



Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials

Susan Armijo-Olivo^{1,2*}, Maria Ospina³, Bruno R. da Costa⁴, Matthias Egger⁵, Humam Saltaji⁶, Jorge Fuentes^{7,8}, Christine Ha⁹, Greta G. Cummings¹

1 CLEAR (Connecting Leadership and Research) Outcomes Research Program, Faculty of Nursing, University of Alberta, Edmonton, Alberta, Canada, **2** Faculty of Rehabilitation Medicine, Department of Physical Therapy, University of Alberta, Edmonton, Alberta, Canada, **3** Emergency Medicine Strategic Clinical Network, Alberta Health Services, Department of Emergency Medicine, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, Alberta, Canada, **4** Department of Physical Therapy, Florida International University, Miami, Florida, United States of America, **5** Institute of Social & Preventive Medicine, University of Bern, Bern, Switzerland, **6** Orthodontic Graduate Program, School of Dentistry, University of Alberta, Edmonton, Alberta, Canada, **7** Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, Alberta, Canada, **8** Catholic University of Maule, Department of Physical Therapy, Talca, Maule, Chile, **9** Rehabilitation Research Center, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, Alberta, Canada

Abstract

Objectives: To test the inter-rater reliability of the RoB tool applied to Physical Therapy (PT) trials by comparing ratings from Cochrane review authors with those of blinded external reviewers.

Methods: Randomized controlled trials (RCTs) in PT were identified by searching the Cochrane Database of Systematic Reviews for meta-analysis of PT interventions. RoB assessments were conducted independently by 2 reviewers blinded to the RoB ratings reported in the Cochrane reviews. Data on RoB assessments from Cochrane reviews and other characteristics of reviews and trials were extracted. Consensus assessments between the two reviewers were then compared with the RoB ratings from the Cochrane reviews. Agreement between Cochrane and blinded external reviewers was assessed using weighted kappa (κ).

Results: In total, 109 trials included in 17 Cochrane reviews were assessed. Inter-rater reliability on the overall RoB assessment between Cochrane review authors and blinded external reviewers was poor ($\kappa = 0.02$, 95%CI: $-0.06, 0.06$). Inter-rater reliability on individual domains of the RoB tool was poor (median $\kappa = 0.19$), ranging from $\kappa = -0.04$ ("Other bias") to $\kappa = 0.62$ ("Sequence generation"). There was also no agreement ($\kappa = -0.29$, 95%CI: $-0.81, 0.35$) in the overall RoB assessment at the meta-analysis level.

Conclusions: Risk of bias assessments of RCTs using the RoB tool are not consistent across different research groups. Poor agreement was not only demonstrated at the trial level but also at the meta-analysis level. Results have implications for decision making since different recommendations can be reached depending on the group analyzing the evidence. Improved guidelines to consistently apply the RoB tool and revisions to the tool for different health areas are needed.

Citation: Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, et al. (2014) Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials. PLoS ONE 9(5): e96920. doi:10.1371/journal.pone.0096920

Editor: Lamberto Manzoli, University of Chieti, Italy

Received: January 30, 2014; **Accepted:** April 13, 2014; **Published:** May 13, 2014

Copyright: © 2014 Armijo-Olivo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This project is funded by the Canadian Institutes of Health Research (CIHR), Alberta Innovates Health solution through a knowledge translation initiative grant, the Knowledge Translation (KT) Canada research Stipend program, and the Physiotherapy Foundation of Canada (PFC) through a B.E. Schnurr Memorial Fund Award. The funding bodies had no input in the design, collection, analysis, and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sla4@ualberta.ca

Introduction

The term "quality assessment" has been used extensively in the literature, particularly in the context of systematic reviews, to refer to the critical appraisal of primary studies. Different approaches to quality assessment have been proposed for assessing the quality of studies [1,2]. A variety of methods (scales and checklists) have been used by different Cochrane Review groups [3,4]; however, because of methodological inconsistencies across quality instruments and the lack of empirical evidence supporting their validity

and reliability [5,6], the use of these methods was explicitly discouraged in Cochrane reviews [3].

In 2008, the Cochrane Collaboration (CC) initiated a shift in the approach to the evaluation of trial quality by linking the concept of quality to the internal validity of a study (risk of bias; the extent to which the design and conduct of a study are likely to prevent bias) [3]. The Cochrane Collaboration developed the Risk of Bias tool (RoB) as a method to assess risk of bias based on study design and conduct rather than relying on general reporting issues of trial characteristics [3]. Since then, the Cochrane Collaboration

has required the use of the RoB tool to establish consistency in the assessment of study quality across Cochrane Review groups.

The RoB tool is based on six domains and 7 items: sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, and “other sources of bias.” Critical assessments of the risk of bias (high, low, unclear) in each domain are made separately for each outcome in a given study. The choice of these components for inclusion in the tool was based on empirical evidence of their association with effect estimates [5,7,8]; Recent research [9,10] recommends further testing of the psychometric properties (i.e., validity, reliability, and responsiveness) of the RoB tool, and evaluations of the tool in a broad range of research fields. In addition, researchers have called for the use of clear and consistent guidelines and classification systems to apply and interpret the RoB tool [11]. This information is essential since differences in the appraisal and interpretation of risk of bias across trials can explain variation in the interpretation of results of studies included in a systematic review, and ultimately impact the conclusions and clinical practice.

Despite the RoB tool being increasingly used in Cochrane reviews; few studies have assessed its psychometric properties, specifically in paediatric trials, general medical and oncology trials [9,10,12,13]. The inter-rater agreement for the individual domains of the RoB tool has been found to range from poor (κ [kappa] = 0.13 for selective reporting) to substantial (κ = 0.74 for sequence generation) [9]. A recent study [13] assessed the reliability of the RoB tool between individual reviewers and across consensus ratings of pairs of reviewers on a sample of 154 and 30 randomized clinical trials (RCTs) published in the general medical literature respectively. The study found that the reliability between pairs of reviewers was “fair” for most of RoB domains with kappa values ranging from 0.2 to 0.34. However, the agreement between consensus ratings was always poorer than the agreement between pairs of reviewers indicating a high variability in interpreting and applying the RoB tool across different systematic review groups and across systematic reviews [13]. This agreement in consensus ratings (across pair of reviewers) was conducted only on 30 trials within a group of reviewers from the same team using guidelines developed specifically for the study.

The reliability of the RoB tool has not been investigated by comparing ratings of an external blinded panel of reviewers with those obtained from authors of Cochrane reviews. This work is of crucial importance for researchers who incorporate risk of bias assessments from Cochrane- and non-Cochrane systematic reviews into meta-epidemiological research approaches, since risk of bias assessments obtained by different research group can lead to different results. Furthermore, the reliability of the RoB in the context of physical therapy (PT) trials has not yet been evaluated. The objectives of this study were to test the inter-rater reliability of the RoB tool applied to PT trials by comparing consensus ratings from Cochrane review authors with those of blinded external reviewers, and to investigate potential sources of disagreements to inform the use of the RoB tool.

Methods

The Cochrane Database of Systematic Reviews (CDSR) was systematically searched from 2005 to May 25 2011 for meta-analyses of PT interventions using the words physical therapy, physiotherapy, rehabilitation, exercise, electrophysical agents, acupuncture, massage, transcutaneous electrical stimulation (TENS), interferential current, ultrasound, stretching, chest therapy, pulmonary rehabilitation, manipulative therapy, mobilization, and related terms. For a detailed search strategy see

Appendix S1. Meta-analyses and their RCTs were included if: 1) the meta-analysis included at least 5 RCTs, with at least one of the interventions being currently or potentially part of PT practice according to the World Confederation for Physical Therapy (WCPT) [14]; 2) the outcome of interest in the meta-analysis (explicitly described as the main outcome or the outcome with the largest number of trials) was continuous; and 3) the RoB tool was used for assessment of individual trials. A unique identifier was assigned to meta-analyses and trials that met the inclusion criteria.

RoB assessments procedure

The risk of bias of individual trials included in the meta-analyses was assessed on 6 domains (7 items) of the RoB tool [15]: sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, selective outcome reporting, and other sources of bias. We followed the guidelines established by the Cochrane Collaboration to perform RoB assessments; however we developed specific decision rules to make decisions (**Appendix S2**). Risk of bias evaluations for blinding and incomplete outcome data were based upon the primary (continuous) outcome of interest selected for meta-analysis in the Cochrane review. If not clearly specified, the outcome was chosen according to the meta-analysis that contained the largest number of trials in the review. The Cochrane guidelines recommend using trial protocols to complete assessments of selective outcome reporting bias. However, due to the low likelihood of locating protocols for trials, we did not search for study protocols [24]. Therefore, for the category of “low” risk of bias, it was required that trial publications reported all primary and secondary outcomes in the methods and results sections, with new outcomes not being added in the results section. If the primary outcome of the trial was not included in the results, there was a high risk of selective outcome reporting bias. In addition we paired outcomes reported in methods and results sections. If more than 70% of the secondary outcomes were not reported in the results or methods sections, then the study was rated as high RoB. For ‘other bias’, we looked at baseline comparability, control for co-interventions (contamination bias) and whether treatment compliance was acceptable. These criteria have been used in the risk of bias assessments of the Cochrane Back Review Group to determine other sources of potential bias [16].

For the overall assessment of RoB, a trial was considered at low risk of bias if it was rated as low risk in all individual domains; if the rating was unclear in at least one domain, and the other domains were unclear or low, the overall assessment of RoB was unclear. Finally, an overall assessment of high risk of bias was considered if at least one domain was rated as high [12,13].

Two independent reviewers (**any of these reviewers: SAO, JF, HS, CH, AC, DP**) blinded to the RoB ratings reported in the Cochrane reviews assessed the risk of bias of all PT trials included in the meta-analyses. **Each pair of reviewers assessed risk of bias in each study and disagreements were resolved by discussion between reviewers until consensus was reached.** If consensus was not achieved, a final decision on RoB assessments was reached after consultation with a third reviewer (first author), **although this was not necessary.** Blinding of the external panel of reviewers was achieved as follows: 1) reviewers were not told the objective of this study; 2) they were not provided with RoB assessments performed by Cochrane reviewers; 3) after the external panel of reviewers completed their assessments, an independent reviewer who was not part of the review panel extracted RoB data assessment performed by Cochrane reviewers (**MO**). The integrity of blinding was assessed by asking the

reviewers post hoc if they had checked the Cochrane RoB assessment. None of them reported that they did.

Data on RoB assessments from Cochrane reviews and other characteristics of reviews and trials were extracted by one reviewer (MO or SAO) and entered directly into a pilot tested electronic form. Consensus assessments between the two reviewers **from our panel** were then compared with the RoB ratings from the Cochrane reviews. In addition, two reviewers independently assessed the RoB at the meta-analysis level for both groups of reviewers (i.e. external panel of reviewers and Cochrane reviewers) using the guidelines established by the Cochrane handbook [15,17]. A low, unclear and high RoB at the meta-analysis level was defined as: “most information is from studies at low, unclear or high risk of bias respectively” [15,17]. Since no further guidance is in the Cochrane handbook, we established an arbitrary cut-off value of 60% to define the “majority of studies”. Assessments were compared and discrepancies were resolved by consensus between reviewers.

Characteristics of the reviewers' panel

Six reviewers with experience in different areas of health sciences research comprised the review panel in this study. Two reviewers had a Bachelor in Health Sciences (**CH, AC**), one had a Masters in Public Health (**DP**), one had a Masters in Dentistry and currently working on a PhD in Orthodontics (**HS**), and two were physical therapists and had Masters and PhD in Rehabilitation sciences (**SAO, JF**) with at least 10 years of experience in critical appraisal and systematic reviews. Four of them (**DP, HS, SAO, and JF**), had formal training in critical appraisal and systematic reviews. The other 2 (**CH, AC**) had at least one year of hands-on experience conducting systematic reviews. **Four of the reviewers (SAO, JF, HS, CH) were part of the research team collaborating in this project and two of them (DP, AC) were hired to perform the data extraction and quality assessments. All of them verbally agreed to participate as reviewers in this study.**

Training process

All reviewers were trained and received guidelines for RoB assessments from the first author (**SAO**) who was a physical therapist by training and had a MSc and PhD in Rehabilitation Sciences and more than 10 years of experience in critical appraisal and systematic reviews. Reviewer training was carried out using 10 trials not included in the study. Results of RoB assessments for these 10 studies were independently reviewed and discussed in a group meeting to determine consistency in ratings. In addition, the team members met on a regular basis to further calibrate RoB assessments throughout the study.

Statistical analysis

Inter-rater reliability of RoB assessments between Cochrane and blinded external reviewers [18–20] and within the panel of external reviewers was assessed using weighted kappa (κ) for categorical data. Inter-rater scores for both individual domains and overall assessments of the RoB tool were considered. Analyses were conducted using STATA (version 12, Stata Corp; College Station, Texas; USA). For raw data for each domain see **Appendix S3**.

Criteria proposed by Byrt [21] were used to interpret kappa values. Values between 0.93–1.00 represented excellent agreement; 0.81–0.92 very good agreement; 0.61–0.80 good agreement; 0.41–0.60 fair agreement; 0.21–0.40 slight agreement, 0.01–0.20 poor agreement; and 0.00 or less were considered to have no agreement.

Results

Literature search

The systematic search of the CDSR resulted in the identification of 3901 Cochrane review titles, with 271 reviews being potentially relevant to physical therapy. Of these, 68 Cochrane reviews included a meta-analysis of at least five studies on PT interventions assessing a continuous outcome. Figure 1 outlines the retrieval of Cochrane reviews and the number of trials included in the analysis. A total of 109 trials included in 17 Cochrane reviews that used the RoB tool were assessed. Table 1 summarizes the characteristics of the Cochrane reviews included in the study.

Characteristics of selected studies

Briefly, the reviews were published between 2008 and 2011 and included meta-analyses of the effectiveness of PT interventions for musculoskeletal (9 reviews [22–30] cardiorespiratory (4 reviews [31–34], neurological (2 reviews [35,36], gynaecological (1 review [37], and general conditions (1 review [38]).

The majority of Cochrane reviews (15 reviews) did not include a formal evaluation of the inter-rater reliability of the RoB assessments. Although the majority of reviews stated that two independent reviewers assessed study RoB, in four reviews, a single reviewer assessed RoB, with verification by a second reviewer. Similarly, twelve of the 17 (71%) Cochrane systematic reviews did not clearly specify the outcome used for the RoB assessments, whereas eight out of 17 (47%) of systematic reviews combined all outcomes into a single bias assessment.

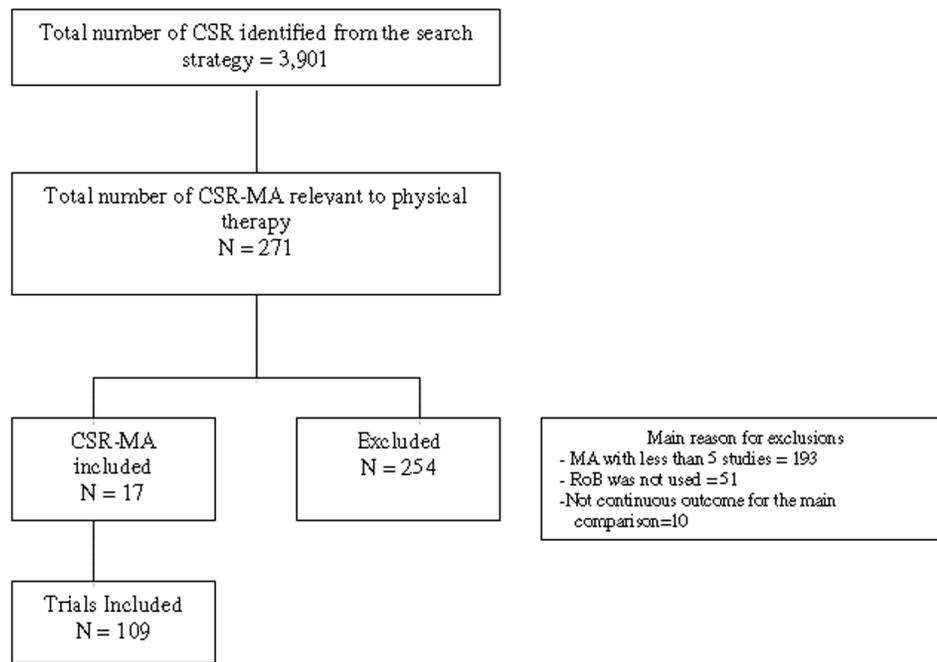
A median number of six trials were included in the meta-analyses (interquartile range: 5, 8). All but one cross-over trial were identified as parallel trials. The majority of trials ($n = 93$) used active controls whereas 15 trials were placebo-controlled. The control group of one trial was not clearly identified. Seventy-five trials were efficacy trials; 26 effectiveness trials, and seven trials combined an evaluation of the efficacy/effectiveness of PT interventions. One trial was not clearly described as an efficacy or effectiveness trial.

The number of trials available for assessing the inter-rater reliability of both individual-domain and overall RoB assessments varied as not all Cochrane reviews reported ratings for all the domains of the RoB tool. Inter-rater reliability of RoB assessments between Cochrane review authors and blinded external reviewers and the inter-rater reliability within the external panel of reviewers are presented in Table 2.

Inter-rater agreement: Cochrane review authors vs. blinded external reviewers

Inter-rater reliability on the overall RoB assessment between Cochrane review authors and blinded external reviewers was poor ($\kappa = 0.02$, 95%CI: $-0.06, 0.06$). Inter-rater reliability on individual domains of the RoB tool was poor (median $\kappa = 0.19$), ranging from $\kappa = -0.04$ (“Other bias”) to $\kappa = 0.62$ (“Sequence generation”). Table 2 displays the inter-rater reliability of the RoB tool between the blinded external review panel versus Cochrane reviewers.

When overall RoB categories assigned by blinded external reviewers were compared to those of Cochrane review authors, we found that the number of trials assessed as “low” risk of bias by Cochrane review authors ($n = 9$) was greater than blinded external reviewers ($n = 3$). Similarly, the number of trials rated as “high” risk of bias by Cochrane review authors ($n = 66$) was greater than blinded external reviewers ($n = 31$). In contrast, blinded external reviewers had a greater number of trials assessed as “unclear” in the overall RoB assessment ($n = 74$) compared to



CSR = Cochrane systematic reviews; MA = meta-analysis; RoB = Risk of Bias

Figure 1. Diagram for the identification of reviews.
doi:10.1371/journal.pone.0096920.g001

Cochrane review authors (n = 33). The main source of disagreement between Cochrane review authors and blinded external reviewers in the overall rating of RoB was due to discrepancies in the classification of “unclear” vs. “high” risk of bias; with 45 trials rated as “high” risk of bias by Cochrane review authors and “unclear” by blinded external reviewers.

Inter-rater agreement within the panel of blinded external reviewers

The inter-rater reliability between blinded external reviewers on the overall RoB rating was fair ($\kappa = 0.55$, 95%CI: 0.40, 0.70). Inter-rater reliability on individual domains of the RoB tool was fair (median $\kappa = 0.56$) ranging from $\kappa = 0.32$ (“Other bias”) to $\kappa = 0.79$ (“allocation concealment”).

Overall RoB at the Meta-analysis level

There was no agreement ($\kappa = -0.29$, 95%CI: $-0.81, 0.35$) in the overall RoB assessment at the meta-analysis level between Cochrane review authors and blinded external reviewers. Cochrane reviewers had evaluated 10 meta-analyses as high RoB while the external panel of reviewers classified them as “unclear”. **Table 3** displays the RoB assessment at the meta-analysis level.

Discussion

Based on the assessment of RCTs included in Cochrane reviews of PT interventions, this study found that the inter-rater reliability of RoB assessments between Cochrane review authors and blinded external reviewers was poor. This result confirms the findings of previous studies regarding the poor reliability of the RoB tool domains in other areas of health research [9,10,12,13]. Our results indicated that RoB assessments in Cochrane reviews could not be

replicated consistently by an external panel of reviewers using consensus RoB assessments.

Consensus ratings are of crucial importance since they are commonly used in systematic reviews. Only one previous study assessed the reliability of the RoB based on consensus assessments across pairs of reviewers from four research centres using a sample of 30 trials indexed in PubMed between 2000 and 2006 [13]. Using a larger number of trials in PT and comparing the RoB consensus ratings between blinded external reviewers and Cochrane reviewers, our study confirmed that agreement across pairs of reviewers is generally lower than agreement between reviewers. Cochrane reviews have long been considered the gold standard for systematic reviews in health care. Results of our study have important implications for the interpretation of results of RoB assessments across Cochrane reviews and produced by different Cochrane Groups. The poor agreement in RoB assessments between Cochrane reviewers and an external panel of reviewers has raised several concerns: 1) RoB assessments cannot be reproduced by different groups of reviewers. If true, it would mean that RoB assessments are not reliable and depend on the reviewers’ level of knowledge and familiarity with the information provided in the individual trials; 2) the RoB tool is a very subjective tool that cannot provide reliable assessments; 3) despite efforts by the Cochrane Collaboration to establish high quality standards for conducting systematic reviews, poor agreement appears to be the norm rather than the exception when conducting RoB assessments. Thus, we pose the following questions: can we trust risk of bias results reported in Cochrane reviews? Can we trust assessments using the RoB tool?

The low reliability of RoB assessments between our panel of blinded external reviewers and Cochrane reviewers has implications for researchers who use bias ratings from Cochrane reviews or other external sources to conduct meta-epidemiological research on the relationship between trial characteristics and over

Table 1. Characteristics of Cochrane systematic reviews on physical therapy interventions that provided trial data for the analysis of inter-rater reliability of RoB.

Author	Review title	Area	Number of RCTs	Outcome Reported for Quality assessment	Outcome selected for External Panel Assessments	Number of RoB assessors in Cochrane Review	Interrater-reliability in Cochrane review
Effing et al., 2009 [33]	Self-management education for patients with chronic obstructive pulmonary disease	Cardiorespiratory PT	6	Not specified	HRQoL	2	Not reported
Sirtori et al., 2009 [36]	Constraint-induced movement therapy for upper extremities in stroke patients	Neurology PT	6	All outcomes	Disability post-intervention	2	Not reported
Taylor et al., 2010 [35]	Home-based versus centre-based cardiac rehabilitation	Cardiorespiratory PT	11	All outcomes	Exercise capacity	1	Not reported
Rutjes et al., 2010 [30]	Transcutaneous electrostimulation for osteoarthritis of the knee	Musculoskeletal PT	12	Pain and Function	Pain	2	Not reported
Orozco et al., 2008 [39]	Exercise or exercise and diet for preventing type 2 diabetes mellitus	General PT (Chronic conditions)	7	All outcomes	Fasting plasma glucose (mmol/L)	2	Not reported
Kramer et al., 2010 [38]	Aerobic exercise for women during pregnancy	Gynaecological PT	6	All outcomes	Birth weight	1	Not reported
Harvey et al., 2010 [25]	Continuous passive motion following total knee arthroplasty in people with arthritis	Musculoskeletal PT	5	Not specified	Active knee ROM	2	Not reported
Rutjes et al., 2010 [29]	Therapeutic ultrasound for osteoarthritis of the knee or hip	Musculoskeletal PT	4	Pain and Function	Pain	2	Not reported
Handoll et al., 2009 [24]	Multidisciplinary rehabilitation for older people with hip fractures	Musculoskeletal PT	7	Function and HRQoL	HRQoL	2	Not reported
Katalinic et al., 2010 [26]	Stretch for the treatment and prevention of contractures	Musculoskeletal PT	7	Not specified	Joint mobility	2	Not reported
Davies et al., 2010 [54]	Exercise based rehabilitation for heart failure	Cardiorespiratory PT	9	All outcomes	HRQoL	1	Not reported
Manheimer et al., 2010 [27]	Acupuncture for peripheral joint osteoarthritis	Musculoskeletal PT	8	Pain and Function (WOMAC)	Pain	2	Not reported
Schaafsma et al., 2011 [31]	Physical conditioning programs for improving work outcomes in workers with back pain	Musculoskeletal PT	5	All outcomes	Time to return to work	2	Between the 2 assessors, there was an average of one or two items of disagreement for every study.
Ostelo et al., 2011 [28]	Rehabilitation after lumbar disc surgery	Musculoskeletal PT	3	All Outcomes	Pain	2	Not reported
Fransen M., et al., 2009 [23]	Exercise for osteoarthritis of the hip	Musculoskeletal PT	5	All outcomes	Pain	3	Not reported
Puhan MA et al., 2010 [34]	Pulmonary rehabilitation following exacerbations of chronic obstructive pulmonary disease	Cardiorespiratory PT	5	HRQoL, Hospital Admission And Walk test	HRQoL, HRQoL	2	Not reported
States et al., 2009 [37]	Overground physical therapy gait training for chronic stroke patients with mobility deficits	Neurology PT	6	Not specified	Gait speed	3	The authors demonstrated 100% agreement on the Cochrane items.

HRQoL = Health-related quality of life; PT = physical therapy; RCT = randomized controlled trial; RoB = Risk of Bias; ROM = range of motion; WOMAC = Western Ontario and McMaster Universities Arthritis Index. doi:10.1371/journal.pone.0096920.t001

Table 2. Reliability between Cochrane Reviewers and External Panel and Reliability for the External Panel.

Domain	N trials included in Analysis	Reliability External Panel vs. Cochrane Reviewers			Within External Panel Reliability		
		Kappa	95%CI	Classification	Kappa	95%CI	Classification
Sequence generation	109	0.62	0.46, 0.76	good agreement	0.71	0.58, 0.84	good agreement
Allocation concealment	108	0.30	0.12, 0.43	slight agreement	0.79	0.65, 0.93	good agreement
Blinding of participants and personnel	86	0.15	0.04, 0.24	poor agreement	0.56	0.54, 0.59	fair agreement
Blinding outcome assessment	97	0.41	0.37, 0.48	slight agreement	0.54	0.36, 0.62	fair agreement
Incomplete data	96	0.24	0.07, 0.32	slight agreement	0.71	0.69, 0.73	good agreement
Selective reporting	87	0.13	0.05, 0.32	poor agreement	0.50	0.29, 0.63	fair agreement
Other bias	78	-0.04	-0.08, 0.04	poor agreement	0.32	0.25, 0.42	fair agreement
Overall rating	108	0.02	-0.06, 0.06	poor agreement	0.55	0.40, 0.70	fair agreement

doi:10.1371/journal.pone.0096920.t002

and under-estimation of treatment effects, since bias ratings obtained by different research group can lead to different results. For example, authors of meta-epidemiological studies [8,39,40], have taken information from external sources (Cochrane assessments, or information provided by authors of reviews). Although using data reported in the reviews, it is a practical and cost-efficient way to obtain information, authors should be aware that these evaluations may be inconsistent and prone to bias due to many factors such as expertise, training, level of education, and other characteristics of reviewers making quality judgements.

Very low agreements among Cochrane reviewers and the external panel were obtained for allocation concealment, blinding of participants, blinding of outcome assessment, and incomplete data. These features of a trial can have a substantial impact on the estimates of treatment effect [5,9,40–42]. **Some studies, for example, have found that** inadequate allocation concealment or lack of double-blinding can overestimate treatment effects on average by 18% and 9%, respectively [5,40,42]. **Nevertheless, other studies have found that trials with adequate allocation concealment and blinding had higher treatment effects than trials that did not accomplish with these methodological features.** [43,44] Similarly, effect sizes from trials that excluded dropouts in the analysis or considered a modified intention to treat (ITT) approach were more likely to show a beneficial effect than trials without exclusions, demonstrating that the ITT principle is important to preserve the benefits of randomization and keep unbiased estimates [45–47]. Overestimates of treatment effects, or bias, at the trial level, can lead to biased or inaccurate results and conclusions in systematic reviews and meta-analyses [40,41,48–50]. In addition, our analyses showed no agreement between decisions made based on RoB assessments at the level of meta-analysis. This means that both groups of reviewers did not agree in the overall quality of the evidence at the meta-analysis level. These factors can ultimately have repercussions on decision-making and quality of patient care since different assessments could lead to different decisions for clinical practice. Therefore, is alarming that the disagreements obtained between the two panels of reviewers are worse when it matters most.

The selection of different outcomes for RoB assessments may have influenced the poor agreement between Cochrane reviewers and a panel of blinded external reviewers. The majority of Cochrane reviews analyzed did not clearly specify the outcome used for RoB assessments. This directly reduces reproducibility of RoB assessment for outcome-dependent domains of the tool. Cochrane reviewers should report RoB assessments separately for each outcome analyzed, or at least for the main outcomes of the review. Half of the systematic reviews included in this study combined all outcomes into a single bias assessment and therefore, it is uncertain for which outcome the RoB assessments were applicable. Cochrane reviewers should clearly state which outcomes were used to perform the RoB assessments, in order to allow reproducibility and comparison.

The RoB has been extensively used by many Cochrane reviews, albeit the information of the inter-rater reliability of RoB is rather limited. To date, five studies [9,10,12,13,51] have investigated the inter-rater reliability of the RoB. One of them [51] did not use the generic RoB tool but a 12-item modified version of the tool developed by the Cochrane Back Review Group. The four other studies were conducted by the same group of researchers. When our inter-rater reliability results for the RoB tool were compared to those of other studies, most kappa values for the RoB domains were similar, except for allocation concealment, incomplete data, selective reporting, and overall rating of the RoB tool. Our kappa

Table 3. Comparison of Overall ratings at the meta-analysis level between external panel and Cochrane reviewers.

Meta-Analysis	Overall RoB meta-analysis level External Panel reviewers	Overall RoB meta-analysis level Cochrane reviewers
Agreements		
Orozco, 2008	unclear	unclear
Sirtory, 2009	unclear	unclear
Davies, 2010	unclear	unclear
Disagreements		
States, 2009	unclear	high
Fransen, 2009	high	low
Handoll, 2009	high	unclear
Effing, 2009	high	unclear
Taylor, 2010	unclear	high
Harvey, 2010	unclear	high
Rutjes, 2010	unclear	high
Katalinic, 2010	unclear	high
Puhan, 2010	high	high
Kramer, 2010	unclear	high
Rutjes, 2010b	unclear	high
Manheimer, 2010	unclear	high
Ostelo, 2011	unclear	high
Schaafsma, 2011	unclear	high

doi:10.1371/journal.pone.0096920.t003

values were much higher than those reported in previous studies (Table 4). We suggest a variety of reasons for these differences. Although we used the Cochrane Handbook guidelines for RoB assessments, we pre-defined specific decision rules to assess the individual domains of the tool. For example, the item of allocation concealment was scored low only when studies used central allocation (including telephone, web-based and centre controlled randomization) or when envelopes with three adequate safeguards were used (sequentially numbered, opaque, and sealed envelopes). If all three safeguards were not described, the item was scored as “unclear”. In addition to the Cochrane guidelines, the RoB item of incomplete data was rated “low” when intention to treat was conducted and the drop-out rate was less than or equal to 20%. When the drop-out rate was higher than 20%, the item was scored as “high” risk of bias since there is evidence that drop-out rates higher than 20% are likely to increase bias in treatment estimates [52,53].

Similarly, we created a precise decision rule for the item of selective reporting, and identified a clear cut off to determine low, unclear and high RoB categories. It is likely that all of these decision rules may have increased the inter-reliability between the blinded external reviewers in the RoB assessments for these domains.

Final ratings of the RoB tool based on the Cochrane reviewers assessments indicated that almost 92% of trials included in the reviews had either high or unclear RoB; a proportion that is similar to those identified in other studies [10,13]. As expressed by other researchers [13], the large number of trials classified as high or unclear RoB casts doubts about the discrimination power of the RoB tool to differentiate between studies with different levels of risk of bias that can explain variability of treatments effects across studies and inform accurately practice based on these assessments. **Thus, it is important to highlight that the overall assessment of the RoB may not be useful to determine**

quality of individual trials. We used the guidelines established by the Cochrane handbook to determine overall RoB. However, these criteria can be considered arbitrary and may not be appropriate. In addition, the items included in the RoB may be insufficient to represent the construct of interest: “Risk of bias”. Other items not considered in this tool may need to be added to provide a more comprehensive evaluation. Some scales commonly used to evaluate the quality of research (e.g. the Jadad scale) use only a limited number of items (3) and have been criticized for their inability to distinguish among good and bad quality studies [54]. This may be a similar problem for the RoB, which may not include all important factors to evaluate the full construct of “risk of bias”. Empirical evidence supports the evaluation of randomization, allocation concealment and blinding of clinical trials, all of which are included in the RoB tool. While there is insufficient evidence to support other domains being included, other methodological factors could be important for evaluating RoB and could be considered for inclusion in the RoB tool after careful empirical evidence testing.

It is recommended that RoB assessments are made by multidisciplinary groups of reviewers, in which epidemiologists, methodologists, and clinicians with expertise in the content area of the review participate in the assessments. Our panel of reviewers had different levels of expertise, with two reviewers having at least 10 years of expertise in performing quality assessments and two of them with expertise in the area of the physical therapy. This might explain in part our higher levels of reliability compared to other studies.

When junior researchers are involved in RoB assessments, it is crucial that training in concepts and guidelines for assessing study bias is provided prior to the start of the review [4]. Training should be intense and monitored in each stage of the review. Previous studies have trained reviewers using an average of 5 trials per study. In contrast, we used 10 studies for training purposes and

Table 4. Inter-rater reliability (kappa values) of the RoB tool reported in the scientific literature.

RoB Domains	Current Study	Hartling et al., 2011 [13]	Hartling et al., 2012 [14]	Hartling et al., 2009 [9]	Graham et al., 2012 [50]
Sequence generation	0.71	0.86	0.79	0.74	0.66
Allocation concealment	0.79	0.54	0.24	0.5	0.76
Blinding of participants and personnel	0.56	0.62	0.33	0.35	0.64
Blinding outcome assessment	0.54	0.62	0.33	0.35	0.5
Incomplete data	0.71	0.44	0.34	0.32	–
Selective reporting	0.50	0.40	0.27	0.13	–
Other bias	0.32	0.52	0.24	0.31	–
Overall rating	0.55	0.41	0.21	0.27	–
Assessment characteristics					
Type of trials	PT trials	Asthma trials	General health	Paediatric trials	Cervical/rehab trials
Number of trials	109	107	154	163	18
Trial evaluation specific to a single SR	No	Yes	No	No	NR
Number of trials used in pilot/training phase	10	??	5	5	NR
Number of reviewers	6	??	12	5	NR
Reviewers expertise	Physical therapy (2), methodology (6), public health (1), dentistry (1) and health related sciences (2). Doctorate (2), PhD candidate (1), Master level (1), undergraduate level (2)	NR	Doctorate (3); Master degree in health (8), epidemiology (1), undergraduate (1)	NR	Clinicians (physiotherapists, chiropractors, physicians), and statistician
Experience time conducting quality assessments	4 months – 10 years	NR	2–10 years	NR	5–50 years
Reviewers with formal training in SRs	3	NR	10	NR	NR
Outcomes used for RoB evaluation	Different outcome measures	Very specific outcomes	Different outcome measures	Different outcome measures	NR

RoB = Risk of Bias; SR = systematic review.
doi:10.1371/journal.pone.0096920.t004

held regular meetings to discuss bias ratings of common papers. These factors may have helped to obtain acceptable levels of reliability between the external reviewer panel for most of the domains of the RoB tool.

Limitations

This study restricted the analysis to a limited number of Cochrane systematic reviews in PT and therefore, the results might not reflect the inter-rater agreement of the RoB tool when applied to Cochrane reviews conducted in other areas of research, or to systematic reviews conducted out of the Cochrane Collaboration. Future studies should further assess potential differences in the inter-rater reliability of the RoB tool by comparing bias ratings of Cochrane reviews and non-Cochrane reviews versus those of independent panels of reviewers.

Future directions

The reliability of RoB assessments applied to clinical trials in systematic reviews needs to be improved. The creation of an international database (a bias assessment bank) in which a

qualified panel of experts (with extensive years of experience in trial methodology and critical appraisal of the scientific literature) contribute with independent RoB assessments of RCTs in a variety of clinical areas would be a promising step in that direction. Thus, researchers conducting systematic reviews and meta-epidemiological studies can use this data bank as a gold standard resource for RoB assessments. It is imperative that if an RoB assessment bank is created, contributors have the proper qualifications and experience to obtain less biased RoB assessments.

Conclusions

As far of our knowledge, this study is the first to demonstrate that risk of bias assessments of RCTs using the RoB tool are not consistent across different research groups contrasting results from Cochrane reviewers with an independent external panel of reviewers. Poor agreement was not only demonstrated at the trial level but also at the meta-analysis level. These results have important implications for decision making since different recommendations can be reached depending on the group

analyzing the evidence. Improved guidelines to apply the RoB tool and revisions to the tool for different health areas are needed. In addition, empirical evidence supporting additional items for the RoB tool needs to be developed. A call is made for the creation of a bank of RoB assessments of trial data, maintained by methodological and clinical experts that can be used as a reliable gold standard resource for RoB assessments. (4453 Words)

Supporting Information

Appendix S1 Search strategy to identify systematic review in physical therapy from the Cochrane Library of Systematic Reviews.

(DOC)

Appendix S2 Guidelines for evaluating the Risk of Bias in PT trials.

(DOC)

Appendix S3 Frequency of responses between Cochrane reviewers and the external panel of reviewers by RoB Domain.

(DOC)

Acknowledgments

Dr. Susan Armijo-Olivo is supported by the Canadian Institutes of Health Research (CIHR) through a full-time Banting fellowship, the Alberta Innovates Health solution through an incentive award, the STIHR

Training Program from Knowledge Translation (KT) Canada, and the University of Alberta.

Dr. Greta Cummings has been funded both provincially with a Population Health Investigator award from the *Alberta Heritage Foundation for Medical Research* (2006–2013), and nationally with a New investigator award from the Canadian Institutes of Health Research (2006–2011). Currently, she holds a Centennial Professorship at the University of Alberta (2013–2020).

Dr. Fuentes is supported by the Government of Chile, University of Alberta through a Dissertation fellowship, and the University Catholic of Maule.

Dr. Humam Saltaji is supported through a Clinician Fellowship Award by Alberta Innovates - Health Solutions (AIHS), the Honorary Izaak Walton Killam Memorial Scholarship by the University of Alberta and the WCHRI Award by the Women and Children's Health.

Research Institute (WCHRI).

In addition, the authors of this study thank the Alberta Research Center for Health Evidence (ARCHE) at the University of Alberta and all research assistants who helped with data collection.

Author Contributions

Conceived and designed the experiments: SA-O. Contributed reagents/materials/analysis tools: GGC ME BRdC. Wrote the paper: SA-O. Contributed to data collection: SA-O JF MO CH HS. Contributed data analysis, and interpretation: SA-O JF MO CH HS. Critically revised the manuscript and provided final approval of the version to be published: SA-O MO BRdC ME HS JF CH GGC. Provided feedback on the concept and research design and participated in interpretation of data: GGC ME BRdC.

References

- Armijo-Olivo S, Macedo LG, Gadotti IC, Fuentes J, Stanton T, et al. (2008) Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review. *Physical Therapy* 88: 156–175.
- Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, et al. (1995) Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials* 16: 62–73.
- Higgins J, Altman DG (2008) Chapter 8: Assessing risk of bias in included studies. In: Higgins J, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.00 [updated February 2008] version 5.0 ed: Available from www.cochrane-handbook.org, February, 2008.
- Lundh A, Goetzsche PC (2008) Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC Medical Research Methodology* 8.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 273: 408–412.
- Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC (1990) An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials* 11: 339–352.
- Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, et al. (1999) Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technology Assessment (Winchester, England)* 3: i–iv.
- Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, et al. (2012) Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Annals of Internal Medicine* 157: 429–438.
- Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, et al. (2009) Risk of bias versus quality assessment of randomised controlled trials: Cross sectional study. *BMJ* 339: 1017.
- Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG (2012) Assessment of study quality for systematic reviews: A comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: Methodological research. *Journal of Evaluation in Clinical Practice* 18: 12–18.
- Boutron I, Ravaud P (2012) Classification systems to improve assessment of risk of bias. *Journal of Clinical Epidemiology* 65: 236–238.
- Hartling L, Bond K, Vandermeer B, Seida J, Dryden D, et al. (2011) Applying the Risk of Bias tool in a systematic review of combination longacting betaagonists and inhaled corticosteroids for persistent asthma. *PLoS Medicine* 6.
- Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, et al. (2012) Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology* 66: 973–981.
- World Confederation for Physical Therapy (2011) Position statement: standards of physical therapy practice. *World Confederation for Physical Therapy*. 1–45 p.
- Higgins JPT, Altman DG, Goetzsche PC, Juni P, Moher D, et al. (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343.
- Furlan AD, Pennick V, Bombardier C, Van Tulder M (2009) 2009 Updated method guidelines for systematic reviews in the cochrane back review group. *Spine* 34: 1929–1941.
- Higgins J, Altman D (2008) Chapter 8: Assessing risk of bias in included studies. In: Higgins J, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions* version 5.0. Chichester, UK: John Wiley & Sons, Ltd.
- Schuck P (2004) Assessing Reproducibility for Interval Data in Health-Related Quality of Life Questionnaires: Which Coefficient Should Be Used? *Quality of Life Research* 13: 571–586.
- Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70: 213–220.
- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.
- Byrt T (1996) How good is that agreement? *Epidemiology (Cambridge, Mass)* 7: 561.
- Fransen M, McConnell S, Hernandez MG, Reichenbach S (2009) Exercise for osteoarthritis of the hip. *Cochrane Database of Systematic Reviews: Reviews* 2009, Issue 3 JohnWiley& Sons, Ltd Chichester, UK DOI: 101002/14651858CD007.
- Handoll-Helen HG, Cameron ID, Mak-Jenson CS, Finnegan TP (2009) Multidisciplinary rehabilitation for older people with hip fractures. *Cochrane Database of Systematic Reviews* 2009, Issue 4 JohnWiley& Sons, LtdChichester, UK DOI: 101002.
- Harvey LA, Brosseau L, Herbert RD (2010) Continuous passive motion following total knee arthroplasty in people with arthritis. *Cochrane Database of Systematic Reviews* 2010, Issue 3 JohnWiley& Sons, LtdChichester, UK DOI: 101002/14.
- Katalinic OM, Harvey LA, Herbert RD, Moseley AM, Lannin NA, et al. (2010) Stretch for the treatment and prevention of contractures. *Cochrane Database of Systematic Reviews* 2010, Issue 9 JohnWiley& Sons, LtdChichester, UK DOI: 101002/14 2010.
- Manheimer E, Cheng K, Linde K, Lao L, Yoo J, et al. (2010) Acupuncture for peripheral joint osteoarthritis. *Cochrane Database of Systematic Reviews: Reviews* 2010 Issue 1 JohnWiley& Sons, Ltd Chichester, UK.
- Ostelo-Raymond WJG, Costa-Leonardo OP, Maher CG, de-Vet-Henrica CW, van-Tulder MW (2008) Rehabilitation after lumbar disc surgery. *Cochrane Database of Systematic Reviews* 2008, Issue 4 JohnWiley& Sons, LtdChichester, UK.
- Rutjes-Anne WS, Nuesch E, Sterchi R, Juni P (2010) Therapeutic ultrasound for osteoarthritis of the knee or hip. *Cochrane Database of Systematic Reviews* 2010, Issue 1 JohnWiley & Sons, LtdChichester, UK DOI: 101002/14651858CD00.

29. Rutjes-Anne WS, Nuesch E, Sterchi R, Kalichman L, Hendriks E, et al. (2009) Transcutaneous electrostimulation for osteoarthritis of the knee. *Cochrane Database of Systematic Reviews* 2009: Issue 4 John Wiley & Sons, Ltd Chichester CD002823.
30. Schaafsma F, Schonstein E, Whelan KM, Ulvestad E, Kenny DT, et al. (2010) Physical conditioning programs for improving work outcomes in workers with back pain. *Cochrane Database of Systematic Reviews* 2010, Issue 1 John Wiley & Sons, Ltd Chichester, UK.
31. Davies P, Taylor F, Beswick A, Wise F, Moxham T, et al. (2010) Promoting patient uptake and adherence in cardiac rehabilitation. *Cochrane Database of Systematic Reviews* 2010, Issue 7 John Wiley & Sons Ltd Chichester, UK.
32. Effing T, Monninkhof-Evelyn EM, Valk-Paul PDL, Zielhuis-Gerhard GA, Walters EH, et al. (2007) Self-management education for patients with chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews* 2007, Issue 4 John Wiley & Sons Ltd Chichester, UK.
33. Puhan MA, Gimeno SE, Scharplatz M, Troosters T, Walters EH, et al. (2009) Pulmonary rehabilitation following exacerbations of chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews* 2009, Issue 1 John Wiley & Sons Ltd Chichester, UK.
34. Taylor RS, Dalal H, Jolly K, Moxham T, Zawada A (2010) Home-based versus centre-based cardiac rehabilitation. *Cochrane Database of Systematic Reviews* 2010, Issue 1 John Wiley & Sons, Ltd Chichester, UK. DOI: 101002/14651858CD.
35. Sirtori V, Corbetta D, Moja L, Gatti R (2009) Constraint-induced movement therapy for upper extremities in stroke patients. *Cochrane Database of Systematic Reviews* 2009, Issue 4 John Wiley & Sons, Ltd Chichester, UK.
36. States RA, Pappas E, Salem Y (2009) Overground physical therapy gait training for chronic stroke patients with mobility deficits. *Cochrane Database of Systematic Reviews* 2009, Issue 3 John Wiley & Sons, Ltd Chichester, UK.
37. Kramer MS, McDonald SW (2006) Aerobic exercise for women during pregnancy. Kramer Michael S, McDonald Sheila W Aerobic exercise for women during pregnancy *Cochrane Database of Systematic Reviews: Reviews* 2006 Issue 3 John Wiley & Sons, Ltd Chichester, UK. DOI: 101002/14651858CD000180pub2.
38. Orozco LJ, Buchleitner AM, Gimenez PG, Figuls M, Richter B, et al. (2008) Exercise or exercise and diet for preventing type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews* 2008, Issue 3 John Wiley & Sons Ltd Chichester, UK.
39. Egger M, Juni P, Bartlett C, Hohenstein F, Sterne J (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health technology assessment* 7: 1–76.
40. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, et al. (2008) Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ* 336: 601–605.
41. Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, et al. (2007) Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *International Journal of Epidemiology* 36: 847–857.
42. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, et al. (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? [see comment]. *Lancet* 352: 609–613.
43. Villari P, Manzoli L, Boccia A (2004) Methodological quality of studies and patient age as major sources of variation in efficacy estimates of influenza vaccination in healthy adults: A meta-analysis. *Vaccine* 22: 3475–3486.
44. Manzoli L, Schioppa F, Boccia A, Villari P (2007) The efficacy of influenza vaccine for healthy children: A meta-analysis evaluating potential sources of variation in efficacy estimates including study quality. *Pediatric Infectious Disease Journal* 26: 97–106.
45. Nuesch E, Trelle S, Reichenbach S, Rutjes AWS, Burgi E, et al. (2009) The effects of excluding patients from the analysis in randomised controlled trials: Meta-epidemiological study. *BMJ* 339: 679–683.
46. Abraha I, Montedori A (2010) Modified intention to treat reporting in randomised controlled trials: Systematic review. *BMJ (Online)* 341: 33.
47. Armijo-Olivo S, Warren S, Magee D (2009) Intention to treat analysis, compliance, drop-outs and how to deal with missing data in clinical research: a review. *Physical Therapy Reviews* 14: 36–49.
48. Hewitt CE, Kumaravel B, Dumville JC, Torgerson DJ (2010) Assessing the impact of attrition in randomized controlled trials. *Journal of Clinical Epidemiology* 63: 1264–1270.
49. Kjaergard LL, Als-Nielsen B (2002) Association between competing interests and authors' conclusions: Epidemiological study of randomised clinical trials published in the *BMJ*. *British Medical Journal* 325: 249–252.
50. Trowman R, Dumville JC, Torgerson DJ, Cranny G (2007) The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. *Journal of Clinical Epidemiology* 60: 1229–1233.
51. Graham N, Haines T, Goldsmith CH, Gross A, Burnie S, et al. (2011) Reliability of three assessment tools used to evaluate randomized controlled trials for treatment of neck pain. *Spine*.
52. Unnebrink K, Windeler J (2001) Intention-to-treat: Methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Statistics in Medicine* 20: 3931–3946.
53. Wright CC, Sim J (2003) Intention-to-treat approach to data from randomized controlled trials: A sensitivity analysis. *Journal of Clinical Epidemiology* 56: 833–842.
54. Herbison P, Hay-Smith J, Gillespie WJ (2006) Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology* 59: 1249.e1241–1249.e1211.