

## Updated Method Guidelines for Cochrane Musculoskeletal Group Systematic Reviews and Metaanalyses

Elizabeth A.T. Ghogomu, Lara J. Maxwell, Rachelle Buchbinder, Tamara Rader, Jordi Pardo Pardo, Renea V. Johnston, Robin D.K. Christensen, Anne W.S. Rutjes, Tania M. Winzenberg, Jasvinder A. Singh, Gustavo Zanolli, George A. Wells, Peter Tugwell and the Editorial Board of the Cochrane Musculoskeletal Group

J Rheumatol 2014;41;194-205  
<http://www.jrheum.org/content/41/2/194>

1. Sign up for our monthly e-table of contents  
<http://www.jrheum.org/cgi/alerts/etoc>
2. Information on Subscriptions  
<http://jrheum.com/subscribe.html>
3. Have us contact your library about access options  
[Refer\\_your\\_library@jrheum.com](mailto:Refer_your_library@jrheum.com)
4. Information on permissions/orders of reprints  
<http://jrheum.com/reprints.html>

*The Journal of Rheumatology* is a monthly international serial edited by Earl D. Silverman featuring research articles on clinical subjects from scientists working in rheumatology and related fields.

# Updated Method Guidelines for Cochrane Musculoskeletal Group Systematic Reviews and Metaanalyses

Elizabeth A.T. Ghogomu, Lara J. Maxwell, Rachelle Buchbinder, Tamara Rader, Jordi Pardo Pardo, Renea V. Johnston, Robin D.K. Christensen, Anne W.S. Rutjes, Tania M. Winzenberg, Jasvinder A. Singh, Gustavo Zanolli, George A. Wells, Peter Tugwell, and the Editorial Board of the Cochrane Musculoskeletal Group

**ABSTRACT.** The Cochrane Musculoskeletal Group (CMSG), one of 53 groups of the not-for-profit, international Cochrane Collaboration, prepares, maintains, and disseminates systematic reviews of treatments for musculoskeletal diseases. It is important that authors conducting CMSG reviews and the readers of our reviews be aware of and use updated, state-of-the-art systematic review methodology. One hundred sixty reviews have been published. Previous method guidelines for systematic reviews of interventions in the musculoskeletal field published in 2006 have been substantially updated to incorporate methodological advances that are mandatory or highly desirable in Cochrane reviews and knowledge translation advances. The methodological advances include new guidance on searching, new risk-of-bias assessment, grading the quality of the evidence, the new Summary of Findings table, and comparative effectiveness using network metaanalysis. Method guidelines specific to musculoskeletal disorders are provided by CMSG editors for various aspects of undertaking a systematic review. These method guidelines will help improve the quality of reporting and ensure high standards of conduct as well as consistency across CMSG reviews. (First Release Dec 1 2013; *J Rheumatol* 2014;41:194–205; doi:10.3899/jrheum.121306)

## Key Indexing Terms:

EVIDENCE-BASED MEDICINE  
ARTHRITIS

SYSTEMATIC REVIEW

METAANALYSIS  
MUSCULOSKELETAL

The aim of the Cochrane Collaboration is to help healthcare providers, patients, patient advocates and carers, and policy makers arrive at well-informed decisions on healthcare treatments by preparing, maintaining, and disseminating methodologically strong systematic reviews<sup>1</sup>.

The systematic review is an essential tool for managing the vast amount of information generated on the etiology, prognosis, incidence/prevalence, diagnosis, prognosis, and treatment of disease. These method guidelines focus on the assessment of treatments (including both benefits and

harms) that are aimed at improving health. Cochrane systematic reviews are increasingly used as the basis for clinical decision support resources such as UpToDate and clinical guidelines [e.g., American College of Rheumatology (ACR) osteoarthritis (OA)<sup>2</sup> and rheumatoid arthritis (RA)<sup>3</sup> guidelines]. Compared to a narrative literature review, the systematic review uses “scientific strategies that are systematic, and designed to limit bias in the assembly, critical appraisal and synthesis of all relevant studies on a specific topic”<sup>4</sup>. Most Cochrane reviews are

*From the Cochrane Musculoskeletal Group (CMSG), Institute of Population Health, University of Ottawa, Ottawa, Ontario, Canada.*

*Supported through a grant from the Canadian Institutes of Health Research.*

*E.A.T. Ghogomu, MD, MSc, Assistant Managing Editor; L.J. Maxwell, MSc, Managing Editor, CMSG, University of Ottawa; R. Buchbinder, MBBS (Hons), MSc, PhD, FRACP, Director, Monash Department of Clinical Epidemiology, Cabrini Hospital, Malvern VIC, Australia, and Professor, Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, and Joint Coordinating Editor, CMSG; T. Rader, MLIS, Knowledge Translation Specialist, CMSG, University of Ottawa; J. Pardo Pardo, BA, Research Coordinator, Centre for Global Health, University of Ottawa; R.V. Johnston, PhD, Managing Editor, CMSG, Monash Department of Clinical Epidemiology, Cabrini Hospital, Monash University; R.D.K. Christensen, BSc, MSc, PhD, Senior Biostatistician, Head of Musculoskeletal Statistics Unit, The Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg, Copenhagen,*

*Denmark; A.W.S. Rutjes, PhD, senior research fellow, Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland; T.M. Winzenberg, MBBS, PhD, FRACGP, MMedSc(ClinEpi), Senior Research Fellow—General Practice at Menzies Research Institute Tasmania, University of Tasmania, Tasmania, Australia; J.A. Singh, MBBS, MPH, Birmingham VA Medical Center, Associate Professor, University of Alabama at Birmingham, Alabama, USA; G. Zanolli, MD, PhD, Casa di Cura SM Maddalena, University of Ferrara Occhiobello, Ferrara, Italy; G.A. Wells, PhD, Professor of Epidemiology and Community Medicine, University of Ottawa; P. Tugwell, MD, MSc, FRCPC, FCAHS, Canada Research Chair, Professor of Medicine, and Epidemiology and Community Medicine, Faculty of Medicine; Principal Scientist, Institute of Population Health, University of Ottawa, Department of Medicine, Ottawa Hospital, Senior Scientist, Ottawa Hospital Research Institute, Joint Coordinating Editor, CMSG.*

*Address correspondence to Dr. E.A.T. Ghogomu, 1 Stewart St., Ottawa, Ontario K1N 6N5, Canada. E-mail: cmsg@uottawa.ca*

*Accepted for publication August 7, 2013.*

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2014. All rights reserved.

quantitative and where appropriate include metaanalysis, but now can include complementary qualitative data to address context and syntheses of how interventions work.

The Cochrane Musculoskeletal Group (CMSG), one of the largest of the 53 international review groups in the Cochrane Collaboration, synthesizes the results of existing studies of appropriate “fit-for-purpose” designs to determine the benefit and harm of interventions for the prevention, treatment, and management of musculoskeletal diseases including various forms of arthritis, soft tissue conditions, and osteoporosis. Conditions specific to the back and musculoskeletal injuries are addressed by the Cochrane Back Group (<http://back.cochrane.org>) and the Cochrane Bone, Joint, and Muscle Trauma Group (<http://bjmtg.cochrane.org>), respectively.

As of Issue 2, 2013, there are 160 completed CMSG reviews and 88 protocols in the Cochrane Database of Systematic Reviews, with its own impact factor of 6.471.

The rigorous, systematic approach used by Cochrane reviews aims to provide a definitive statement on the effects of healthcare treatments. This is useful when there are no other systematic reviews but may also help clarify confusion arising from single studies or systematic reviews with conflicting results.

The 2011 updated Cochrane Handbook for Systematic Reviews of Interventions<sup>5</sup> reflects major methods advances; this has now been complemented by a set of essential and desirable methodological standards for Cochrane systematic reviews endorsed at the 2011 Madrid Cochrane Colloquium<sup>6</sup>. The methods have changed substantively since our original methods guidance paper published in 2006<sup>7</sup>.

Although the updated Cochrane Handbook is comprehensive and easily accessible, its size is daunting to new authors and a number of review groups have found it helpful to tailor a short summary for their clinical area<sup>8,9,10,11,12</sup>. This tailored guidance is in alignment with the Cochrane Handbook and should improve consistency among authors and thus facilitate comparison across reviews.

Many clinicians and their patients find it difficult to master the key features of systematic reviews and communicate the key issues to each other and to others such as family members. This paper will help the reader master the key features. The companion article will provide a primer on options for translating results from CMSG reviews into “usable” and “useful” formats with the advent of summary of findings tables, podcasts, videos, decision aids, phone apps, and cloud technology<sup>13</sup>.

These guidelines, prepared by the CMSG editors (who have a combination of clinical, knowledge translation, methods and statistics expertise), are intended to complement the handbook and not to substitute for the handbook. We will use specific examples from musculoskeletal reviews to illustrate recommendations from the handbook.

*Defining the question.* First, the research question needs to be clearly formulated using the “PICOS” framework, i.e., a clinically relevant or policy-relevant question that takes into account the patient/population, intervention, comparison, outcomes, and study design, and includes both the benefit and harm of the intervention being studied.

Priority topics for new and updated reviews have been identified by CMSG editors and consumers based on criteria including the burden of disease, equity, identification of new interventions, number of new studies, and frequency of access for existing reviews.

*Literature search and study selection.* The complete search strategy for each database searched is defined *a priori* and is documented in the review appendices with the date so that the search can be duplicated<sup>14,15</sup>. The search strategy frequently needs tailoring to the topic, so it is reviewed before implementation by the CMSG Trial Search Coordinator and peer reviewed by information science specialists using the PRESS checklist (Peer Review of Electronic Search Strategies)<sup>16</sup>.

It is recommended that, at a minimum, the following databases and trial registers be searched: MEDLINE, EMBASE, the Cochrane Central Register of Controlled Trials (CENTRAL), ClinicalTrials.gov, and the World Health Organization International Clinical Trial Registry Platform portal. In addition, we recommend checking the references in identified relevant systematic reviews and individual studies that meet the review’s inclusion criteria. For systematic reviews of drugs, authors should search for adverse effects in Websites of regulatory authorities such as the US Food and Drug Administration (FDA) and the European Medicines Evaluation Agency (EMA)<sup>17</sup>. For example, the FDA Website contains important trial and observational study data on tuberculosis and fungal infections from the use of biologics, which were included in the Cochrane and BMJ reviews on biologics overview and network metaanalysis<sup>18,19</sup>. The Trial Search Coordinator may suggest additional sources, such as conference abstracts from ACR and the European League Against Rheumatism (EULAR), depending on the topic.

We do not recommend excluding trials in languages other than English<sup>20</sup>. Some topics such as studies of the effects of medicinal plants may have a significant number of trials published in another language and the CMSG can assist with translation when necessary.

Two people should independently screen the titles and abstracts from the results of the searches for the selection of trials meeting the predefined inclusion criteria. The full text of those articles that appear to meet the inclusion criteria should then be obtained and assessed for eligibility. Those full-text studies that do not meet the inclusion criteria should be added to the Table of Excluded Studies and a reason provided for their exclusion.

## INCLUSION CRITERIA

The minimum criteria for trial inclusion in the systematic review should be defined in advance and address several items using the PICOS framework.

**Population.** Participants of trials should be defined by acceptable diagnostic criteria where possible, such as the ACR criteria for OA, RA, gout, systemic lupus erythematosus, and fibromyalgia (FM)<sup>21,22,23,24,25,26</sup>. Specific exclusions, such as age, sex, and condition, must be detailed.

Trials may report “mixed populations” in which patients with different conditions are enrolled. For example, for reviews on knee OA, randomized controlled trials (RCT) included patients with both hip and knee OA<sup>27</sup>. If such situations are anticipated, review authors should define in advance how to handle these reports. Rather than excluding, a rule may be chosen to include those trials, requiring that at least a given percentage, such as 75%, of patients meet the inclusion criteria. It is also desirable to contact trial authors, to obtain data for the population of interest.

**Intervention.** Glasziou, *et al* have pointed out the importance of having sufficient information in papers to be able to apply the intervention to patients<sup>28</sup>. The intervention must be explicitly described. If applicable, the route of administration, dose, timing, duration of treatment, and concomitant treatments should be outlined.

An example of a definition of type of intervention could read like this<sup>29</sup>: Trials were included that investigated treatment with adalimumab 40 mg subcutaneously every week to every other week, alone or in combination with disease-modifying antirheumatic drugs (DMARD) for a minimum of 12 months.

**Comparator.** The comparator intervention should be explicitly defined (e.g., placebo, another treatment).

An example of a definition of type of comparator<sup>30</sup>: Studies were included comparing leflunomide treatment (as monotherapy or in combination with other DMARD) at a dose of 20 to 25 mg/day (with or without a loading daily dose of 100 mg given in the first 1 to 3 days) to placebo or other DMARD.

**Outcomes.** Cochrane reviews are only as useful as their outcomes are relevant and accurate. Cochrane reviews now report results by outcome. They should include all outcomes that are likely to be meaningful, and not include trivial outcomes. At the time of the title registration, the authors should list all patient-important outcomes, including both benefits and harms relevant to the intervention, organized from the most important to the least important. The major outcomes (up to a maximum of 7) to be presented in the “summary of findings” (SoF) table are selected at the protocol stage.

Review authors must choose at the protocol stage what they consider the main timepoint of interest for each

outcome. This does not imply that they should extract only 1 timepoint; to the contrary, analyzing and depicting results over time is very informative. However, defining the timing in advance forces the review author to think about short-term and longterm effects and to consider whether both are relevant for their intervention in question. It helps to plan statistical analyses and to define the focus of the review.

The SoF table (Figure 1), which may be created using GRADEProfiler software<sup>31</sup>, is now shown on the first page of every Cochrane review, along with the matching abstract and plain language summary. Although review authors may complete an SoF table for each major comparison in their review, to best convey the main “evidence-based actionable message” to users, the Cochrane Library format places the single most important SoF table on the first page of the review. This means that the accuracy and consistency of the numbers and wording across the SoF table, plain language summary, and abstract are pivotal.

The CMSG is in the process of developing default outcome templates for SoF tables for classes of interventions for each condition. Standardizing the outcomes presented in these tables will improve consistency for readers and also permit easier production of overviews of reviews using network metaanalyses. The CMSG editorial team has drafted preliminary default guidelines for which outcomes should be included in SoF tables for pharmacologic and complementary interventions in the following conditions: RA, OA, FM, and ankylosing spondylitis. The preliminary default templates for pharmacologic and complementary interventions for RA and OA are shown in Table 1. These may need tailoring for specific interventions or for specific research questions and may require a different set of major outcomes. For example, in a review on a biologic for RA<sup>32</sup>, radiographic progression would be an important outcome of interest; in a review of arthroplasty, the imaging outcomes would be very different; while in a review on the effect of patient education programs for RA<sup>33</sup>, imaging changes may not be of key interest.

Common core sets established and validated by groups such as OMERACT (Outcome Measures in Rheumatology — an international initiative to improve outcome measurement in rheumatology) and their associated groups are encouraged. This has yet to be done for regional musculoskeletal disorders such as shoulder and elbow disorders, where currently a set of standardized measures does not exist; until then we favor a description of the most relevant ones from the patient’s perspective. The CMSG has a joint working group with OMERACT and their partners to develop standardized outcomes by both condition and intervention for use in CMSG SoF tables.

The CMSG accepts surrogate outcomes if other outcomes are not available and if they meet the following conditions: (1) they have been shown to be on the causal

Abatacept (2 and 10 mg/kg) + DMARD/biologic versus placebo + DMARD/biologic for rheumatoid arthritis							
Patient or Population: patients with rheumatoid arthritis							
Settings: International; clinic/hospital							
Intervention: Abatacept (2 and 10 mg/kg) + DMARD/biologic							
Comparison: Placebo +DMARD, /biologic							
Outcomes	Placebo +DMARD/biologic	Abatacept (2 & 10 mg/kg) +DMARD/biologic	Relative Effect (95% CI)	No of Participants (Studies)	Quality of Evidence (GRADE)	Comments (95% CI)	
ACR 50% improvement Follow-up: 12 months	168 per 1000	371 per 1000 (291 to 474)	RR 2.21 (1.73 to 2.82)	993 (3)	Moderate	Absolute difference =21% (16% to 27%). NNT=5 (4 to 7) Relative percent change =121% (73% to 182%)	
Pain measured at end of study on a 100 mm visual analog scale. Scale from 0 (better) to 100 (worse). Follow-up: 12 months.	Mean pain in the control group was 49.24 mm	Mean pain in the intervention group was 10.71 mm lower (12.97 to 8.45)		1425 (1)	Moderate	Absolute difference=-11% (-13% to -8.5%) NNT=5 (4 to 6) Relative percent change =18% (-22% to -1.4%)	
Improvement in physical function (HAQ: greater than 0.3 increase from baseline, 0-3 scale) Follow-up: 12 months	393 per 1000	637 per 1000 (531 to 766)	RR 1.62 (1.35 to 1.95)	638 (1)	Moderate	Absolute difference = 24% (16% to 32%) NNT=5 (4 to 7) Relative percent change =62% (35% to 195%)	
Achievement of low disease activity state (DAS 28 less than 3.2, scale 0-10) Follow-up: 12 months	98 per 1000	424 per 1000 (278 to 646)	RR 4.33 (2.84 to 6.59)	683 (1)	Moderate	Absolute difference = 33% (26% to 39%) NNT=4 (3 to 5) Relative percent change =333% (184% to 559%)	
Total serious adverse events Follow-up: 6 to 12 months	121 per 1000	127 per 1000 (105 to 155)	RR 1.05 (0.87 to 1.28)	3151 (6)	Moderate	Absolute difference = 1% (-2% to 3%). NNT=n/a Relative percent change =5% (-14% to 29%)	
Long-term serious adverse events Follow-up: 2 years	See comment	See comment	Not estimable	950 (2)	Low	Number of patients with SAE: Genovese 2005: 103/357; 23.4 SAE/100 patient-years; 70% completed the LTE Kremer 2006: 149/593; 16.3 SAE/100 patient-years; 90.5% completed the LTE	

Figure 1. Summary of findings table<sup>32</sup>. DMARD: disease-modifying antirheumatic drug; ACR: American College of Rheumatology; HAQ: Health Assessment Questionnaire; DAS: Disease Activity Score; RR: risk ratio; NNT: number needed to treat; SAE: severe adverse events; n/a: not applicable; LTE: longterm extension; GRADE: Grading of Recommendation Assessment Development and Evaluation. From Maxwell L, Singh JA. Cochrane Database Syst Rev 2009(4):CD007277; with permission.

Table 1. Rheumatoid arthritis (RA) and osteoarthritis (OA) preliminary default outcome templates for pharmacologic and complementary interventions.

RA: Major Outcomes for the SoF Table	OA: Major Outcomes for the SoF Table
ACR 50	Pain
DAS (low or remission)	Physical function
HAQ for function	Quality of life
Radiograph or appropriate imaging changes*	Radiograph or appropriate imaging changes*
Patients who withdraw because of adverse events	Patients who withdraw because of adverse events
Serious adverse events	Serious adverse events
Longterm adverse events	Longterm adverse events

\* Where possible, biomarkers such as radiographs should be transformed into an outcome. ACR: American College of Rheumatology; SoF: summary of findings; DAS: Disease Activity Score; HAQ: Health Assessment Questionnaire.

pathway between the disease and target patient-important outcome<sup>34,35,36</sup>. A high correlation would be indicative of this<sup>37</sup>; (2) the change in the surrogate largely identifies the intervention's effect on the patient-important outcome<sup>34,35,36</sup>. For example, it is tempting to make inferences about the anti-fracture efficacy of pharmacotherapies on the basis of their effects on bone mineral density (BMD). However, there are many limitations associated with using BMD for this purpose. Studies using efficacy estimates from metaanalyses of RCT of antiresorptive therapies to explore the relationship between BMD and fractures using logistic regression analysis have demonstrated that the increases in BMD do not adequately explain the reduction in fracture risk<sup>38,39</sup>. It is important that the limitations of surrogate outcomes be clearly outlined in the review, and these are reflected in the subsequent Grading of Recommendation Assessment Development and Evaluation (GRADE) evidence profile.

Information on harms as well as benefits must be included. The Cochrane Adverse Effects Methods Group has developed guidelines for evaluating adverse effects<sup>17</sup>. The minimum recommendation is to collect the adverse events/effects reported in the trials. Adverse events are usually generic. If they do not differ across clinical indications, authors are encouraged to pool these from the other indications, when possible<sup>18</sup>. Where rare or delayed serious toxicity is a major concern, it is appropriate to do a more comprehensive review of adverse effects by including data from observational studies and high-quality registries. This is a daunting task when there are numerous observational studies, so at a minimum, information from regulatory authority Websites should be searched and a focused, selective review carried out. One example is the review by Rostom, *et al*,<sup>40</sup> in which unpublished RCT safety data from the FDA Website were included in the metaanalyses.

## STUDY DESIGN

Review authors should consider what study designs are likely to provide valid data to answer their questions. The study designs included will depend upon the question, the context, and the resources of the systematic review team. Reviews should define selection criteria for study designs according to their "fitness for purpose" for the research question being posed, rather than just follow an evidence hierarchy<sup>5</sup>. The rationale for the fitness for purpose should be clearly stated and explained.

RCT, where 2 or more groups are formed by randomly allocating participants so that any differences between groups can be attributed to the intervention, should always be included. Controlled clinical trials (CCT) are trials where allocation to treatment and control groups is quasi-random, for example, alternation, date of birth, or case record number. In some treatment settings, such as educational interventions, it is not possible to randomize individuals because of the risk of 1 group receiving some or all of the intervention of the other group (i.e., contamination) as a result of being in the same setting or place; cluster RCT overcome this contamination by randomizing the different individual practices to different groups<sup>41</sup>. Crossover RCT, in which each patient is allocated to a sequence of treatment and control interventions, can also be included, but their analyses need special attention<sup>5</sup>.

Other study designs can also be included, whenever possible, according to their "fitness for purpose". In 2009, 6% of Cochrane reviews included nonrandomized study designs<sup>42</sup> and it has become a priority to develop the skills and best-practice methods to ensure that this component of systematic reviews is useful. The Cochrane Non-Randomised Trial Methods Group has developed guidelines for nonrandomized studies to standardize searching and assessment of rare and delayed adverse effects that will not be detected in short-term trials<sup>43</sup>. A series of 6 papers has been published that provides an update on the increasing consensus on how these studies should be assessed and synthesized. The last article provides useful checklists to help authors<sup>44</sup>.

An example of study design<sup>45</sup>: To assess benefits and harms we included RCT. To further assess harms, we included the following types of studies as long as they reported at least 1 year of followup for patients taking anti-tumor necrosis factor agents, had a sample size > 100 patients<sup>46</sup>, and reported an *a priori*-selected adverse effect outcome: CCT, cohort studies (prospective, e.g., longterm extension of RCT, or retrospective), case-control studies, case series, and published registry data.

*Assessment of risk of bias.* In RCT, because an included study may be performed to the highest possible standards but some individual patient-important outcomes may be underpowered and/or still have an important risk of bias, the 2011 handbook also recommends the assessment of risk of bias by outcomes<sup>5</sup>.

Risk of bias is assessed in a 2-step process (Figure 2, Figure 3).

Step 1: The risk-of-bias tool addresses 7 different domains: (1) sequence generation; (2) allocation concealment; (3) blinding of participants and personnel; (4) blinding of outcome assessment; (5) incomplete outcome data; (6) selective outcome reporting; and (7) “other sources of bias”. Review authors need to specify in their protocol which issues they will consider for “other sources of bias”. Other potential sources of bias should address issues that may affect the internal validity of the study. The handbook provides further details on these issues such as significant baseline imbalances between groups, or situations where a cointervention is not administered evenly between groups. Each domain includes 1 or more specific entries in a risk-of-bias

table. Within each entry, the first part of the tool involves assigning a judgment of low risk, high risk, or unclear risk, and the second part involves providing an explanation of the judgment. A summary table of review authors’ judgments for each risk-of-bias item for each study is shown in Figure 2 for the “Abatacept for rheumatoid arthritis” review. Independent assessment of risk of bias should be undertaken by at least 2 review authors. Where differences in assessment cannot be resolved, arbitration by a third person is warranted.

Step 2: The handbook suggests summarizing risk of bias for each important outcome within and across studies using 3 categories — low, unclear, and high risk of bias. Within a single trial, different outcomes may be at different risk of bias given that different studies may contribute to each

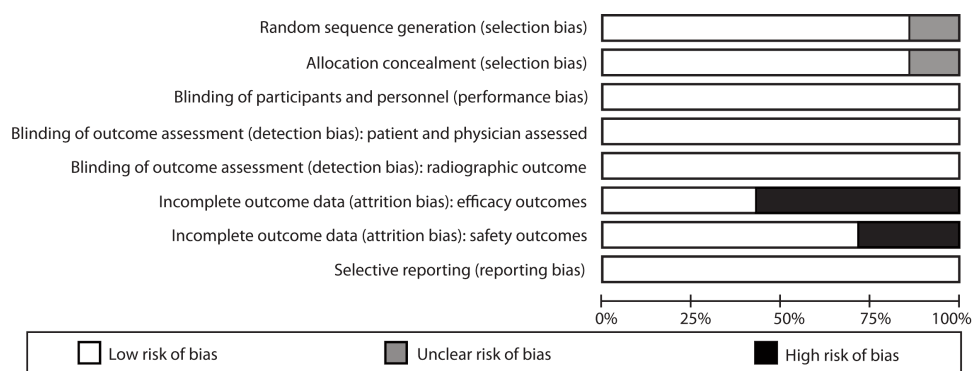


Figure 2. Risk-of-bias summary: review authors’ judgments about each methodological item for each included study in the “Abatacept for rheumatoid arthritis review”<sup>47</sup>. From Maxwell L, Singh JA. Cochrane Database Syst Rev 2009(4):CD007277; with permission.

Item	Weinblatt 2007	Weinblatt 2006	Schiff 2008	Moreland 2002	Kremer 2006	Kremer 2003	Genovese 2005
Random sequence generation (selection bias)	+	+	?	+	+	+	+
Allocation concealment (selection bias)	+	+	?	+	+	+	+
Blinding of participants and personnel (performance bias)	+	+	+	+	+	+	+
Blinding of outcome assessment (detection bias): Patient and physician assessed	+	+	+	+	+	+	+
Blinding of outcome assessment (detection bias): Objective (radiographic outcomes)	+	+	+	+	+	+	+
Incomplete outcome data (attrition bias): Efficacy outcomes	-	-	+	+	-	-	+
Incomplete outcome data (attrition bias): Safety outcomes	-	+	+	+	+	-	+
Selective reporting (reporting bias)	+	+	+	+	+	+	+

Figure 3. Risk-of-bias graph: review authors’ judgments about each risk-of-bias item presented as percentages across all included studies in the “Abatacept for rheumatoid arthritis” review<sup>47</sup>. Maxwell LJ, Singh JA. Abatacept for rheumatoid arthritis: a Cochrane systematic review. From Maxwell L, Singh JA. Cochrane Database Syst Rev 2009(4):CD007277; with permission.

outcome. Figure 3 shows a plot of the distribution of review authors' judgments across studies for each risk-of-bias item in the "Abatacept for rheumatoid arthritis" review.

In nonrandomized (observational) studies, assessment of risk of bias is more difficult than assessment in an RCT. Risk-of-bias assessment methods for systematic reviews of nonrandomized studies are under development for the Cochrane Collaboration. Meanwhile, 6 existing useful tools have been identified<sup>3,48</sup>. One tool is the Newcastle-Ottawa Scale<sup>49</sup> ([www.ohri.ca/programs/clinical\\_epidemiology/oxford.htm](http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm)), which assesses cohort and case-control studies and takes 5–10 min to complete. The second, by Downs and Black<sup>50</sup>, is a longer tool taking about 10–20 min to complete. CMSG authors should also consider the methodological checklists developed by the Scottish Intercollegiate Guidelines Network ([www.sign.ac.uk/methodology/checklists.html](http://www.sign.ac.uk/methodology/checklists.html)).

### DATA COLLECTION AND ANALYSIS

The CMSG recommends that at least 2 review authors independently extract data from included studies. Data collection forms should be used on all CMSG reviews and it is recommended that they be piloted on a sample of studies. It is important that key characteristics and contextual factors of each study be identified for entry into the "Table of Included Studies." At the editorial office we have developed data collection forms that can be modified for new reviews.

Priority is given to extracting the information on up to 7 outcomes predefined for inclusion in the SoF table that is presented on the first page of the Cochrane review; however, the review authors should be alert to the possibility of important, unexpected findings, particularly serious adverse effects that may need to be recorded.

It is usually desirable to collect summary data separately for each intervention group and to enter these into RevMan (Review Manager, the software used for preparing and maintaining Cochrane Reviews), where effect estimates can be calculated. Examples are frequency summary data upon which effect estimates such as risk ratio (RR), OR, and risk difference (RD) can be calculated, or mean and SD for continuous data upon which effect estimates such as mean difference (MD) and standardized mean difference (SMD) can be calculated. Chapter 7 in *The Cochrane Handbook* describes how data should be extracted and converted when necessary to obtain an effect estimate.

Cluster randomized trials are often incorrectly analyzed<sup>41</sup>. When including cluster randomized trials in a review, we recommend that a CMSG statistical editor be consulted.

*Metaanalysis.* Because we want to provide the best numerical estimate of the probability of each patient-important outcome, metaanalysis should be undertaken when data are sufficiently clinically homogeneous. Straightforward statistical analyses should be performed

using RevMan, if data are available and sufficiently similar. The timing of outcome measures should be provided for the most clinically relevant time frame. It may be appropriate to provide summary estimates for short, medium, or long term, depending on the intervention. For example, in the CMSG network metaanalysis of biologics for RA<sup>18</sup>, the methods section defined the following timing of outcomes: short ( $\leq 6$  mo), intermediate ( $> 6$  to 12 mo), or long duration ( $> 1$  yr). Estimates based on these different timings were presented in the SoF tables.

*Effect measures.* The "effect sizes" for dichotomous outcomes may be expressed in RR, OR, Peto OR, or RD. Although absolute differences in important outcomes is the preferred way of presenting the magnitude of the benefit or harm of an intervention to patients and their clinicians, an absolute measure such as the RD is very vulnerable to heterogeneous baseline rates; a relative effect size measure is more stable, so these are preferred for the statistical estimation. The CMSG recommends that RR be used to express dichotomous outcomes because they are easier to understand<sup>51</sup>. When events are rare, the Peto OR is recommended<sup>52</sup>. There is no consensus on the definition of "rare;" a working rule of an event rate of  $< 10\%$  can be used, but special care is needed when studies within the metaanalysis do not have any events.

For continuous outcomes, relative differences are again used, such as MD between the postintervention values, or the difference between baseline values and postintervention values, of the intervention and control groups. SMD should be used when results for continuous outcomes measuring the same concept are presented on different scales; for example, visual analog scale (VAS) and Likert pain scales. One important caveat associated with the use of SMD values in metaanalyses is that few clinicians, patients, journalists, or policy makers understand how to interpret them; we recommend transforming them back to a well-known scale for the SoF table, abstract, and plain language summary (e.g., VAS pain). For examples of this conversion, see Bliddal and Christensen<sup>53</sup>.

Although relative difference metrics are used in the RevMan statistic calculations, patients and their clinicians also need to be provided with absolute differences in the patient-important benefits and harms as listed in the SoF table. The frequency of events without treatment (i.e., the baseline prevalence) makes a marked difference. For example, a relative 50% success in achieving a patient-important reduction in severe pain in a group of 100 patients, 90 of whom report severe pain without the treatment of interest, gives an absolute patient-important reduction in severe pain in 45 patients out of 100 patients with pain. This is substantively different from the same relative 50% success only providing a patient-important reduction in severe pain in 5 of a group of 100 patients when only 10 report severe pain without the treatment of interest.



Another way to interpret an SMD value is to convert it into a number needed to treat (NNT) through a transformation to an OR<sup>54,55</sup>. This OR can then be combined with an assumed control group risk to obtain an absolute benefit as in NNT<sup>56</sup>. To do this for continuous or categorical data, the review authors need to estimate a reasonable control event rate — the percentage of patients who would be expected to respond to placebo/sham therapy.

*Subgroup analyses.* Subgroups are frequently of clinical or policy importance, e.g., to determine the effects of dosage or disease severity on the response to treatment. The problem is that they may show spurious differences, i.e., by chance alone. However, if the review authors have some clear objectives that justify this in advance — to confirm clinically sound hypotheses — the CMSG endorses this prespecified behavior. Therefore, as few subgroups as possible should be prespecified. These should be justified against the criteria proposed by Sun, *et al*<sup>57</sup> (Table 2).

*Diversity/heterogeneity of effect sizes across available studies.* Following the terminology of the handbook, the terms “heterogeneity” or “diversity” may be used to describe variability among studies included in a systematic review. Clinical diversity (or heterogeneity) is the most important — that is, the variability in participants, interventions (e.g., dose), context, comparator (including differences in “usual care”), and outcomes (both surrogate and clinical). Variability in study design is termed “methodological diversity (or heterogeneity)”. “Statistical heterogeneity” (or conventionally just heterogeneity) is the term used when the variation in intervention effects between studies is greater than that expected by chance. “Inconsistency” is the term used for quantifying the effect of heterogeneity on the metaanalysis.

This issue is characterized by the expression “one cannot combine apples and oranges”. It is important to take an initial look at the results for both clinical diversity and methodological diversity. Clinical diversity is assessed by checking that the patients, interventions, and comparators

are not too different from each other such that combining them is clinically useless. Methodological diversity means checking that the studies are similar in terms of study design and risk of bias. Once satisfied that the studies are minimally diverse and that it makes sense to combine them in a metaanalysis, an assessment of the statistical heterogeneity must be undertaken by examining the forest plot and result of the  $I^2$  statistic and the  $\tau^2$  statistic, described below.

A forest plot provides a visual sense of heterogeneity because one can easily see whether the different point estimates of the effect size of each trial all show either a benefit or harm. RevMan calculates  $I^2$  and  $\tau^2$  statistics, used to indicate the presence of statistical heterogeneity. The  $\tau^2$  statistic provides an estimate of the between-study variance. The  $I^2$  statistic describes the percentage of total variation across studies due to heterogeneity, and it does not inherently depend on the number of studies in the metaanalysis<sup>58</sup>, although the size of trials included in the metaanalysis should be taken into consideration for proper interpretation<sup>59</sup>. In Figure 4, at 12 months the effect size has an  $I^2$  of 0%, which is consistent with a  $\tau^2$  of 0.

If the effects observed across trials are inconsistent and vary to a large extent (say,  $I^2 > 50\%$ ), it is important that the review authors explore the results again and try to assess whether the differences can be explained by some clinical or methodological heterogeneity<sup>60</sup>. Inconsistency that cannot be explained (i.e., reduced) by prespecified stratified analyses will lead to an overall estimate with less confidence when interpreting the inference from the metaanalysis. In this case, instead of the fixed-effect approach, a suitable, more conservative approach would be a random-effects metaanalysis, so that the between-study variance is considered and the uncertainty of the effect estimate is reflected in wider CI in the model.

*Sensitivity analyses.* Sensitivity analyses should be performed to examine the strength of the results to risk of bias and the influence of other variables. Authors should prespecify in their protocol which key domains of the

Table 2. Criteria to assess the credibility of subgroup analyses. The greater the extent to which these criteria are met, the more plausible the putative subgroup effect.

Design	<ul style="list-style-type: none"> <li>• Is the subgroup variable a characteristic measured at baseline or after randomization?</li> <li>• Is the effect suggested by comparisons within rather than between studies?</li> <li>• Was the hypothesis specified <i>a priori</i>?</li> <li>• Was the direction of the subgroup effect specified <i>a priori</i>?</li> <li>• Was the subgroup effect one of a small number of hypothesized effects tested?</li> </ul>
Analysis	<ul style="list-style-type: none"> <li>• Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?</li> <li>• Is the significant subgroup effect independent?</li> </ul>
Context	<ul style="list-style-type: none"> <li>• Is the size of the subgroup effect large?</li> <li>• Is the interaction consistent across studies?</li> <li>• Is the interaction consistent across closely related outcomes within the study?</li> <li>• Is there indirect evidence that supports the hypothesized interaction (biological rationale)?</li> </ul>

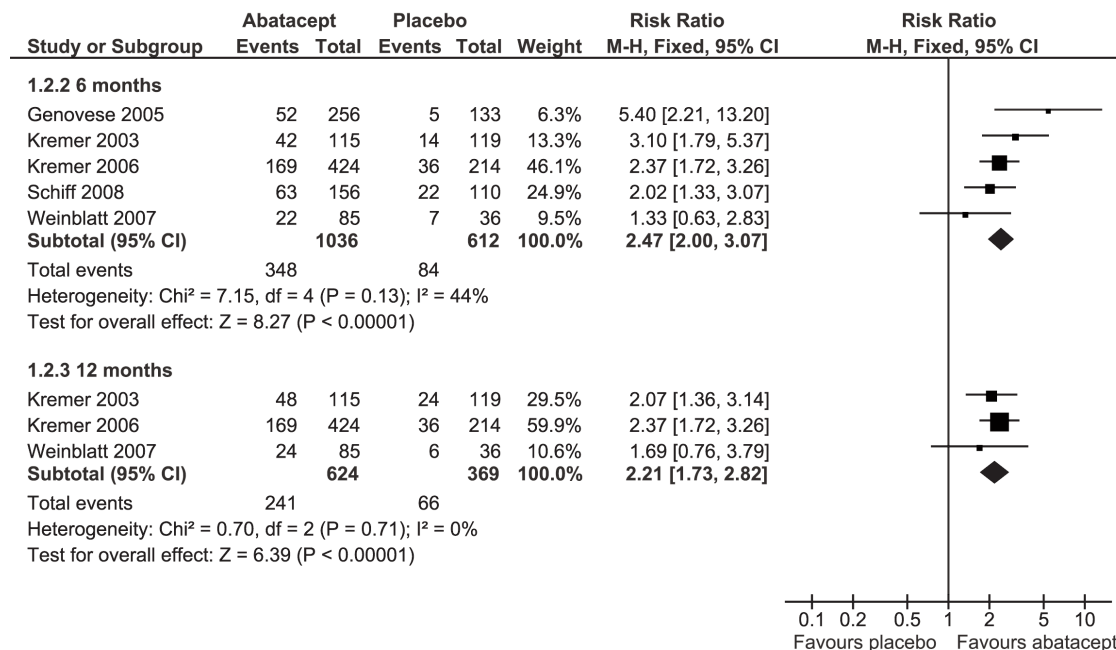


Figure 4. Forest plot from RevMan for abatacept for treating rheumatoid arthritis<sup>47</sup>. M-H: Mantel-Haenszel methods. From Maxwell L, Singh JA. *Cochrane Database Syst Rev* 2009(4):CD007277; with permission.

risk-of-bias criteria will be used to perform a sensitivity analysis, by outcome. For example, for each major outcome, those studies contributing data to that outcome that are judged at low risk of bias for the domains of allocation concealment, blinding of patients and outcome assessors, and incomplete outcome data may be compared with all studies to check the strength of the result of including all studies versus a restricted set of studies with a stronger methodological design. Effectiveness/pragmatic studies may need additional sensitivity analyses of considerations such as different populations, differences in interventions, or patient adherence<sup>61</sup>.

**Forest plots.** Using RevMan, the results of individual studies should be presented graphically in forest plots (Figure 4). The overall effect size is shown as a diamond (individual studies as a square), and the horizontal points of the diamond (horizontal line in an individual study) illustrate the 95% CI. The treatment effect is determined by the location of the square in relation to the vertical middle line that indicates the null hypothesis; an effect size is considered to have no statistical significance when the CI crosses the vertical middle line. When appropriate, data from more than 1 trial may be combined in a metaanalysis, and the diamond at the bottom of the graph provides an estimate of effect of this pooled data.

**Grading of the evidence.** In an effort to make it easier for the end user to understand the quality of the evidence or the “degree of confidence” in the reported results included in the review, we recommend that a rating or grade of the evidence for each major outcome be provided in each

review. The GRADE approach<sup>31,62,63</sup> now replaces the simplified grading system that was derived by the editors of *Evidence-based Rheumatology*<sup>14</sup>.

The GRADE approach specifies 4 levels of quality: high, moderate, low, and very low to quantify the “degree of confidence” in the reported results per outcome (Table 3). Note that this requires a decision for all the studies included in an SoF table (and hence is distinct from the assessment of the risk of bias or methodological strength of the individual studies). The highest quality rating is for a body of evidence based on data from randomized trials without important limitations and the lowest quality rating is for a body of evidence based on case series/case reports.

The quality of the body of evidence involves consideration of 5 factors (Table 4) that allow authors of systematic reviews to make a transparent judgment in how they downgrade the quality rating. With observational studies or downgraded randomized trials, 3 factors permit upgrading to moderate or even high quality (Table 4).

A detailed description of the factors that reduce or increase the quality of the evidence is provided in Chapter 12 of the *Cochrane Handbook*<sup>31</sup>.

Systematic reviews need to be conducted according to high methodological standards. Designed to accompany the detailed *Cochrane Handbook for Systematic Reviews of Interventions*<sup>5</sup>, this report provides guidelines tailored to authors undertaking a review within the *Cochrane Musculoskeletal Group* scope. These guidelines are consistent with the November 2011 *Methodological Standards for the Conduct of Cochrane Intervention*

Table 3. Levels of quality and definition of a body of evidence in the Grading of Recommendation Assessment Development and Evaluation (GRADE) approach<sup>31,62,63</sup>.

Underlying Methodology	Quality of Rating or Grade	Definition
Randomized trials; or double-upgraded observational studies	High	Further research is very unlikely to change our confidence in the estimate of effect
Downgraded randomized trials; or single upgraded observational studies	Moderate	Further research may change the estimate (and is likely to have an important influence on our confidence in the estimate of effect)
Double-downgraded randomized trials; or observational studies	Low	Further research is likely to change the estimate (and is very likely to have an important influence on our confidence in the estimate of effect)
Triple-downgraded randomized trials; or single downgraded observational studies; or case series/case reports	Very low	Any estimate of effect is very uncertain

Table 4. Factors that may influence the quality level of a body of evidence.

Decrease Quality	Increase Quality
<ol style="list-style-type: none"> <li>1. Within study, risk of bias causes the overall analysis to be biased</li> <li>2. Indirectness of evidence (i.e., considering each of the following 4 PICO letters — participant, intervention, control, and outcome, especially surrogates). If any of the PICO factors are not directly clinically relevant, the evidence might be judged appropriate for downgrading</li> <li>3. Unexplained heterogeneity or inconsistency of results (including problems with subgroup analyses, making it difficult to interpret the overall metaanalysis)</li> <li>4. Lack of precision of the overall effect estimates (wide CI), when the overall analysis does not confirm that the effect size is evident</li> <li>5. Risk of publication bias</li> </ol>	<ol style="list-style-type: none"> <li>1. A large magnitude of effect</li> <li>2. All plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results show no effect</li> <li>3. Evidence of a dose-response gradient</li> </ol>

Reviews<sup>6</sup>. These guidelines on developing and performing a systematic review will help improve the quality of reporting and promote high standards of conduct as well as consistency across CMSG reviews.

## ACKNOWLEDGMENT

CMSG editors who contributed to the preparation of this manuscript: Isabelle Boutron, Angela Busch, Ernest Choy, Robin Christensen, Rob de Bie, Rhian Goodfellow, Tracey Howe, Anne Lyddiatt, Mário Lenza, Philippe Ravaud, Raphaële Seror, Beverley Shea, Maria Suarez-Almazor, and Karine Toupin-April.

## REFERENCES

1. The Cochrane Collaboration. The Cochrane Policy Manual. 2012 (updated 26 April 2012). [Internet. Accessed Sept 13, 2013.] Available from: [www.cochrane.org/policy-manual/welcome](http://www.cochrane.org/policy-manual/welcome).
2. Hochberg MC, Altman RD, April KT, Benkhalti M, Guyatt G, McGowan J, et al. American College of Rheumatology 2012 recommendations for the use of nonpharmacologic and pharmacologic therapies in osteoarthritis of the hand, hip, and knee. *Arthritis Care Res* 2012;64:465-74.
3. Singh JA, Furst DE, Bharat A, Curtis JR, Kavanaugh AF, Kremer JM, et al. 2012 Update of the 2008 American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. *Arthritis Care Res* 2012;64:625-39.
4. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *Review. J Clin Epidemiol* 1995;48:167-71.
5. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions* version 5.1.0 (updated March 2011). The Cochrane Collaboration. [Internet. Accessed Sept 13, 2013.] Available from: [www.cochrane-handbook.org](http://www.cochrane-handbook.org).
6. Methodological Expectations of Cochrane Intervention Reviews (MECIR). [Internet. Accessed Sept 13, 2013.] Available from: [www.editorial-unit.cochrane.org/mecir](http://www.editorial-unit.cochrane.org/mecir)
7. Maxwell L, Santesso N, Tugwell PS, Wells GA, Judd M, Buchbinder R. Method guidelines for Cochrane Musculoskeletal Group systematic reviews. *J Rheumatol* 2006;33:2304-11.
8. van Tulder M, Furlan A, Bombardier C, Bouter L. Editorial Board of the Cochrane Collaboration Back Review Group. *Cochrane Collaboration Back Review Group*. 2003;28:1290-9.
9. Furlan AD, Pennick V, Bombardier C, van Tulder M; Editorial Board, Cochrane Back Review Group. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review

- Group. *Spine* 2009;34:1929-41.
10. Cochrane Pain, Palliative and Supportive Care Group. Resource hub. [Internet. Accessed Sept 13, 2013.] Available from: <http://papas.cochrane.org/resources-review-authors>
  11. Cochrane Anaesthesia Group. Resources for review authors. [Internet. Accessed Sept 13, 2013.] Available from: <http://carg.cochrane.org/resources-review-authors>
  12. Cochrane Neuromuscular Disease Group. Statistical Guidelines. [Accessed Sept 13, 2013]; Available from: <http://neuromuscular.cochrane.org/>
  13. Rader T, Pardo Pardo J, Stacey D, Ghogomu E, Maxwell LJ, Welch VA, et al. Update of strategies to translate evidence from Cochrane Musculoskeletal Group systematic reviews for use by various audiences. *J Rheumatol* 2014;41:206-15.
  14. McGowan J. Literature searching. In: Evidence-based rheumatology. Tugwell P, Shea B, Boers M, Brooks P, Simon LS, Strand V, et al, eds. London: BMJ; 2004.
  15. Lefebvre C, Manheimer E, Glanville J. Searching for studies. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 5.1.0*. (updated March 2011). The Cochrane Collaboration. [Internet. Accessed Sept 13, 2013.] Available from: [www.cochrane.org/training/cochrane-handbook](http://www.cochrane.org/training/cochrane-handbook)
  16. Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. *J Clin Epidemiol* 2009;62:944-52.
  17. Loke Y, Price D, Herxheimer A. Chapter 14: Adverse effects. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 5.1.0*. (updated March 2011). The Cochrane Collaboration. [Internet. Accessed Sept 13, 2013.] Available from: [www.cochrane.org/training/cochrane-handbook](http://www.cochrane.org/training/cochrane-handbook)
  18. Singh JA, Christensen R, Wells GA, Suarez-Almazor ME, Buchbinder R, Lopez-Olivo MA, et al. Biologics for rheumatoid arthritis: an overview of Cochrane reviews. *Cochrane Database Syst Rev* 2009(4):CD007848.
  19. Tugwell P, Singh JA, Wells GA. Biologics for rheumatoid arthritis. *BMJ* 2011;343:d4027.
  20. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, LeLorier J, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996;347:363-6.
  21. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum* 2010;62:2569-81.
  22. Wallace SL, Robinson H, Masi AT, Decker JL, McCarty DJ, Yu TF. Preliminary criteria for the classification of the acute arthritis of primary gout. *Arthritis Rheum* 1977;20:895-900.
  23. Wolfe F, Smythe HA, Yunus MB, Bennett RM, Bombardier C, Goldenberg DL, et al. The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. *Arthritis Rheum* 1990;33:160-72.
  24. Tan EM, Cohen AS, Fries JF, Masi AT, McShane DJ, Rothfield NF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1982;25:1271-7.
  25. Altman R, Alarcon G, Appelrouth D, Bloch D, Borenstein D, Brandt K, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum* 1991;34:505-14.
  26. Altman R, Asch E, Bloch D, Borenstein D, Brandt K. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the knee. *Arthritis Rheum* 1986;29:1039-49.
  27. Rutjes AW, Nuesch E, Sterchi R, Kalichman L, Hendriks E, Osiri M, et al. Transcutaneous electrostimulation for osteoarthritis of the knee. *Cochrane Database Syst Rev* 2009(4):CD002823.
  28. Glasziou P, Chalmers I, Altman DG, Bastian H, Boutron I, Brice A, et al. Taking healthcare interventions from trial to practice. *BMJ* 2010;341:c3852.
  29. Navarro-Sarabia F, Ariza-Ariza R, Hernandez-Cruz B, Villanueva I. Adalimumab for treating rheumatoid arthritis. *J Rheumatol* 2006;33:1075-81.
  30. Osiri M, Shea B, Robinson V, Suarez-Almazor M, Strand V, Tugwell P, et al. Leflunomide for the treatment of rheumatoid arthritis: a systematic review and metaanalysis. *J Rheumatol* 2003;30:1182-90.
  31. Schunemann HJ, Oxman AD, Higgins JPT, Vist GE, Glasziou P, Guyatt GH. Chapter 11: Presenting results and 'Summary of findings' tables. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 5.1.0*. (updated March 2011). The Cochrane Collaboration. [Internet. Accessed Sept 13, 2013.] Available from: [http://handbook.cochrane.org/chapter\\_11/11\\_presenting\\_results\\_and\\_summary\\_of\\_findings\\_tables.htm](http://handbook.cochrane.org/chapter_11/11_presenting_results_and_summary_of_findings_tables.htm)
  32. Maxwell LJ, Singh JA. Abatacept for rheumatoid arthritis: a Cochrane systematic review. *J Rheumatol* 2010;37:234-45.
  33. Riemsma RP, Kirwan JR, Taal E, Rasker JJ. Patient education for adults with rheumatoid arthritis. *Cochrane Database Syst Rev* 2003(2):CD003688.
  34. Li Z, Chines AA, Meredith MP. Statistical validation of surrogate endpoints: is bone density a valid surrogate for fracture? Review. *J Musculoskelet Neuronal Interact* 2004;4:64-74.
  35. Lassere MN, Johnson KR, Boers M, Tugwell P, Brooks P, Simon L, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *J Rheumatol* 2007;34:607-15.
  36. Fleming TR, DeMets DL. Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med* 1996;125:605-13.
  37. Burzykowski T, Buyse M. Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat* 2006;5:173-86.
  38. Cummings SR, Karpf DB, Harris F, Genant HK, Ensrud K, LaCroix AZ, et al. Improvement in spine bone density and reduction in risk of vertebral fractures during treatment with antiresorptive drugs. *Am J Med* 2002;112:281-9.
  39. Sarkar S, Mitlak BH, Wong MM, Stock JL, Black DM, Harper KD. Relationships between bone mineral density and incident vertebral fracture risk with raloxifene therapy. *J Bone Miner Res* 2002;17:1-10.
  40. Rostom A, Muir K, Dubé C, Jolicoeur E, Boucher M, Joyce J, et al. Gastrointestinal safety of cyclooxygenase-2 inhibitors: a Cochrane Collaboration systematic review. *Clin Gastroenterol Hepatol* 2007;5:818-28.
  41. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702-8.
  42. Konnerup M, Kongsted H. Do Cochrane reviews provide a good model for social science? The role of observational studies in systematic reviews. *Evid Policy* 2012;8:79-96.
  43. Reeves B, Deeks JJ, Higgins J, Wells GA. Chapter 13: Including non-randomized studies. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 5.1.0*. (updated March 2011). The Cochrane Collaboration. [Internet. Accessed Sept 13, 2013.] Available from: [http://handbook.cochrane.org/chapter\\_13/13\\_including\\_non\\_randomized\\_studies.htm](http://handbook.cochrane.org/chapter_13/13_including_non_randomized_studies.htm)
  44. Wells G, Shea B, Higgins J, Stern J, Tugwell P, Reeves B. Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Res Synth Methods* 2013;4:63-77.

45. Maxwell LJ. Assessment of intra- and inter-individual variability of outcome measures in ankylosing spondylitis and the efficacy and adverse effects of anti-TNF therapy. [Order No. MR86827, University of Ottawa (Canada)]. ProQuest Dissertations and Theses, 2011. [Internet. Accessed Sept 27, 2013.] Available from: [www.ruor.uottawa.ca/fr/handle/10393/20093](http://www.ruor.uottawa.ca/fr/handle/10393/20093)
46. Nüesch E, Trelle S, Reichenbach S, Rutjes AW, Tschannen B, Altman DG, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ* 2010;341:c3515.
47. Maxwell L, Singh JA. Abatacept for rheumatoid arthritis. *Cochrane Database Syst Rev* 2009(4):CD007277.
48. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovich C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;7:iii-x, 1-173.
49. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. [Internet. Accessed Sept 13, 2013.] Available from: [www.ohri.ca/programs/clinical\\_epidemiology/oxford.htm](http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm).
50. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377-84.
51. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002;21:1575-600.
52. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med* 2007;26:53-77.
53. Bliddal H, Christensen R. The treatment and prevention of knee osteoarthritis: a tool for clinical decision-making. *Expert Opin Pharmacother* 2009;10:1793-804.
54. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000;19:3127-31.
55. da Costa BR, Rutjes AW, Johnston BC, Reichenbach S, Nüesch E, Tonia T, et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. *Int J Epidemiol* 2012;41:1445-59.
56. Osiri M, Suarez-Almazor ME, Wells GA, Robinson V, Tugwell P. Number needed to treat (NNT): implication in rheumatology clinical practice. *Ann Rheum Dis* 2003;62:316-21.
57. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.
58. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
59. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
60. Thompson SG. Systematic review - why sources of heterogeneity in metaanalysis should be investigated. *BMJ* 1994;309:1351-5.
61. 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. *J Clin Epidemiol* 2009;62:464-75.
62. Schünemann HJ, Oxman AD, Vist G, Higgins J, Deeks JJ, Glasziou P, et al. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions* 5.1.0. (updated March 2011). The Cochrane Collaboration. [Internet. Accessed Sept 13, 2013.] Available from: [http://handbook.cochrane.org/chapter\\_12/12\\_interpreting\\_results\\_and\\_drawing\\_conclusions.htm](http://handbook.cochrane.org/chapter_12/12_interpreting_results_and_drawing_conclusions.htm)
63. Schünemann HJ, Brozek J, Oxman AD. GRADE handbook for grading quality of evidence and strength of recommendation: The GRADE Working Group; 2009.