Applying the PRECIS criteria to describe three effectiveness trials of weight loss in obese patients with comorbid conditions. Health services research. 2012;47:1051-67.
[7] Selby P, Brosky G, Oh PI, Raymond V, Ranger S. How pragmatic or explanatory is the randomized, controlled trial? The application and enhancement of the PRECIS tool to the evaluation of a smoking cessation trial. BMC Med Res Methodol. 2012;12:101.
[8] Tosh G, Soares-Weiser K, Adams CE. Pragmatic vs explanatory trials: the pragmascope tool to help measure differences in protocols of mental health randomized controlled trials. Dialogues in clinical neuroscience. 2011;13:209-15.
[9] Bratton DJ, Nunn AJ, Wojnarowska F, Kirtschig G, Sandell A, Williams HC. The value of the pragmatic-explanatory continuum indicator summary wheel in an ongoing study: the bullous pemphigoid steroids and tetracyclines study. Trials. 2012;13:50.
[10] Wider B, Pittler MH, Thompson-Coon J, Ernst E. Artichoke leaf extract for treating hypercholesterolaemia. The Cochrane database of systematic reviews. 2013;3:CD003335.
[11] Panebianco M, Sridharan K, Ramaratnam S. Yoga for epilepsy. The Cochrane database of systematic reviews. 2015;5:CD001524.
[12] Manheimer E, Cheng K, Linde K, Lao L, Yoo J, Wieland S, et al. Acupuncture for peripheral joint osteoarthritis. The Cochrane database of systematic reviews. 2010:CD001977.
[13] Linde K, Allais G, Brinkhaus B, Fei Y, Mehring M, Vertosick EA, et al. Acupuncture for the prevention of episodic migraine. The Cochrane database of systematic reviews. 2016:CD001218.
[14] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin. 1979;86:420-8.
[15] Loudon K, Zwarenstein M, Sullivan F, Donnan P, Treweek S. Making clinical trials more relevant: improving and validating the PRECIS tool for matching trial design decisions to trial purpose. Trials. 2013;14:115.
[16] Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. A simple and valid tool distinguished efficacy from effectiveness studies. Journal of clinical epidemiology. 2006;59:1040-8.

[17] Institute of Medicine. What is Comparative Effectiveness Research? In: INitial National Priorities for Comparative Effectiveness Research. Washington D.C.2009. p. 29 p.
[18] Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2009;180:E47-57.
[19] Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. Bmj. 2008;337:a2390.
[20] Roland M, Torgerson DJ. What are pragmatic trials? Bmj. 1998;316:285.
[21] Bratton DJ, Nunn AJ. Alternative approaches to tuberculosis treatment evaluation: the role of pragmatic trials. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease. 2011;15:440-6.

Figure 1. Visual presentation of the efficacy-effectiveness of trials within a systematic review



What is new

Key findings

- We developed Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES), a tool to rate the evidence from trials included in systematic reviews along a continuum between maximum efficacy and maximum effectiveness.

What this adds to what was known

- Trials are often characterized as designed to produce information more related to effectiveness or to efficacy. Decision-makers reading systematic reviews may consider it important to understand whether the evidence provided by the included trials is information about the efficacy or the effectiveness of an intervention.
- RITES is the first tool systematically designed specifically for characterizing evidence from completed trials along an efficacy-effectiveness continuum for retrospective use in systematic reviews.

What is the implication, what should change now

- We are developing additional guidance on how to carry out ratings. We are also working on clarifying how the ratings can be of practical use to readers of systematic reviews.

Appendix - Results from Piloting of RITES

A. Ten trials from 3 different reviews, each trial rated by 10 raters

1. Participants

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Artichoke leaf extract for hypercholesterolaemia						
Trial 1	2.7	0.9	1	4	10	0
Trial 2	3.0	0.8	2	4	10	0
Trial 3	1.9	1.5	1	5	9	1
SUMMARY	2.5					
Yoga for epilepsy						
Trial 4	3.7	1.0	2	5	10	0
Trial 5	3.6	1.1	2	5	10	0
SUMMARY	3.6					
Acupuncture for hip osteoarthritis						
Trial 6	2.2	0.6	1	3	10	0
Trial 7	4.0	1.2	1	5	10	0
Trial 8	3.9	0.9	2	5	10	0
Trial 9	3.3	1.1	2	5	10	0
Trial 10	4.1	0.9	2	5	10	0
SUMMARY MEAN	3.5					

2. Trial Setting

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Artichoke leaf extract for hypercholesterolaemia						
Trial 1	2.7	1.4	1	5	10	0
Trial 2	2.9	0.7	2	4	10	0
Trial 3	1.7	1.2	1	3	4	6
SUMMARY	2.4					
Yoga for epilepsy						
Trial 4	3.3	1.6	1	5	10	0
Trial 5	2.4	0.7	2	4	10	0
SUMMARY	2.9					
Acupuncture for hip osteoarthritis						
Trial 6	1.8	1.2	1	4	8	2
Trial 7	3.3	1.4	1	5	10	0
Trial 8	3.3	0.5	3	4	10	0
Trial 9	3.1	1.0	3	4	9	1
Trial 10	4.4	0.5	4	5	10	0
SUMMARY MEAN	3.5					

3. Flexibility of Intervention

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Artichoke leaf extract for hypercholesterolaemia						
Trial 1	2.1	1.4	1	5	10	0
Trial 2	2.0	0.8	1	3	10	0
Trial 3	1.6	0.8	1	3	8	2
SUMMARY	1.9					
Yoga for epilepsy						
Trial 4	2.9	0.8	2	4	9	1
Trial 5	2.5	0.9	1	4	9	1
SUMMARY	2.7					
Acupuncture for hip osteoarthritis						
Trial 6	2.3	0.8	1	4	10	0
Trial 7	2.8	0.8	2	4	10	0
Trial 8	3.2	1.1	2	5	10	0
Trial 9	2.6	1.3	1	4	10	0
Trial 10	4.3	0.7	3	5	10	0
SUMMARY MEAN	3.0					

4. Clinical Relevance

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Artichoke leaf extract for hypercholesterolaemia						
Trial 1	1.8	1.0	1	4	10	0
Trial 2	1.4	0.7	1	3	10	0
Trial 3	1.8	1.4	1	5	10	0
SUMMARY	1.7					
Yoga for epilepsy						
Trial 4	4.2	1.0	2	5	10	0
Trial 5	2.7	1.2	1	5	10	0
SUMMARY	3.4					
Acupuncture for hip osteoarthritis						
Trial 6	2.2	1.3	1	5	10	0
Trial 7	4.0	0.7	3	5	10	0
Trial 8	4.0	1.1	2	5	10	0
Trial 9	2.3	1.7	1	5	10	0
Trial 10	4.3	0.7	3	5	10	0
SUMMARY MEAN	3.4					

Correlation coefficient according to Sprout and Fleiss

RITES	Inter-rater-correlation
Participants	0.25823
Trial Setting	0.23498
Flexibility of Intervention	0.26437
Clinical Relevance	0.44532

B. Eight trials from 1 review, each trial rated by 2 authors of the review

1. Participants

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Acupuncture for migraine						
Trial 1	--	--	2	2	1	1
Trial 2	2	0	2	2	2	0
Trial 3	1.5	0.7	1	2	2	0
Trial 4	2.5	0.7	2	3	2	0
Trial 5	--	--	2	2	1	1
Trial 6	1.5	0.7	1	2	2	0
Trial 7	5	0	5	5	2	0
Trial 8	4	0	4	4	2	0
SUMMARY MEAN	2.75					

2. Trial Setting

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Acupuncture for migraine						
Trial 1	--	--	1	1	1	1
Trial 2	2	0	2	2	2	0
Trial 3	2	0	2	2	2	0
Trial 4	--	--	2	2	1	1
Trial 5	1.5	0.7	1	2	2	0
Trial 6	1	0	1	1	2	0
Trial 7	5	0	5	5	2	0
Trial 8	4.5	0.7	4	5	2	0
SUMMARY MEAN	2.66					

3. Flexibility of Intervention

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Acupuncture for migraine						
Trial 1	--	--	2	2	1	1
Trial 2	1.5	0.7	1	2	2	0
Trial 3	1	0	1	1	2	0
Trial 4	1.5	0.7	1	2	2	0
Trial 5	1	0	1	1	2	0
Trial 6	1	0	1	1	2	0
Trial 7	4	0	4	4	2	0
Trial 8	5	0	5	5	2	0
SUMMARY MEAN	2.14					

4. Clinical Relevance

Trial	Mean	SD	min	max	Number of raters providing ratings	Number of raters unable to rate
Acupuncture for migraine						
Trial 1	--	--	1	1	1	1
Trial 2	1.5	0.7	1	2	2	0
Trial 3	1.5	0.7	1	2	2	0
Trial 4	3.5	0.7	3	4	2	0
Trial 5	1.5	0.7	1	2	2	0
Trial 6	1.5	0.7	1	2	2	0
Trial 7	4.5	0.7	4	5	2	0
Trial 8	5	0	5	5	2	0
SUMMARY MEAN	2.71					

Correlation coefficient according to Sprout and Fleiss

RITES	Inter-rater-correlation
Participants	0.86437
Trial Setting	0.92162
Flexibility of Intervention	0.94225
Clinical Relevance	0.85143