



**UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI MEDICINA E CHIRURGIA**

**DOCTORAL PROGRAM IN
EPIDEMIOLOGY, ENVIRONMENT AND PUBLIC HEALTH**

XXXI Cycle

Department of Clinical Sciences and Community Health

**“Investigating Innovative Evidence Synthesis Methods:
the Trial Sequential Analysis, the GRADE system
and the Network Meta-Analysis”**

PhD candidate: **CASTELLINI Greta**

Matricola N°: R11431

Tutor: Prof. Francesco **AUXILIA**

Prof. Lorenzo **MOJA**

Coordinator: Prof. Carlo **LA VECCHIA**

Academic Year 2018/2019

Index

LIST OF ABBREVIATIONS	4
EXECUTIVE SUMMARY.....	5
BACKGROUND AND RATIONALE	7
The systematic review and its role.....	7
Issues in meta-analysis: power, precision of findings and multiple comparisons	8
RCT as unit of analysis: power and relevance of findings	10
Research questions	12
Organization of the dissertation	13
SECTION 1.....	14
ARE RCTS INCLUDED IN SYSTEMATIC REVIEWS AND META-ANALYSIS ADEQUATELY REPORTED IN TERMS OF POWER AND RELEVANCE OF FINDINGS?	14
Background.....	14
Reporting of power and sample size calculation	15
Reporting of statistical significance and clinical relevance	16
Aim.....	17
Chapter 1. Improving Power and Sample Size Calculation in Rehabilitation Trial Reports: A Methodological Assessment	18
Abstract.....	18
Method.....	19
Results.....	20
Discussion.....	26
Conclusions.....	29
Chapter 2. Rehabilitation interventions in randomized controlled trials for low back pain: proof of statistical significance often is not relevant	30
Abstract.....	30
Methods.....	31
Results.....	33
Discussion.....	40
Conclusions.....	43
General Conclusion.....	44
SECTION 2.....	45
WHICH IS THE LIKELIHOOD OF A META-ANALYSIS TO BE UNDERPOWERED, INCONCLUSIVE AND IMPRECISE?	45
Background.....	45
Aim.....	46
Chapter 1. Trial Sequential Analysis	47
The Trial Sequential Analysis – method.....	49
Limitation.....	56

Alternative methods to the TSA – Examples.....	57
Case analysis 1: TSA in low back pain rehabilitation	61
Case analysis 2: TSA in cardiovascular diseases.....	74
General conclusion.....	81
Chapter 2. Precision of results: a focus on the GRADE system	82
Introduction.....	82
GRADE and TSA in systematic reviews and meta-analysis	88
Imprecision Assessment: A Comparison between the GRADE System and the Trial Sequential Analysis.....	90
Abstract	90
Aim.....	91
Methods.....	91
Results.....	94
Discussion	101
Conclusions.....	104
General conclusion.....	105
SECTION 3.....	106
IS IT POSSIBLE TO SIMULTANEOUSLY COMPARE MULTIPLE INTERVENTIONS IN A SYSTEMATIC REVIEW?.....	106
Background	106
What is a network meta-analysis?.....	106
The graphical representation: network diagrams	109
Ranking of the intervention effects.....	110
Case example of NMA: effectiveness of treatments for LBP interventions.....	111
Methods.....	111
Preliminary Results: back pain and specific functional status at 1 week of follow up.....	118
Discussion	132
Conclusion	133
CONCLUSIONS	134
APPENDIX	137
Appendix 1.....	137
Appendix 2.....	139
Appendix 3.....	141
BIBLIOGRAPHY	145

List of abbreviations

CI: Confidence interval

DALYs: disability adjusted life-year

FU: follow up

GRADE: Grading of Recommendations Assessment, Development and Evaluation

IQR: Interquartile range

LBP: Low back pain

MD: Mean difference

NMA: Network meta-analysis

NSAIDs: Non-Steroidal Anti-Inflammatory Drugs

OIS: Optimal information size

PRISMA-P: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocol

RCT: Randomized controlled trial

RoB: Risk of bias

RRR: Relative risk reduction

SD: Standard deviation

SMD: Standardized mean difference

SoF: Summary of findings

SR: Systematic reviews

SUCRA: Surface under cumulative ranking area

TSA: Trial Sequential Analysis

Executive summary

Systematic reviews and meta-analysis are the most appropriate and preferred study designs to inform clinical decision-making. Patients, clinicians, researchers and policymakers rely on this type of design since it aggregates several studies at the same time, increasing internal and external validity. Furthermore meta-analyses can increase statistical validity, reaching higher levels of power and precision as compared to single studies. This often means better understanding the magnitude of the treatment effect, and identifying the best intervention in pairwise or network (multiple) comparisons.

Despite these advantages, meta-analyses also have limitations and shortcomings. Power and precision depend on several elements, such as size of studies, number of events, sample variability and underlying heterogeneity. These elements are precondition and cannot be modified by reviewers. If one or more of these elements are problematic, the meta-analysis will replicate the same problem, despite the severity of the problem might be diminished by the co-presence of multiple studies. For instance meta-analyses including several small studies will be prone to several biases, eg, small study effect and publication biases. Moreover actual studies often explore modest intervention effects, which are difficult to be identified: even limited perturbations of study data can result in biases that can hide or inflate intervention effects, sabotaging the decision-making process of health professionals.

Against this background, we first wondered if RCTs included in systematic reviews and meta-analysis are adequately reported in terms of power and relevance/conclusiveness of findings. Secondly we explored how detect and assess underpowered, inconclusive and imprecise meta-analyses, using and comparing two modern approaches - Trial Sequential Analysis and the GRADE (Grading of Recommendations Assessment, Development and Evaluation). Third, we explored new meta-analytic techniques to contrast multiple interventions through direct and indirect evidence, in an extreme attempt to solve major limitations of a priori literature and lack of head to head trials.

In order to answer these research questions, we moved from the unit of analysis of systematic reviews, the randomized controlled trial, to the methods to cumulate evidence. Inadequate reporting, underpowered meta-analyses and conflicting direct and indirect evidence beyond head-to-head comparisons can alter the clinical decision-making process unless proper assessment, analysis and critical interpretation are put in place.

My dissertation is organized in three main Sections:

- Section 1: We focused on how sample size calculations were reported in RCTs exploring the efficacy of low back pain rehabilitation interventions and, among those adequately reported, how findings were interpreted in terms of statistical significance and clinical relevance.
- Section 2: We explored Trial Sequential Analysis and the GRADE approach in several medical areas. We compared the agreement of the approaches in evaluating the overall quality of evidence.
- Section 3. We finally combined direct and indirect evidence on a sample of RCTs assessing rehabilitation interventions for low back pain through a network meta-analysis.

This dissertation focuses on innovative advanced methods used in evidence synthesis science. These methods are a partial answers to the need for precise and reliable results, standard and transparent methods to assess the body of the evidence, and comparisons of multiple interventions in addition of pair-wise comparisons.

Background and Rationale

The systematic review and its role

Clear and comprehensive summaries of medical literature information is needed to inform health professionals and policy makers about the best available treatments (Sackett, Rosenberg et al. 1996). Systematic reviews (SRs) and meta-analysis of randomized controlled trials (RCTs) are recognized by the scientific community to be the gold standard in demonstrating the efficacy of an intervention (Cook, Hislop et al. 2015). Systematic reviews have gained momentum in the medical field thanks to their efficiency in keeping up-to-date the accumulation of the evidence and to be a key document for clinical practice guideline developers that used them as a starting point for guideline production, identifying those treatments that insure the best payback of researches (Grimshaw, Eccles et al. 2012).

The Cochrane Collaboration defines a systematic review as “an attempts to collect all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question” (Higgins JPT and Green S 2011) using systematic methods to select, critically assess and provide synthesis of characteristics and findings of the studies included in the reviews (Oxman and Guyatt 1993). Indeed, a systematic review provides a qualitative, and possibly a quantitative, synthesis of the evidence. Other definitions consistently mention key elements central to systematic reviews: searching for evidence, appraising its quality and synthetizing the results across studies (Moher, Tetzlaff et al. 2007).

A qualitative synthesis is based on a comprehensive assessment of the methodological quality of trials included in the review, i.e. the study’s internal validity related to whether it answers its research question ‘correctly’, free from bias. Whereas, the quantitative synthesis is obtained by combining information from all relevant studies: a meta-analysis, a key element of most systematic reviews, is the statistical synthesis of the results of independent studies pooling the effect sizes of each study (Higgins JPT and Green S 2011).

Decision makers, clinicians, and patients have high expectations, i.e. results unbiased and consistent among reports (Whiting, Savovic et al. 2016). However, the inconsistent quality of systematic reviews and meta-analyses of RCTs that are published in the literature has been previously documented (Shea, Grimshaw et al. 2007, Gagnier and Kellam 2013, Tunis, McInnes et al. 2013, Gianola, Castellini et al. 2016). SRs are not immune to methodological flaws and can reach invalid and discordant

conclusions, indeed. Major flaws can be related to: (i) the quality of the included RCTs since the quality of a SR is strongly related to it, with primary studies at high risk of bias; then, (ii) different methods can be applied to same data, reaching contrasting conclusions: the statistical methods used generate quantitative synthesis can be underpowered, imprecise, unreliable without providing a conclusive evidence on a wide range of interventions comparisons.

Thus, even though SRs and RCTs are usually considered high quality study designs, the body of the evidence could be generally of low or very low quality so that insufficient definitive conclusions could be reached and reported (Buchbinder, Maher et al. 2015).

Quantitative synthesis of the evidence in a SR: the meta-analysis

The meta-analysis of RCTs is considered the best approach to identify the actual benefit of a health intervention. This statistical method have several advantages (Higgins JPT and Green S 2011):

To increase power. Power is the chance of detecting a real effect as statistically significant if it exists. Many individual studies are too small to detect small effects, but when several are combined there is a higher chance of detecting an effect.

- To improve precision of the effect of an intervention: when several studies are combined to detect a small effect, the chance of detecting it increases.
- To detect intervention effects when controversies arise from conflicting studies or to generate new hypotheses.

A meta-analysis can also take into consideration the risk of bias of each single study, or across studies, and random errors limiting and balancing their affection on results.

Nevertheless, like any tool, it can be misused and it might not guarantee valid and reliable results. Indeed, 'The best available evidence' may not be synonymous with 'sufficient evidence' or 'strong evidence'.

Issues in meta-analysis: power, precision of findings and multiple comparisons

Even if when systematic reviews and meta-analysis are methodologically great, most may still be not informative: "weak or insufficient evidence" is a common conclusion which makes the review not informative on what the best interventions is (Ioannidis 2016). Indeed, a meta-analysis is not

infallible: it has limitations, which have to be cautiously considered when interpreting the clinical findings.

It has been showed in literature that the majority of published Cochrane meta-analyses including RCTs might be underpowered (Brok, Thorlund et al. 2009, Thorlund, Imberger et al. 2011, Imberger, Gluud et al. 2015). Much attention should be paid to potentially overestimated or underestimated conclusions. When a meta-analysis has not enough statistical power to detect an effect, the lack of power and precision can amplify the chances of random error, leading neutral or negative (non-positive) findings (Brok, Huusom et al. 2012). Trial Sequential Analysis (TSA) is a frequentist methodology which can handle the risk of overestimating or underestimating clinical findings: it combines the calculation of a required information size for a meta-analysis and an adjusted threshold for a statistically significant treatment effect (Thorlund K, Wetterslev J et al. 2011). The information size required for more reliable and conclusive meta-analysis results may be assumed to be at least as large as the sample size of a single well-powered randomised clinical trial to detect or reject an anticipated intervention effect. TSA' conclusions have the potential to be more reliable than those using traditional meta-analysis techniques. The calculation of the information size and the adjusted significance thresholds can indeed eliminate early false positive findings due to imprecision and repeated significance testing in meta-analyses. Methods as TSA emphasize the fact that an accurate and critic assessment, both qualitative and quantitative, of the body of the evidence has become fundamental in clinical research. It has been progressively important to reliably provide the extent of the confidence (the uncertainty) of the benefit or harm of the effect of an intervention. A good clinical decision requires accurate estimation of uncertainty as much as we would like to reach a definitive conclusion. Therefore, communicate greater error more accurately than to infer less error inaccurately seems a better solution. Trial Sequential Analysis is useful to handle this uncertainty.

Another approach has become internationally and widely adopted for conveying the uncertainty associated with findings and conclusions: The Grading of Recommendations Assessment, Development and Evaluation (GRADE) system. The GRADE system is an international standard to assess the strength of body of the evidence, informing transparently and explicitly the confidence that researchers have on the results. (Guyatt, Oxman et al. 2011). GRADE uses a framework of information about risk of bias, imprecision, inconsistency, indirectness and publication bias. Particularly, they define the issue of imprecision for SRs as the confidence in the estimate of the overall effect (Guyatt, Oxman et al. 2011). Imprecision also encompasses the size or importance of an effect, and it is influenced by the magnitude of the sample size or the number of events. The assessment of this domain is complex since it requires to balance the magnitude of the effect derived

from meta-analysis results, the achievement of an optimal information size and the clinical relevance of the effect.

Both TSA and the GRADE approach, with the imprecision domain assessment, aim to define uncertainty of the evidence: clinicians and researchers need to find the most accurate method to be confident in clinical findings.

An additional key limitation of a meta-analysis is that it can compare only two treatments at a time, yielding only partially information that clinicians, patients and policy-makers need. Since usually more than two treatment options are available in a real clinical setting for certain conditions, a meta-analysis can offer just a part of the possible wealth of treatment options and might be not supportive for an optimal clinical decision-making. In the last decade, a new meta-analytic technique called network meta-analysis (NMA) has been developed to assess the effectiveness of several interventions and offer the synthesis of the evidence across a network of randomized trials. The advantage of this statistical method over standard pairwise meta-analysis is that it enables indirect comparisons of multiple interventions that have not been studied in a head-to-head fashion. The interest for this technique has been increased across researchers and over the time. The PRISMA Statement for Reporting of Systematic Reviews published an extension incorporating the guideline for reporting Network Meta-analyses of Health Care Interventions (Hutton, Salanti et al. 2015).

Over the last 20 years, the literature describing methods used to quantitatively summarize the evidence has rapidly expanded with the above-mentioned innovative techniques in order to overcome meta-analysis limitations. This emphasises an endless need to find and provide the best and the most reliable available evidence for the clinical decision making process and the health care planning.

RCT as unit of analysis: power and relevance of findings

A randomized controlled trial is the unit of analysis for reviews and meta-analyses addressing the efficacy of an intervention. Meta-analysis can include one or more RCTs relatively small and not powered enough to detect a modest intervention effect. Small RCTs tend to show greater intervention effects than larger studies leading to the so-called 'small-study effects'. Usually, meta-analysis with findings from small studies are more prone to publication bias (Schork 2003). Hypothetically, a meta-analysis should include adequately powered studies, both to get rid of publication bias and to discourage future researchers from conducting small studies (Turner, Bird et al. 2013).

It has been demonstrated that most meta-analyses published by Cochrane reviews are made up by underpowered studies: Turner et al reported that the 70% of 14,886 meta-analyses across different medical field included underpowered studies and trials' power was low across all areas (Turner, Bird et al. 2013). Thus, the robustness of the conclusions of a meta-analysis should be assessed with careful interpretation of trials results.

A review can draw conclusions about the efficacy and effectiveness of an intervention only collecting data and results from RCTs expected to be reliable and valid. The relevance of findings and quality of conduct in clinical trials can be judged almost only on the basis of what is reported in the published version of the article (Simera, Altman et al. 2008). Therefore, investigating the extent of good or poor reporting to evaluate the quality of trial conduction is essential to understand its validity and avoid distortions during evidence synthesis.

Unfortunately, not all published manuscripts provide the essential and necessary information that allow readers, researchers and clinicians to assess the methodological quality of studies and interpret the findings (Hopewell, Dutton et al. 2010, Hoffmann, Eructi et al. 2013). It has been demonstrated that, despite the existence of reporting guidelines for 21 years, there is still a suboptimal uptake and a not correct usage of reporting guidelines (Jin, Sanger et al. 2018). An inadequate reporting is linked to poorly conducted research and clinical findings resulted from methodologically flawed randomized trials are more likely to comprise biased or misleading estimates of treatment effects (Schulz, Chalmers et al. 1995, Simera, Altman et al. 2008).

The assessment of RCTs validity is an essential component when included in a review because it should influence the analysis, interpretation and conclusions of the review itself and be an unreliable unit of analysis for a meta-analysis.

Research questions

Based on this background, I have focused my PhD program on three research questions, moving from the unit of the analysis of the SR, i.e. the RCT, to the methods to pool the evidence, i.e. meta-analyses and modern techniques. The research questions are:

1. ***Are RCTs included in systematic reviews and meta-analysis adequately reported in terms of power and relevance of findings?***

Trials failing to detect a real difference between treatment effects may inflate the results of meta-analyses, obfuscating the clinical decision-making process. The possible lack of good quality in reporting of RCTs makes me unsure about not only the quality with which researchers report details in the published manuscripts but also whether authors perform an adequate sample size calculation at protocol stage. I wondered if RCTs included in Cochrane review are adequately powered and reported reliable results. Otherwise, if not, how can the estimates in the meta-analysis be adequate and precise?

(First year of my PhD program)

2. ***Which is the likelihood of meta-analysis to be underpowered, inconclusive and imprecise?***

Meta-analysis underpowered and imprecise exist. The TSA can handle the risk of overestimating or underestimating clinical finding. Moreover, the GRADE approach assess imprecision as part of the quality of the evidence.

The TSA and the GRADE approach both take into consideration the imprecision domain in systematic reviews. Consequently, I ended up investigating how TSA can yield a different interpretation of this domain in meta-analysis results compared with those obtained by the GRADE system.

(Second year of my PhD program)

3. ***Is it possible to simultaneously compare multiple interventions in a systematic review?***

Head-to-head comparisons limits the relevance of findings generated by a systematic review. Network meta-analyses allow to better understand potentials of modern evidence syntheses applied to the medical field.

(Third year of my PhD program)

Organization of the dissertation

This dissertation is organized into two sections:

Section 1: Research question 1.

We focused on a particular quality aspect: how sample size calculations were reported in RCTs published in low back pain rehabilitation and, among those adequately reported, how findings were interpreted in terms of statistical significance and clinical relevance.

Section 2: Research questions 2

We investigated key topics as uncertainty and imprecision in meta-analysis. We explored the Trial Sequential Analysis and the GRADE system in medical field. Thus, we showed how the above-mentioned techniques are implemented in a modern evidence synthesis generation.

Section 3. Research question 3.

How to combine direct and indirect evidence on a sample of RCTs in rehabilitation field through a network meta-analysis was deepened.

A general conclusion follows in the last section.

SECTION 1.

Are RCTs included in systematic reviews and meta-analysis adequately reported in terms of power and relevance of findings?

Background

Because of my background as physiotherapist, I addressed my first research question on RCTs published in the field of rehabilitation of mechanical low back pain where several gaps in conduction and reporting exist.

In many countries mechanical low back pain is one of the most common causes of disability and lost work days and it is ranked as the greatest contributor to global disability (years lived with disability) (Hoy, March et al. 2014, March, Smith et al. 2014). Mechanical low back pain imposes significant economic and social burdens. Although several interventions are used to treat mechanical low back pain, including medicines and rest, rehabilitation plays a central role. Therefore, many researchers have devoted time and effort to examining the efficacy and safety of various rehabilitation interventions aimed at decreasing the impact of this condition.

Although the number of reports of RCTs in rehabilitation has been increasing (Castellini, Gianola et al. 2016), most studies are empiric based on clinical observations with small sample sizes and inadequate reporting of essential information (Abdul Latif, Daud Amadera et al. 2011).

We demonstrated that dimensions as description of interventions or adequate reporting of the outcomes are poorly reported across RCTs included in Cochrane systematic reviews published on LPB rehabilitation (Gianola, Castellini et al. 2016, Gianola, Frigerio et al. 2016). Based on this background, other dimensions should be assessed to correctly interpret the magnitude and accuracy of the effect of an intervention and guarantee its validity and generalizability: patients and sample size, statistical significance and clinical relevance.

Over the last decade, several initiatives have promoted the accurate, complete, and transparent reporting of clinical studies to support research reproducibility and usefulness (Simera, Altman et al. 2008). The CONSORT (Consolidated Standards of Reporting Trials) Statement has been properly introduced in 1996 to promote the reporting of findings of randomized controlled trials. It offers a

standard minimum set of reporting items, as the population included in the RCT, the characteristics of the intervention, the performed comparison, the chosen outcome measure, the sample size calculation (Kessler 2002, Boutron, Moher et al. 2008). Unfortunately, not all published manuscripts provide the essential and necessary information that allow readers, researchers and clinicians to assess the methodological quality of studies and interpret the findings (Hopewell, Dutton et al. 2010, Hoffmann, Eructi et al. 2013). It has been highlighted that inadequate reporting is linked to poorly conducted research and clinical findings resulted from methodologically flawed randomized trials are more likely to comprise biased or misleading estimates of treatment effects (Schulz, Chalmers et al. 1995, Simera, Altman et al. 2008). As a consequence, investigating the extent of good or poor reporting is essential to evaluate the quality of trial conduction whose results can inflate the findings of a systematic review.

The quality of the reporting of the following domains were judged: description of the interventions, outcomes assessment, reporting of sample size calculation and reporting of statistical significance and clinical relevance. I have selected RCTs on the efficacy of interventions for low back pain (a non-pharmacologically intervention), where the reporting and description of the above-mentioned domains are considered fundamental.

I have participated at the conduction and publishing of two manuscripts on the reporting of description of intervention and outcomes (Gianola, Castellini et al. 2016, Gianola, Frigerio et al. 2016). Overall, it has been confirmed a lack of quality of reporting of these domains.

I found very interesting investigating how sample size calculation is stated and described in RCTs besides, how results are interpreted in terms of statistical significance and clinical relevance. Sample size calculation is essential to demonstrate that a trial is adequately designed to detect a likely real effect or association, if such exists, in a given population. To understand the validity of a RCT, the assumptions made in the power analysis should be reported in a transparent fashion.

Reporting of power and sample size calculation

Well-designed, properly executed RCTs provide the most reliable evidence on the effectiveness of health care interventions (Calvert, Blazeby et al. 2013). The validity of an RCT depends on several key factors that should be adequately reported: the sample size calculation is one of them. Sample size is related to statistical power, which derives from b error or type II error (Schulz and Grimes 2005, McKeown A 2015): it represents the likelihood of failure to reject the null hypothesis when, in

fact, it should be rejected. The investigator's aim is to minimize this type of error by increasing the sample size. Sample size calculation is essential in study design because a low-power study may fail to yield significant results and detect relevant clinical effects. Its description is fundamental in any published report so that readers can base their assessment on what is reported rather than rely on assumptions about how the study authors arrived at their results. However, sample size calculation is not always adequately reported (Ayeni, Dickson et al. 2012, Koletsi, Pandis et al. 2014, Rutterford, Taljaard et al. 2014). To ensure quality in trial conduction, the Consolidated Standards of Reporting Trials (CONSORT) 2010 statement recommends that authors provide a clear description of sample size calculation methods and assumptions as follows: the estimated outcomes in each group (minimum important treatment effect or effect size), the level of significance (α or type I error), the statistical power ($1 - \beta$ or type II error), and, for continuous outcomes, the assumed SD of the measurements (Antes 2010, Moher, Hopewell et al. 2010). In addition, the CONSORT guidelines recommend reporting the primary outcome using which important differences between 2 groups are determined. Authors should therefore decide and state a priori the fixed values for parameter assumptions. Although the number of reports of RCTs in rehabilitation has been increasing (Castellini, Gianola et al. 2016), most studies are based on clinical observations with small sample sizes and inadequate reporting of essential information (Abdul Latif, Daud Amadera et al. 2011).

Reporting of statistical significance and clinical relevance

As we already stated above, to understand the validity of RCT's findings, the assumptions made in the power analysis should be reported in a transparent fashion. We wondered if among those trials adequately reporting the sample size calculation, their findings were interpreted taking into consideration the clinical relevance required in the calculation.

Randomized controlled trials (RCTs) aim to show differences in an outcome measurement between two or more groups of patients undergoing different interventions (Hoffmann, Thomas et al. 2014). Authors of RCTs usually report findings in term of statistical significance (i.e., with a p-value < 0.05 the intervention is more effective than the comparison). However, the p-value indicates the chance of the observed effect, does not consider the magnitude of benefits (or harm) and indeed the clinical relevance. This is defined as the estimate of the smallest treatment effect between groups that people would consider important and is often called minimally important difference (MID)(Beaton, Bombardier et al. 2000).

In rehabilitation, RCTs are often based on small sample sizes (Abdul Latif, Daud Amadera et al. 2011) and most outcomes are patient-reported associated with small clinical changes (e.g., pain reduction from moderate to low). Recognizing small but clinically relevant effects requires clinical trials with large sample sizes. This scenario leads to two problems. First, when a small trial in rehabilitation achieves statistical significance, false positive outcomes may occur. Second, it should not be assumed that trial results which are statistically significant are also clinically relevant (Freiman, Chalmers et al. 1978). Even if the pre-specified value of success for the primary outcome has been met for the difference in treatment effects (usually a p-value of less than 0.05), it does not necessarily imply that the difference matters to patients (Wright 1996, van der Roer, Ostelo et al. 2006). For example, a recent re-analysis of data of a published Cochrane review on Multidisciplinary Biopsychosocial Rehabilitation (MBR) for LBP showed how findings were highly significant but irrelevant in practice. Across studies, pain was reduced by less than one-third of 1 MID unit on a numerical rating scale (0.27 MID units, confidence interval 0.07–0.48) (Gianola, Andreano et al. 2018). A MID of 2 points out of 10 is usually considered meaningful (Ostelo, Deyo et al. 2008).

Aim

The purpose of this section is to systematically assess the quality of reporting of power and sample size calculation in RCTs included in the Cochrane Systematic Reviews comparing mechanical LBP interventions.

Moreover, among those trials adequately reporting the elements of the sample size calculation, we assessed whether treatment effects of RCTs for LBP are both statistically significant and clinically relevant. We also investigated if trials were powered to achieve clinically relevant outcomes assessing the risk of possible false-negative results (i.e., missing an effect that is actually there).

We wanted to demonstrate that RCTs not adequately reported can reflect serious implications on research results as in meta-analysis.

Chapter 1. Improving Power and Sample Size Calculation in Rehabilitation Trial Reports: A Methodological Assessment

Published as: Improving Power and Sample Size Calculation in Rehabilitation Trial Reports: A Methodological Assessment. Castellini G, Gianola S, Bonovas S, Moja L.

Arch Phys Med Rehabil. 2016 Jul;97(7):1195-201.

Abstract

Objective: To systematically assess the reporting of sample size calculation in randomized controlled trials (RCTs) on rehabilitation interventions for mechanical low back pain.

Data Sources: The Cochrane Database of Systematic Reviews was searched through February 2015.

Study Selection: We conducted an electronic database search for RCTs published from January 1, 1968 to February 28, 2015 and included in the Cochrane Systematic Reviews.

Data Extraction: Two investigators independently used an ad hoc 6-item checklist derived from the Consolidated Standards of Reporting Trials (CONSORT) 2010 statement recommendations to extract data on sample size calculation. The primary outcome was the proportion of RCTs that reported sample size calculation; the secondary outcome was the completeness of sample size analysis reporting. We also evaluated improvement in reporting of sample size calculation over time.

Data Synthesis: Sample size calculation was reported in 80 (36.0%) of the 222 eligible RCTs included in 14 Cochrane Systematic Reviews. Only 13 (16.3%) of these RCT reports gave a complete description, and about half reported ≥ 4 of the 6 elements of sample size calculation (median, 4; interquartile range, 3-5). Completeness of reporting of sample size calculation improved from 1968 to 2013; since 2005, the number of RCTs reporting sample size calculation has increased compared with the number of RCTs not reporting it.

Conclusions: Despite improvement, reporting of sample size calculation and power analysis remains inadequate, limiting the reader's ability to assess the quality and accuracy of rehabilitation studies.

Archives of Physical Medicine and Rehabilitation 2016;97:1195-201

Method

Search strategy

We conducted an electronic database search for systematic reviews published between 1968 and February 2015 limited to the Cochrane Database of Systematic Reviews. Search terms back pain and rehabilitation were run in “title, abstract, keywords” search tab in advanced search strategy. We included a systematic review if the title or the abstract presented mechanical low back pain as the disease target and the intervention was rehabilitative, as defined by the National Library of Medicine. We did not take into account interventions other than therapeutic rehabilitation (i.e. prevention) or involving population subgroups (i.e. pregnancy). From the eligible systematic reviews, we extracted all included trials with a randomized study design and published in English, Italian, Spanish, or French. After removing duplicates of RCTs, 2 researchers (G.C., S.G.) independently screened the title and abstract of all potentially eligible RCTs. Disagreements were resolved by consensus.

Data extraction

We extracted the general characteristics of RCTs: year of publication, number of authors, first author’s geographic region (Europe, North and South America, Asia, Australia), journal that published the study, and funding source. We developed an ad hoc checklist derived from the CONSORT checklist to extract data on sample size calculation. The checklist was uploaded on DistillerSR, a web-based database for data management. We examined whether the RCT report included a power of the sample size calculation was compliant with CONSORT guidelines. Following the CONSORT checklist (Moher, Hopewell et al. 2010), we assessed the description for reporting of 6 sample size calculation elements: (1) type I or alpha error; (2) type II or beta error, or power; (3) assumption of the expected treatment effect of the intervention (ie, the difference between group means as effect size or minimal important difference and relative risk); and (4) the assumed variability expressed as an SD, a variance, or an intraclass correlation coefficient. We also looked for (5) the outcome on which sample size calculation was based and (6) whether there was an adjustment to accommodate attrition rate. In addition, we extracted from the Methods section the sample size planned (ie, as resulted from the sample size calculation procedure) and from the Results section the actual number of participants randomized (N) according to the CONSORT flow diagram. If there was no statement or CONSORT flow diagram reporting the number of patients randomized, we extracted it from implicit information (ie, “enrolled” or “included”). When articles reported the sample size

calculation, we examined whether there was a discrepancy between the sample size planned and the number of participants randomized. Moreover, we asked whether sample size reporting might be affected by the funding status of the RCT. Data extraction was performed independently by 2 reviewers (G.C., S.G.). Disagreements were resolved by consensus.

Statistical methods

Descriptive statistics are presented as median (interquartile range) or number (percentage), when appropriate. The nonparametric Wilcoxon matched-pair signed-rank test and the chi-square test were used for statistical evaluations. For hypothesis testing, a probability value of $<.05$ was considered statistically significant. All statistical tests were 2-sided. Stata statistical software was used for all statistical analyses.

Results

Study selection

We identified 14 relevant Cochrane systematic reviews in the Cochrane Library (Heymans, van Tulder et al. 2004, Urrutia, Burton et al. 2004, Hayden, van Tulder et al. 2005, Clarke, van Tulder et al. 2007, Furlan, Imamura et al. 2008, Khadilkar, Odebiyi et al. 2008, Yousefi-Nooraie, Schonstein et al. 2008, Henschke, Ostelo et al. 2010, Rubinstein, van Middelkoop et al. 2011, Rubinstein, Terwee et al. 2013, Wegner, Widyahening et al. 2013, Ebadi, Henschke et al. 2014, Kamper, Apeldoorn et al. 2014). Sixty of 301 RCTs included in these 14 systematic reviews were excluded because they were duplicates or multiple publications of the same RCT; 7 (2.3%) were excluded because their full text could not be retrieved; and 12 (3.9%) were excluded because they did not satisfy the language criterion. A total of 222 RCTs (73.7%) was included in our review (fig 1).

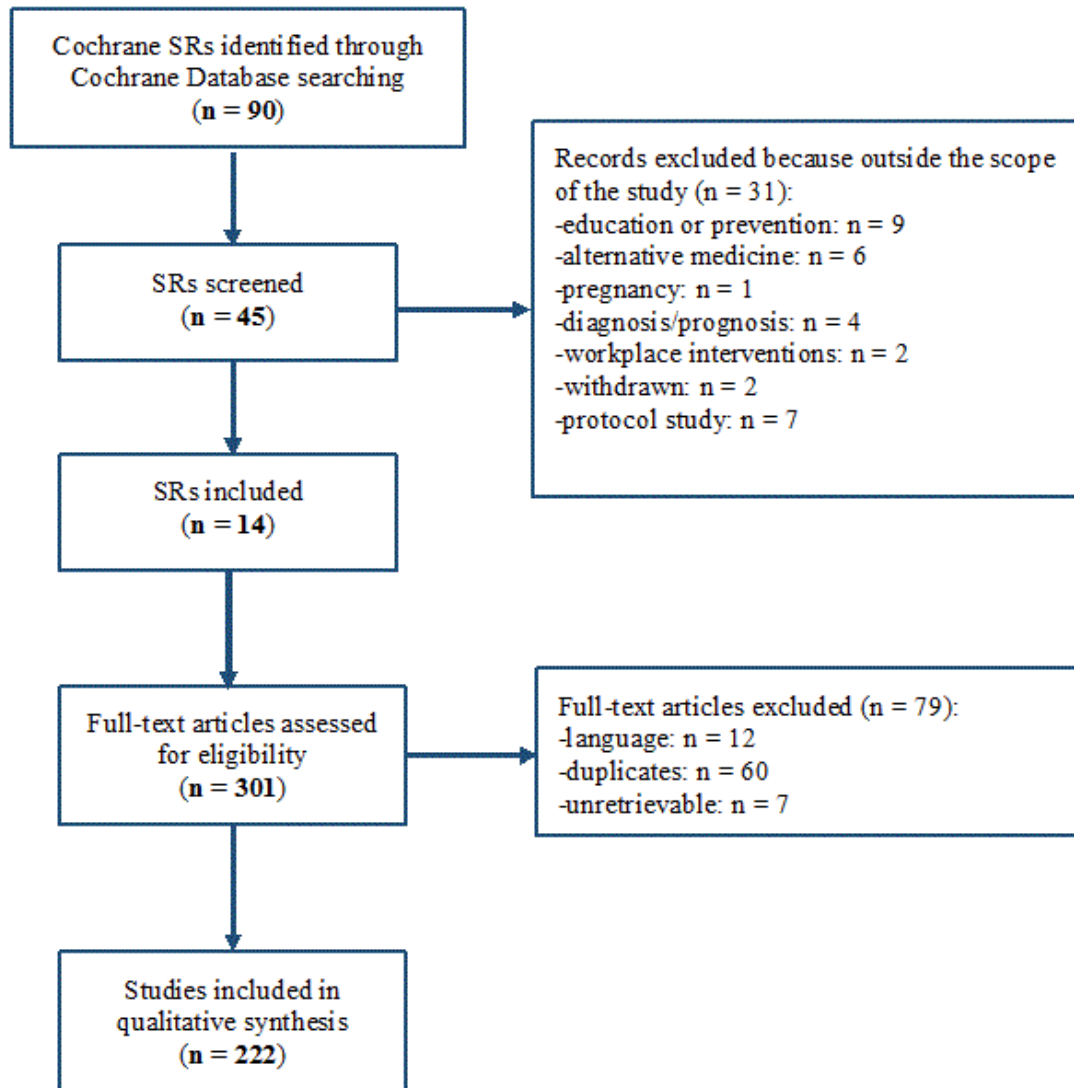


Figure 1. Flow diagram.

General characteristics

The 222 eligible RCT reports were published in 78 journals. Most were published in *Spine* (22.5%, n=50), followed by *Journal of Manipulative and Physiological Therapeutics* (4.5%, n=10), *Pain*, *British Medical Journal*, and *Archives of Physical Medicine and Rehabilitation* (4.1%, n=9), and *Clinical Journal of Pain* (3.6%, n=8). Some 32 countries were indicated as countries of publication, with the top 3 countries being the United States (18.9%, n=42), the United Kingdom (13.1%, n=29), and the Netherlands (9.9%, n=22); most studies were published (59.5%, n=132) by European researchers. The period of RCT publication was from 1968 to 2013. The characteristics of the RCTs are reported in **table 1**.

Table 1 General characteristics of the RCTs

Characteristic	Value
No. of countries	32
Top 9 countries	
United States	42 (18.9)
United Kingdom	29 (13.1)
The Netherlands	22 (9.9)
Norway	15 (6.8)
Sweden	14 (6.3)
Finland	12 (5.4)
Australia	10 (4.5)
Canada	10 (4.5)
Turkey	10 (4.5)
No. of journals	78
Most frequent journals	
<i>Spine</i>	50 (22.5)
<i>Journal of Manipulative and Physiological Therapeutics</i>	10 (4.5)
<i>Pain, British Medical Journal, Archives of Physical Medicine and Rehabilitation</i>	9 (4.1)
<i>Clinical Journal of Pain</i>	8 (3.6)
No funding reported	97 (43.7)
No. of authors	5 (1–12)
Year of publication of the trial report	2000 (1968–2013)

NOTE. Values are n (%) or median (interquartile range).

Sample size calculation

Reporting

Only 80 (36.0%) of the 222 RCTs reported sample size calculation. However, there was a significant improvement in reporting of sample size calculation over time (**fig 2**). We found that 13.3% (11 of 83) of trials published on or before 1996 reported sample size calculation compared with 49.6% (69 of 139) of trials published on or after 1997 ($\chi^2_1= 29.85$; $P<.001$). Furthermore, we found an association between reporting of a funding source and reporting of sample size calculation. In particular, 48.8% (61 of 125) of the trials reporting a funding source were also reporting a sample size calculation as compared with only 19.6% (19 of 97) of the trials not reporting a funding source ($\chi^2_1= 20.22$; $P<.001$). This association was strong in the post-CONSORT era, with 61.4% (54 of 88) of the trials reporting a funding source also reporting a sample size calculation versus 29.4% (15 of 51) of the RCTs not reporting a funding source ($\chi^2_1=13.19$; $P<.001$). However, it was not significant in the pre-CONSORT era (18.9% vs 8.7%; $\chi^2_1=1.86$; $P=.17$), but data were scarce.

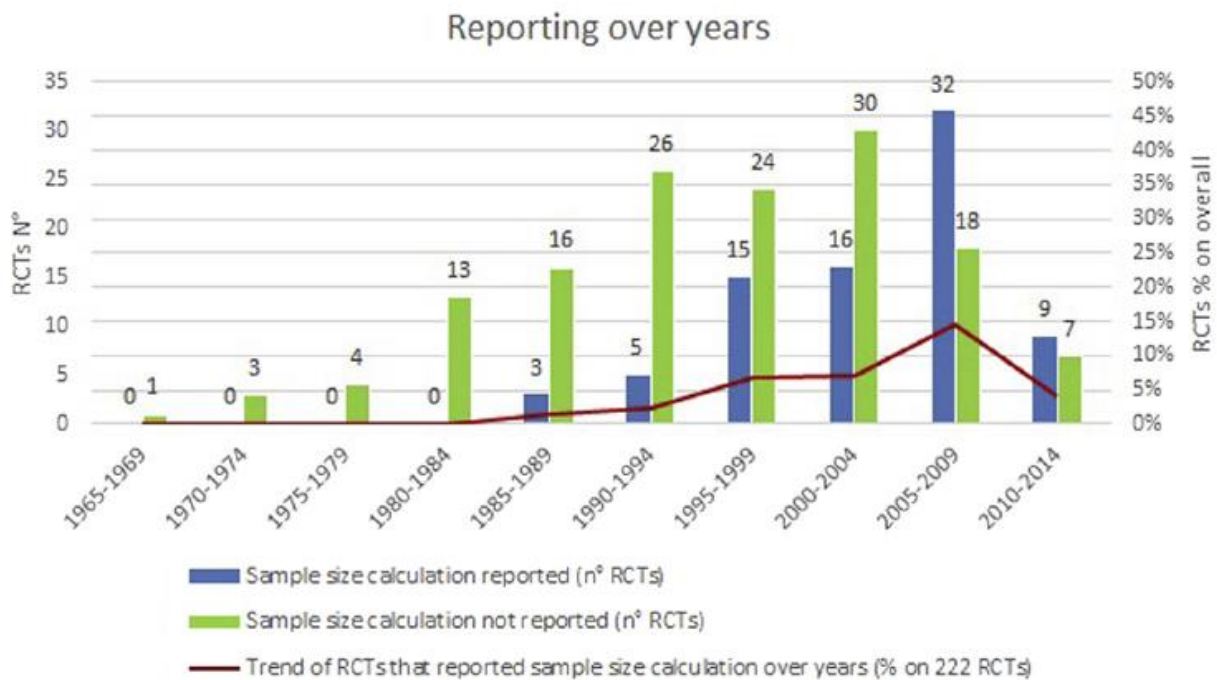


Figure 2. Trend of improvement in reporting of sample size calculation over time.

Complete description of sample size calculation

Thirteen (16.3%) of the 80 RCTs reporting sample size calculation gave an adequate description of the a priori sample size calculation, with all 6 elements provided in compliance with CONSORT guidelines. Half of the RCTs reported at least 4 of 6 elements (fig 3). Of the 6 CONSORT elements required for sample size calculation, the 3 most frequently reported were power (91.3%, n=73), assumption of the expected treatment effect of the intervention (86.3%, n=69), and alpha or type I error (85.0%, n=68). Adjustment to accommodate attrition rate was the least frequently reported element (32.5%, n=26).

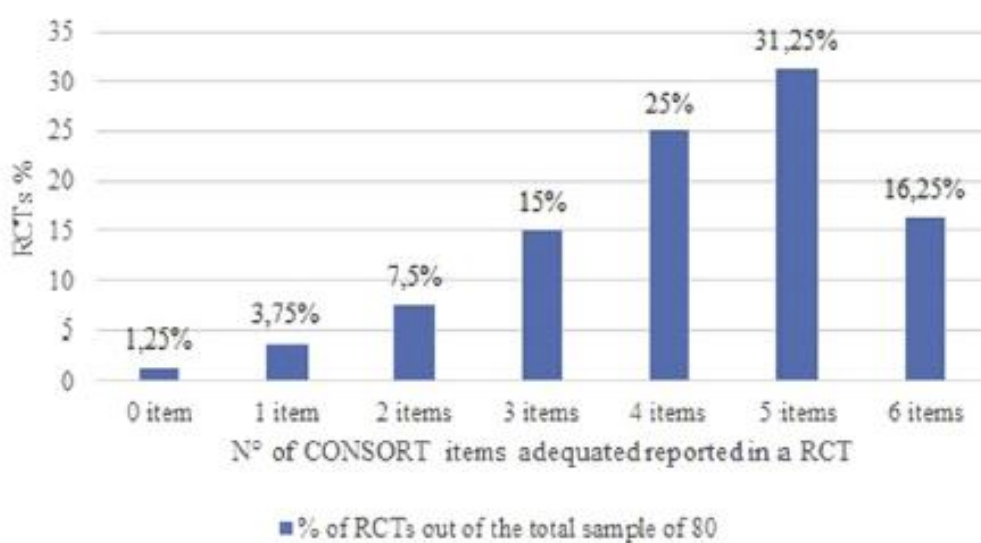


Figure 3. Completeness of sample size calculation description.

Characteristics of each element reported

Each element could be expressed in a different way; **table 2** present common expressions for elements. Power was usually defined as 1-beta (82.5%, n=66). The minimal important difference was the assumed value for the detection of the treatment effect most often reported in the 80 trials (46.3%, n=37). Concerning the outcome on which the calculation was based, all RCTs evaluated continuous outcomes: disability was the one most often reported (42.5%, n=34), followed by pain (22.5%, n=18).

Table 2 Commonly reported elements for sample size calculation

Element for Sample Size Calculation	n (%)
Level of significance	
α or type I error	68 (85)
Power	
β or type II error	10 (12.5)
$1-\beta$	66 (82.5)
Total	73 (91.3)
Assumption of treatment effect	
MID	37 (46.3)
Effect size	9 (11.3)
Other (ie, reduction in percentage)	24 (30)
Total	69 (86.3)
Assumption of variability	
SD	28 (35)
Other (ie, variance)	7 (8.8)
Total	35 (43.8)
Correction for losses to follow-up	26 (32.5)
Outcome considered for sample calculation	
Disability	34 (42.5)
Pain	18 (22.5)
Other (ie, recovery rate and work days)	19 (23.8)
Total	63 (78.8)

Abbreviation: MID, minimal important difference.

Discrepancy between the sample size planned and the sample size randomized

The sample size planned was reported in 72 of 80 RCTs. In the remaining 8 (10.0%) RCTs that reported the sample size calculation, the number of participants planned was not stated. The median number of participants needed to prove sufficient power was 120 (interquartile range, 17-2000), whereas the median number of participants randomized among these 72

RCTs was 133 (interquartile range, 21-741). The number of participants randomized was lower than the number of participants planned in 17 RCTs (23.6%), equal in 13 (18.0%), and higher in 42 (58.4%); figure 4 showed the discrepancy between the sample size planned and the number of participants randomized when the number obtained by the sample size calculation increased.

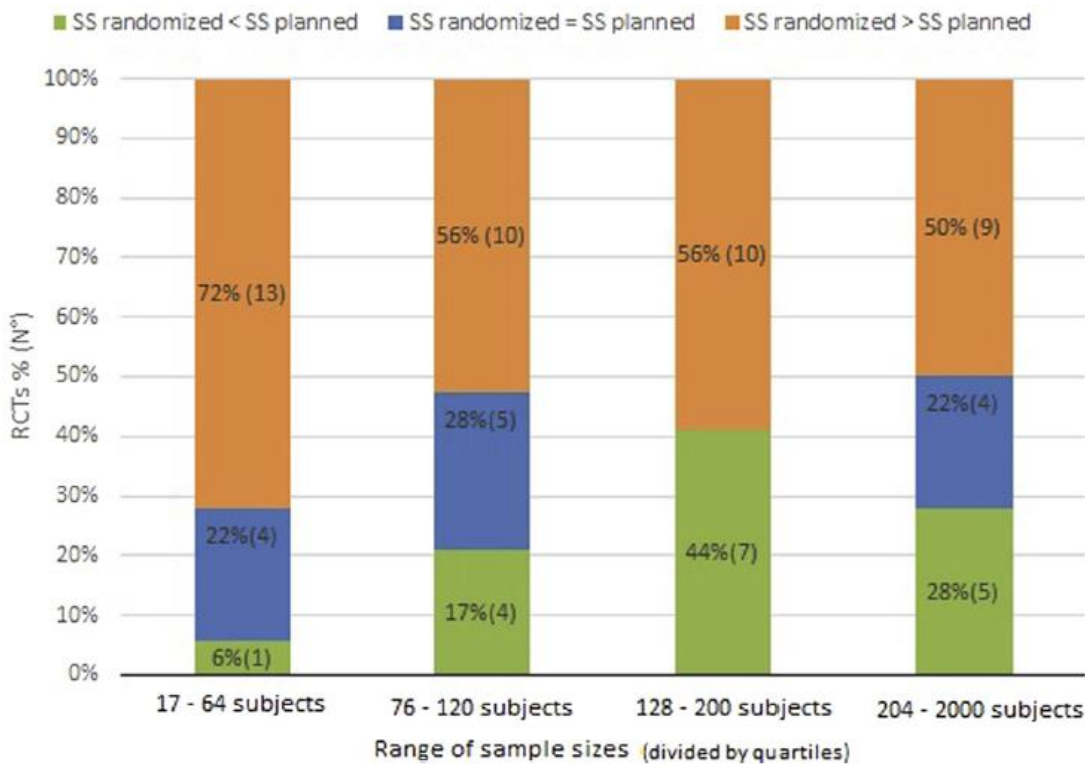


Figure 4. discrepancy between the sample size planned and the sample size randomized.

Discussion

Reporting of sample size calculation in RCTs on low back pain rehabilitation is often incomplete. We found that numerous RCTs published between the 1960s and the present failed to report a priori sample size calculation, barring readers from understanding whether calculation was done and whether it was done correctly. Among the RCTs reporting a priori sample size calculation, only a minority gave a complete description of the elements used. Nevertheless, the reporting of sample size calculation and its elements has increased over years; since 2005, more RCTs report sample size calculation than those that do not. Moreover, our results showed that the publication of the CONSORT statement has increased authors' awareness of high-quality reporting as compared to that in the pre-CONSORT era. Despite this, assessing the quality of the reporting does not necessarily reflect the quality of the underlying research: it is fundamental, distinguishing between "what researchers do" and "what researchers report." For instance, the assessment of risk of bias in a RCT leads to ambiguity between the quality of reporting and the quality of the research (Higgins JPT and Green S 2011). Our findings are consistent with a previous review of the general medical literature that described poor compliance by authors with CONSORT guidelines. Similarly, a review of physical medicine and rehabilitation trials published between 1998 and 2008 found that reporting had

somewhat improved, with only slightly more than half of the articles (57.3%) published in 2008 reporting sample size analysis (Abdul Latif, Daud Amadera et al. 2011). Conducting responsible research entails complete accurate reporting in a transparent fashion according to international guidelines. To ensure high quality in conducting a clinical trial, it is not sufficient to state the sample size without giving a description of how it was calculated. More than half of the RCTs with a priori sample size analysis included in our review reported fewer than 4 of the 6 elements required for replication of calculations. A recent review (ACTION systematic review) found that half of the published analgesic clinical trials gave an incomplete description of sample size calculation (McKeown A 2015).

Sample size calculation is usually based on a single outcome, chosen as a primary measure: specifying it helps researchers to clarify the initial basis upon which an RCT is built, besides simplifying interpretation, judgment, and use of findings (Cook, Hislop et al. 2015). We noted that more than half of the RCTs stated the primary endpoint, similar to the rates reported in a previous review in physical medicine trials (Abdul Latif, Daud Amadera et al. 2011). In the literature, disability and pain are the most frequently investigated outcomes in low back pain rehabilitation: several authors have recommended including these measurements in the back-specific core outcome sets because they are most relevant to patients, health care practitioners, regulators, industry representatives, and policymakers (Froud, Patterson et al. 2014). They were also the elective outcome measures most often used in RCTs according to our and recent review that found a low frequency of reporting outcome and intervention descriptions, reflecting a multidimensional lack of quality in rehabilitation RCTs (Gianola, Castellini et al. 2016).

Among the RCTs in which a power analysis was performed, 72 reported the sample size planned. In 2 out of 3 of these RCTs, the sample size randomized was larger than the sample size planned, and in a small proportion (30%) the sample size randomized was smaller than the sample size planned. Although authors are always encouraged to include more than the minimum number of participants to compensate for loss to follow-up, over recruitment to account for attrition is unjustifiable both economically and ethically - economically unsound because of the high costs of clinical trials and ethically questionable because of potential harm to patients. Except for trials on rare diseases or early-phase trials, underpowered studies are unethical because they may fail to yield significant results, are more likely to be inconclusive, and produce more false negatives (Maggard, O'Connell et al. 2003, Charles, Giraudeau et al. 2009, Calvert, Blazeby et al. 2013). However, trials with an overly large sample size may waste resources in terms of patients, time, and funding beside offering significant but not clinically relevant results. Authors should aim to achieve robust research findings by

calculating an adequate sample size by using time and resources in the best cost-effective manner (Fitzner and Heckinger 2010) and in collaboration with experienced biostatisticians and methodologist researchers (Ioannidis, Greenland et al. 2014).

Our results show that funding status affects the quality of reporting. Building a sustainable funding scheme for clinical comparative research in areas less explored, that is, the “orphan areas” such as anesthesiology or orthopedics, is critical to support evidence-based practice in medical research (2010). Funding is fundamental to obtain more resources in terms of personnel and to make the research process more efficient. Economic support is important in both pharmacological research and research areas where public health needs are changing. For example, rehabilitation for low back pain has increased its importance in both primary care, where rehabilitation as intervention plays a central role in low back pain management, and research (Castellini, Gianola et al. 2016); therefore, evidence-based rehabilitation has grown. When the aim is to translate results from research to practice, it is essential to focus on how the evidence is generated. RCT reports should provide essential information so that readers can make better decisions in clinical practice, especially in the rehabilitation of low back pain, an increasingly common health problem with a substantial community and financial burden (Maniadakis and Gray 2000, March, Smith et al. 2014).

Future studies should assess the quality of reporting of other essential elements for clinicians in rehabilitation. For instance, an adequate and satisfied description of the experimental intervention should be crucial, as well as the description of the target population and the outcome selection. Maybe a multidimensional lack of reporting of information exists, reflecting difficulties in transferring the research’s results in clinical practice.

Study limitations

This study focused only on the reporting of sample size calculation and its elements as described in the Methods section of RCTs. It would have been interesting to compare the final publication with the published protocol to explore whether the absence of some elements was limited to the research article or were included in the research protocol. This was not possible because our sample comprised a wide range of RCTs published from 1968 to 2013.

Conclusions

Sample size calculation is essential to demonstrate that a trial is adequately designed to detect a likely real effect or association, if such exists, in a given population (Fitzner and Heckinger 2010). Although some elements are difficult to define, the assumptions made in the calculation should be reported in a transparent fashion. The CONSORT statement provides a standard guidance for authors to prepare reports of trial findings and to facilitate their complete and transparent reporting. In addition, the SPIRIT (Standard Protocol Items: Recommendation for Interventional Trials) initiative has recently strengthened the purpose to improve transparency in the trial protocols (Chan, Tetzlaff et al. 2015). Furthermore, Cook et al (Cook, Hislop et al. 2015) have just created a more extensive set of elements for adequate reporting of this process in trial protocols and results, also providing justifications for the assumption of sample size calculation. Just as researchers should be encouraged to use these guidelines so, too, journal editors and peer reviewers should impose stricter criteria for adequate and transparent reporting. In addition, the sharing of software could help to simplify sample size calculation. Improving the methodological quality of RCTs, and all types of trials, will go some way to ensure the validity of results, reproducibility of research, and dissemination of results from research to practice.

Chapter 2. Rehabilitation interventions in randomized controlled trials for low back pain: proof of statistical significance often is not relevant

Published as: Rehabilitation interventions in randomized controlled trials for low back pain: proof of statistical significance often is not relevant. Gianola S, Castellini G, Corbetta D, Moja L.

Health and Quality of Life Outcomes; 17; July 2019

Abstract

Objective: We aimed to assess if treatment effects of randomized controlled trials (RCTs) for low back pain (LBP) are statistically significant and clinically relevant, and if RCTs were powered to achieve clinically relevant differences on continuous outcomes.

Methods: We searched for all RCTs included in Cochrane Systematic Reviews focusing on the efficacy of rehabilitation interventions for LBP and published until April 2017. RCTs having sample size calculation and a planned minimal important difference were considered. In the primary analysis, we calculated the proportion of RCTs classified as “statistically significant and clinically relevant”, “statistically significant but not clinically relevant”, “not statistically significant but clinically relevant”, and “not statistically significant and not clinically relevant”. Then, we investigated how many times the mismatch between statistical significance and clinical relevance was due to inadequate power.

Results: From 20 eligible SRs including 101 RCTs, we identified 42 RCTs encompassing 81 intervention comparisons. Overall, 60% (25 RCTs) were statistically significant while only 36% (15 RCTs) were both statistically and clinically significant. Most trials (38%) did not discuss the clinical relevance of treatment effects when results did not reach statistical significance. Among trials with non-statistically significant findings, 60% did not reach the planned sample size, therefore being at risk to not detect an effect that is actually there (type II error).

Conclusions: Only a minority of positive RCT findings was both statistically significant and clinically relevant. Scarce diligence or frank omissions of important tactic elements of RCTs, such as clinical relevance, and power, decrease the reliability of study findings to current practice.

Keywords: Epidemiologic Methods, Trials, Randomized Clinical Minimal Clinically Important Difference, Patient Outcome Assessment, Data Interpretation, Statistical, Sample size

Methods

This is a retrospective cohort study, building on a previous published research.¹⁵ We updated the search strategy, adopted the same eligibility criteria and re-run the same selection process. Here, methods are briefly reported.

Literature search

We moved from Cochrane Systematic Reviews (SRs) for selecting trials since they are usually considered of high quality, and adopt extensive search strategies. For the identification of Cochrane SRs on LBP, we updated the previous search strategy to April 2017.¹⁵

RCTs eligibility criteria

From the eligible Cochrane SRs, we extracted all trials. We considered a trial eligible if it met all the following criteria: (i) was a RCT; (ii) identified a primary outcome and determined the sample size on the basis of the primary outcome; (iii) considered continuous outcomes (e.g., pain, disability) leaving out any binary outcomes (e.g., fall or not fall); (iv) identified a priori planned MID for the primary outcome measure in the sample size calculation; and (v) the language of publication was English or Italian.

Data collection

We developed an *ad hoc* data extraction form. For each trial, we collected general information (e.g., country and year of publication) and specific information. Specific information included: the primary continuous outcome (e.g., pain), scales used for the outcome assessment (e.g., numeric pain rating scale), details on measurement scoring (e.g., 0-10 points), planned sample size, planned MID, any bibliographic reference and/or explanation of the rationale for the choice of the MID (e.g. anchor/distribution or other methods), follow-ups, number of randomized patients, number of patients at any follow up. When the time of follow-up analysis was not specified in the sample size calculation, we arbitrarily selected the follow-up time point closest to the end of the intervention.

In addition, we classified the type of intervention as “active treatment” or “inert treatment”, the second used when the expected responses could not be attributed to the investigated interventions (e.g. lack of biological plausibility of an effect). More precisely, we considered the interventions such as manipulation as “active treatment”, while placebo or sham control treatments as “inert treatment”.

Referring to estimates of effect sizes, we noted the mean difference (MD) of the primary outcome and its 95% confidence intervals (CIs), or any other available data to estimate the effect size and its imprecision (e.g. standard errors).

Determination of statistical significance. For every comparison between the intervention and control, we dichotomized the statistical significance as ‘achieved’ or ‘not achieved’ according to the pre-specified significance level (i.e., when pre-specified significance level was less than 5%, $p < 0.05$, statistical significance was classified as ‘achieved’).

Determination of the clinical relevance. Between-group differences were compared with the planned MID reported in the sample size calculation, determining if the effect size reached clinical relevance. We classified clinical relevance as ‘achieved’ if the point estimate of the MD was equal or greater than the a priori planned MID, and ‘not achieved’ in the other case.

Determination of study powered. We defined a study as “powered” if the sample size was equal or greater than the sample size originally planned.

Finally, we screened all RCTs to determine how often authors discussed trial’ findings related to the clinical relevance. We revised all full-text sections and we classified each trial according to the attempt to interpret differences as clinically relevant as “clinical relevance discussed” or “clinical relevance not discussed”. Two reviewers conducted the screening independently and a third author was consulted in case of disagreements.

Data analysis

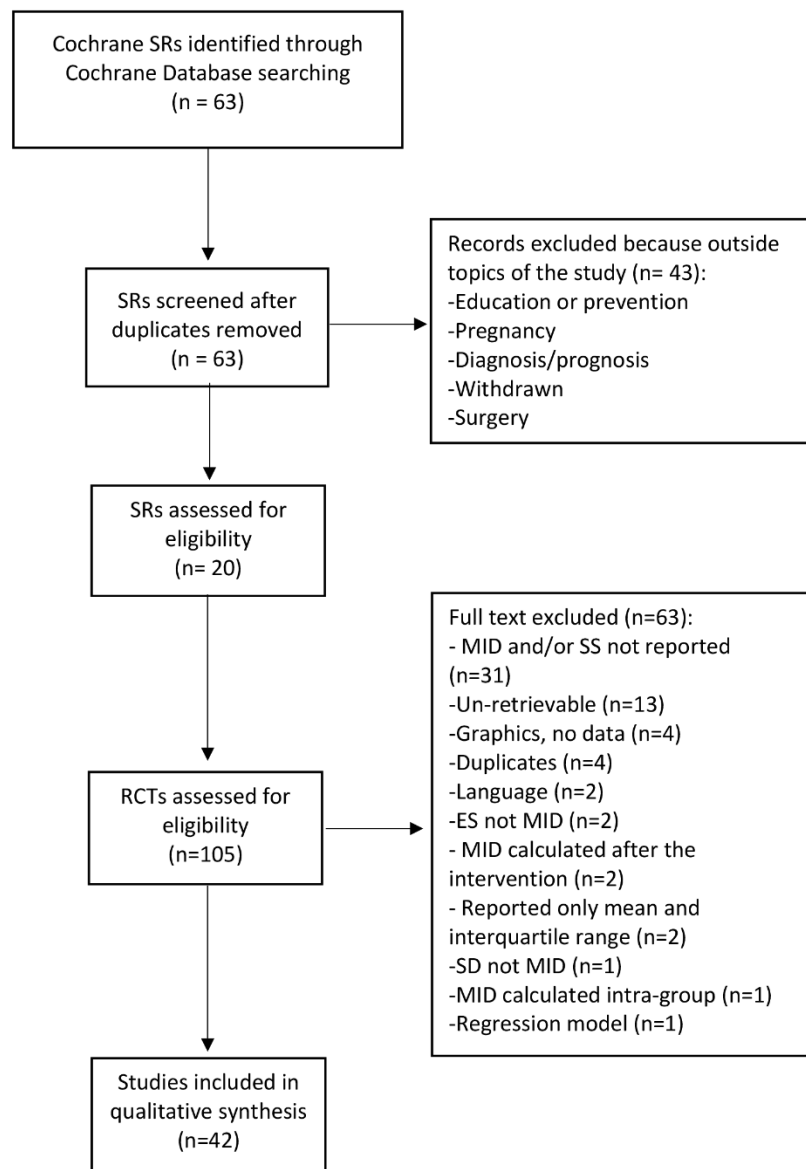
We reported data of continuous variables by medians and interquartile range (IQR), and data of categorical variables by frequencies and relative percentages. We computed the number of RCTs falling in each of the following four categories: “*statistically significant and clinically relevant*”, “*statistically significant but not clinically relevant*”, “*not statistically significant but clinically relevant*”, and “*not statistically significant and not clinically relevant*”. Whenever multiple arm comparisons were presented, in a primary analysis, we considered the whole trial as *statistically*

significant if at least one comparison was statistically significant. In a secondary analysis, we considered each multiple arm comparison as independent.

Results

Study selection

We identified sixty-three Cochrane SRs. After the selection process, 20 SRs were considered for the identification of eligible trials. One hundred-five RCTs were eligible from the included Cochrane SRs but only 42 (40%) met the inclusion criteria and were finally included in our study. The study selection process is shown in **Figure 1**.



Trial's characteristics

The 42 included RCTs were published in 19 journals. Most of these were published in *Spine* (n=13, 31%), in *The British Medical Journal* (n=5, 12%), and in the *Clinical Journal of Pain* (n=5, 12%). Thirteen countries were designated as publishing countries, of which the most frequent are United States (n=11 RCTs, 26%), United Kingdom (n=9 RCTs, 21%), Norway and the Netherlands (n=4 RCTs, 10%). The publication period runs from 1996 to 2014 (median = 2006; IQR = 2003 - 2008). Most RCTs reported the funding source (81%). One-fifth of the studies was multi-arm and 29% of trials calculated the sample size based on a composite outcome. One-third of trials (32%) investigated comparisons against an inert intervention. All general characteristics are reported in **Table 1**.

Table 1. General characteristics.

	n° of RCT (n=42)	%
<i>N° of countries (n=14)</i>		
United states	11	26
United Kingdom	9	21
Norway	4	10
Netherland	4	10
Brazil	3	7
Australia	3	7
Finland	2	5
Spain	1	2
Sweden	1	2
Switzerland	1	2
Italy	1	2
Thailand	1	2
Taiwan	1	2
<i>N° of journals (n=19)</i>		
<i>Most frequent journals</i>		
Spine	13	31
Clinical Journal of Pain	5	12
British Medical Journal	5	12
Journal of Manipulative and Physiological Therapeutics	3	7
<i>N° of reported funding</i>	34	81
<i>Multi-arm trials</i>	8	19
	n° of comparisons (n=81)	%
<i>Comparisons</i>		
active treatment <i>versus</i> active treatment	55	68
active treatment <i>versus</i> inert treatment	26	32

Clinical relevance characteristics

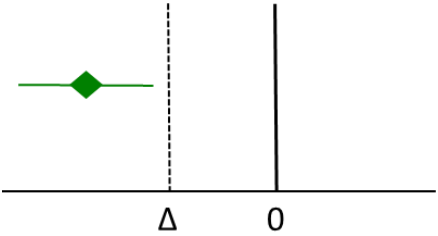
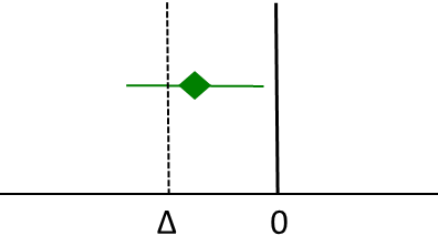
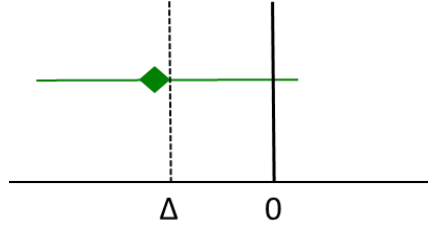
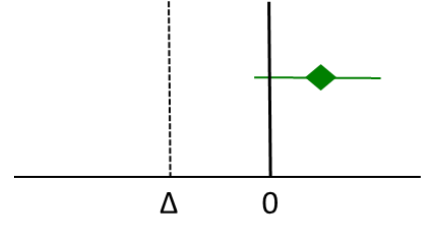
The majority of the included studies (n=37, 88%) reported MID as an absolute value, while the remaining studies reported it as a percentage of improvement over the baseline. Only a half of the included studies (n=20, 48%) referenced the source used to calculate the MID. Eliminating duplicates, 16 different method sources were found and examined. Of these, 6 were anchor-based, one distribution-based, one based on consensus (e.g. expert panel), three cited other articles, three were not clearly described and information was not found in two studies.

Is the effect always clinically relevant?

Table 2 shows the main findings for statically significant and clinically relevant results. We found that almost a half of trials (n=16, 40%) were “not statistically significant and not clinically relevant” and more than a half (n=25, 59%) were statistically significant. Out of these 25, 15 trials (36% of all included trials) were both “statistically significant and clinically relevant” and 10 trials, (24 % of all included trials) were “statistically significant but not clinically relevant”. One trial was classified as “not statistically significant but clinically relevant”.

Considering all comparisons of multiple arm trials (n=81) the four scenarios were similar to those reported in **table 2**. However, among statistical positive findings (scenario A and B, n=42 out of 81), a consistent part of the comparisons were against an inert treatment (40%) as compared to active head-to-head comparisons (60%).

Table 2. Statistically significance and clinically relevance on continuous outcomes of LBP. Δ is the MID. Negative values means improvement (for example, greater pain reduction in the treatment vs. control group).

Scenario	N° of trials (%) (total=42)	N° of comparisons (%) (total=81)
 <p>A) statistically significant and clinically relevant</p>	15 (36%)	20 (25%)
 <p>B) statistically significant but not clinically relevant</p>	10 (24%)	22 (27%)
 <p>C) not statistically significant but clinically relevant</p>	1 (2%)	1 (1%)
 <p>D) not statistically significant and not clinically relevant</p>	16 (38%)	38 (47%)

Is the clinical relevance always discussed?

Eighteen out of 42 trials (43%) did not report or discuss the clinical relevance of their results, even when clinical relevance was demonstrated (5%). **Table 3** lists most frequent types of omissions and embellishments characterising reporting of results when clinical relevance was not reached and considered. Full omission for the primary outcome was the main strategy (11 trials).

Table 3. Types of omissions and embellishments in reporting RCT findings when clinical relevance was not reached and considered (n = 42)

Clinical Relevance discussion	Strategy for specific reporting	No. (%)
Clinical relevance discussed		24 (57)
Clinical relevance not discussed		18 (43)
not reached		16 (39)
	Full omission for the primary outcome	7 (44) *
	Full omission for all primary outcomes used in the sample size calculation**	4 (25) *
	Clinical relevance discussed only as within-group improvements	4 (25) *
	Clinical relevance discussed at follow-ups not declared in the sample size calculation	1(6) *
reached		2 (5) *

* The Total refers to 16 trials that not discussed the clinical relevance.

**Composite outcomes.

Were non-statistically significant interventions powered?

Four studies did not report the number obtained in the sample size calculation. For the other 38 studies, the median of the sample sizes planned a priori was 125 subjects, while the median of the actual enrolled sample sizes was 133 subjects. Nevertheless, 14 trials out of 38 (37%) reached the planned sample size, while remaining were low-powered. Sixteen out of 38 trials (42%) do not achieve the statistical significance but less than half of these (n=6, 40%) have an adequate and powered sample size (**Table 4**, scenarios C and D).

Table 4. Statistical and clinical effects according to planned a priori sample size achievement.

Scenario per n° of trials (total=38*)	N° of powered trials (%)	N° of unpowered trials (%)
A) statistically significant and clinically relevant (n=14)	7 (50)	7 (50)
B) statistically significant but not clinically relevant (n=8)	1 (12)	7 (87)
C) not statistically significant but clinically relevant (n=1)	0	1 (100)
D) not statistically significant and not clinically relevant (n=15)	6 (40)	9 (60)
Totals	14 (37)	24 (63)

* The total number of trials is 38 because four studies did not report the patients number obtained from the sample size calculation (1 study belongs to scenario A, 2 belong to scenario B and 1 study belongs to scenario D).

Discussion

Based on the RCTs included in our retrospective cohort, we found a poor reporting of rehabilitation interventions for LBP in terms of validity and clinical relevance. In fact, only a half of the trials reported the source reference of the adopted planned MID as measure of validity of the clinical relevance and, among studies achieving statistically significant results ($n=25/42$), only 60% ($n=15/25$) achieved the planned clinical relevance. This means that 1 out of 2 studies reaching a statistically significant difference favoring a treatment has results that cannot be truly relevant for stakeholders, clinicians and patients. This result could also be overestimate because the 29% of trials reported the sample size based on composite outcomes at risk of falling in type I error. Less than a half of RCTs in our sample (43%) did not discuss their findings from a clinical perspective, mainly by omitting information, particularly when the clinical relevance was not reached. Finally, we found that in trials reaching statistically significant and clinically relevant results the majority of multi-arm comparisons were against inert treatments. All these findings support the hypothesis that the efficacy of rehabilitation interventions for LBP tend to be overestimated, or potentially underestimated if we considered that 63% trials with not statistically significant and not clinically relevant results did not have an adequate and powered sample size.

Our results are coherent with the literature where the reporting of results in terms of clinical relevance is sparsely used across trials.(Hoffmann, Thomas et al. 2014) We confirm the preliminary results published by Van Tulder et al. focused on exercise therapy for chronic LBP reporting that less than half of studies (39%) with positive conclusions shown clinically important differences(van Tulder, Malmivaara et al. 2007). A general poor reporting of clinical relevance is also present across pharmacological interventions with a discussion of results in clinical terms ranging from 24% to 46% of the samples. (Pocock, Geller et al. 1987, Moher, Dulberg et al. 1994, Chan, Man-Son-Hing et al. 2001, Molnar, Man-Son-Hing et al. 2009, Hoffmann, Thomas et al. 2014)

When results of trials do not achieve a statistically significant and/or a clinically relevant difference among treatments, authors tends to discuss and shape the impression of their results for readers. In scientific writing, this is called “to spin” the scientific report(Junger 1995). In our sample of RCTs, the most frequently adopted strategy to spin the report was the under-reporting of the clinical relevance of statistically significant results. One possible reason at the basis of this phenomenon can be found in the publication process of biomedical research that tends to favor the publication of positive results.(Chalmers and Matthews 2006) To some extent, similarly for statistical significance, it can happen that reports of RCTs with clinically relevant results are published more often than those

with not-clinically relevant results.(Rising, Bacchetti et al. 2008, Turner, Matthews et al. 2008)·(Chan, Hrobjartsson et al. 2004, Chan and Altman 2005, Al-Marzouki, Roberts et al. 2008, Dwan, Altman et al. 2008, Song, Parekh et al. 2010)

Clinical relevance can influence not only the statistically significant results but also the non-statistically ones. In fact, the aim to detect a MID between the intervention and control group determines the power of study in the sample size calculation. A study conducted in 2008 in the field of physical medicine and rehabilitation reported that, of the 82 articles reviewed, 57% reported sample size calculation and 13% of them without sufficient information about the parameters required for a priori calculation.(Abdul Latif, Daud Amadera et al. 2011) In a more recent published study on low back pain rehabilitation we denounced a low frequency of trials reporting all elements needed for sample size calculation.(Castellini, Gianola et al. 2016) Anyway, also taking into account all trials with all the elements for sample size calculation, we now found a very low percentage of powered trials that clinically interpreted their findings on the light of the planned clinical relevance. This issue encompass other healthcare professions than the physical rehabilitation in the evidence based care: a large proportion of the existing trials are poorly designed and underpowered.(Geha, Moseley et al. 2013) The potential weakness in small-size “negative” clinical trials was already reported and pointed out fifty year ago.(Freiman, Chalmers et al. 1978) In a planning phase of a clinical trial, scientific ethics committees should be more rigid on the sample size definition requiring its “a priori” calculation and its complete reporting. Ethics committees should mandatory require researchers to provide the preliminary data or a referenced study assessing the MID of the outcome used for the determination of the sample size calculation. This is expected to happen also for pilot studies, even if exposed to unexplored knowledge. The size of a pilot study should be calculated in relation to the desired level of confidence for the SD and the chosen power and significance level of analysis in the main study. At high level of confidence, a pilot study of at least $n=50$ is advisable in many circumstances.(Sim and Lewis 2012) The same process should be followed during the editorial assessment of a scientific report before its publication. We also suggest researchers to select a single primary endpoint for formal statistical inference otherwise involving several outcomes conventional significance testing can seriously inflate the overall type I error rate.(Pocock, Geller et al. 1987) Finally, an accurate replication of the sample size should be done, prior to the approval of experimental study, in order to avoid approximated, wrong and unfounded assumptions.

Implication for clinical practice

A very low proportion of trials research (2.6%) reflects the priorities of patients and clinicians showing an important mismatch between wishes of patients and evaluations of researchers.(Crowe, Fenton et al. 2015) Treatments efficacy should be useful in terms clinical relevance. This would allow a better informing patients strategy on the possible benefits and harms of the intervention, as well as their size or likelihood, costs, and inconveniences of the intervention for a tailored therapy. The shared decision-making approach should encompass the patient's preferences and values into the discussion in a perspective evidence based health-care.(Sackett, Rosenberg et al. 1996) A treatment leading to non-relevant results for patients is often an unsuccessful treatment, resulting in frustration, discontinuation of therapies and waste of resources. An approach focused on the achievement clinical relevance of a treatment will increase awareness of condition and the participation of each patient in the managing of their benefit-harm trade-off tailored, limiting the burden of physical rehabilitation conditions for obsolete or harmful or discontinued treatments.

Implication for research

We call for more adherence to reporting of planned sample size including the clinical relevance with the clinical interpretation of the effects. Without the complete information, the reader is unable to fully interpret the results of a study.(Hoffmann, Thomas et al. 2014) On one hand, authors have to report all elements used for sample size calculation, including the clinical relevance. Furthermore, they have the duty to interpret observed effects on the light of this threshold coherently. Otherwise, sample size calculation does not make any sense. On the other hand, editors and reviewers have to enforce authors to provide sufficient details about clinical relevance and sample size calculation (sample planned, randomized and reached) for the primary outcome.

This would prevent an unusable treatment for its non-interpretable effects and leading to promote treatments clinically relevant. Then, the actual guidelines for the reporting of patient-reported outcome in RCTs, endorsed through the initiatives of the Consort Statement,(Boutron, Moher et al. 2008, Calvert, Blazeby et al. 2013) promote the discussion (item 22) of a minimal important change in the interpretation of patient-reported outcome results. Actually, clinical relevance is not explicitly contemplated in the planning, the current item 7 only describe "how sample size was planned". The reporting of patient-reported outcome in RCTs must consider to expand the item regarding the sample size definition introducing a dedicate section for the declared a priori clinical relevance and reporting of its validity (e.g., by citation of references). This approach can avoid too positive results improving

the assessment of the imprecision of quality of evidence and standardizing the process. If findings are statistically significant but not clinically important, the quality of evidence in the meta-analyses will potentially change the conclusions.(Guyatt, Oxman et al. 2011)

Strength and limitations

The major strength of this study is to assess the clinical relevance of results assuming the MID declared in the sample size of each study and not a standardized MID as already previously investigated.(van Tulder, Malmivaara et al. 2007) However, some limitations are present. The sample of trials only included non-pharmacological LBP interventions and our findings may not be extended to other trials published on different interventions (e.g., pharmacological interventions). Moreover, we did not assess the risk of bias in each trial that could have been correlated the quality of study to the interpretation of results.

Conclusions

Authors' conclusions are usually too positive and the clinical relevance is not yet fully considered as a valid measure reported in the sample size, and in the interpretation of findings of RCTs in LBP rehabilitation. If authors of trials reported adequately the a priori sample size and commented their results in term of clinical relevance, the threshold for efficacy could identify the true effect. We also called for powered trials, particularly optimizing the sample size recruitment and for more attention in interpretations of findings from a clinical perspective. Even if studies with larger sample sizes are more onerous in terms of both time and money, they can give results that are more reliable if adequately powered and precise. Sample size calculation must be performed before conducting a trial in order to ensure to have a sufficiently large sample size to be able to draw meaningful conclusions. Without a complete, corrected and powered sample size, the common justification of “not enough power to detect a significant efficacy of the intervention” is always justifiable for negative studies.

General Conclusion

It has been showed that the amount of research on low back pain has been increasing over the years (Castellini, Gianola et al. 2016). However, it seems that limited efforts were directed to improve its quality and validity. Indeed, a high prevalence of poor reporting in RCTs in low back pain rehabilitation was found in our research publications. This results in publishing research not reliable and inaccurate, with a production of wasteful research and consequent useless treatments in clinical practice and results not transferable to clinical practice.

When a literature synthesis is needed to provide evidence about the effectiveness of an intervention, the RCT is the unit of analysis from which researchers, clinicians and stakeholders should start to look for it. Nevertheless, a valid and reliable RCT is essential to provide reliable evidence synthesis. When an RCT demonstrate to fail to report essential elements, as we showed, results of systematic reviews and meta-analysis should be cautiously critically considered and interpreted.

SECTION 2

Which is the likelihood of a meta-analysis to be underpowered, inconclusive and imprecise?

Background

Even though a meta-analysis is often considered the best approach to quantitatively synthesize the evidence, in this section we are going to show methods useful to optimize meta-analysis results and overcome its limitations. In particular, (a) the likelihood to be underpowered, not conclusive and imprecise and (b) the limit to provide just head-to-head comparisons.

Underpowered and imprecise meta-analysis

Random errors frequently cause erroneous estimation of treatment effect when meta-analyses are small (Thorlund, Imberger et al. 2011). The lack of statistical power and precision can amplify the random error in a meta-analysis, leading to neutral or negative (non-positive) findings. The TSA is a frequentist methodology which can handle the risk of overestimating or underestimating clinical findings: it combines the calculation of a required information size (RIS) for a meta-analysis and an adjusted threshold for a statistically significant treatment effect. TSA's conclusions have the potential to be more reliable than those using traditional meta-analysis techniques. Methods as TSA emphasize the fact that an accurate and critical assessment, both qualitative and quantitative, of the body of the evidence has become fundamental in clinical research.

During my first year I have deepened the TSA methodology spending a period at the Copenhagen Trial Unit, between April 2016 and September 2016. During this working experience and studying the TSA methodology, I came across a publication by Jakobsen and colleagues which suggested to include the TSA as a supplement of imprecision assessment of the GRADE system (Jakobsen, Wetterslev et al. 2014).

The GRADE system is an international standard to assess the strength of body of the evidence, informing transparently and explicitly the confidence that researchers have on the results. (Balshem,

Helfand et al. 2011). GRADE uses a framework of information about risk of bias, imprecision, inconsistency, indirectness and publication bias. Particularly, it defines the issue of imprecision for SRs as the confidence in the estimate of the overall effect (Guyatt, Oxman et al. 2011). Imprecision also encompasses the size or importance of an effect, and it is influenced by the magnitude of the sample size or the number of events.

I have with interest read the article by Jacobsen et al. on how the TSA and the GRADE assessment take both into consideration the imprecision domain in systematic reviews. Therefore, I have investigated how TSA can yield a different interpretation of this domain in meta-analysis results compared with those obtained by the GRADE system. Imprecision is a matter of interest in clinical decision making process. How much be confident in systematic reviews and trials results is what influence a clinician in his clinical decision making process.

Multi comparisons of interventions in a meta-analysis

A meta-analysis usually compares two interventions at a time, being of limited use since it provides only a partial view of the whole picture of treatment options for a given condition. Clinicians are always asked to provide the best evidence-based available treatment to their patients. A new meta-analysis technique, called network meta-analysis (or multiple treatments meta-analysis or mixed-treatment comparison), has been developed to assess the efficacy of several interventions and synthesize evidence across a network of randomized trials. I ended up with a case example on the effectiveness of interventions in low back pain rehabilitation developed in collaboration with a group of researchers.

Aim

In chapter 1, the TSA method was explained in theory and case example of its application were performed. Then, in chapter 2, we focused on imprecision domain of meta-analyses' results and we compared the imprecision assessment of the GRADE approach to that obtained by a TSA application. Secondly, in chapter 3, we studied and applied the network meta-analysis in low back pain rehabilitation field. Indeed, a meta-analysis usually offers a comparison between 2 interventions at a time, limiting the ability of clinicians, researchers and stakeholders to choose the best intervention among all the available for a specific condition, which is often the case in real clinical practice.

Chapter 1. Trial Sequential Analysis

The meta-analysis is an essential tool to detect intervention effects when controversies arise from conflicting studies. This statistical method can increase the power of results: when several studies are combined to detect a small effect, the chance of detecting it increases. A meta-analysis can also improve the precision of the intervention effect estimate since a great amount of information is collected (Higgins JPT and Green S 2011). However, it is not an infallible tool.

Underpowered meta-analysis

RCTs that fail to detect a real difference between treatment effects due to risks of biases may inflate the results of a meta-analysis. Our recent review of RCTs published in low back pain rehabilitation field showed that 64% of RCTs did not report transparently the power analysis (Castellini, Gianola et al. 2016). This has increased concerns not only about the quality with which researchers report details in the published manuscripts but also whether authors perform an adequate sample size calculation at protocol stage. Turner et al. found that, using the conventional power analysis and testing for a relative risk reduction of 30%, the 78% of Cochrane meta-analysis with a dichotomous outcome have a power <80% and the 50% have a power <27%. We can assume that the majority of published Cochrane meta-analyses might be underpowered and much attention should be paid to their overestimated conclusions (Turner, Bird et al. 2013, Imberger, Gluud et al. 2015, Imberger, Thorlund et al. 2016).

From simulation studies we know that random errors frequently cause erroneous estimation of treatment effect when meta-analyses are small (Thorlund, Imberger et al. 2011). The lack of statistical power and precision can amplify the random error in a meta-analysis, leading neutral or negative (non-positive) findings. Recently, *Imberger et colleagues* demonstrated that in a sample of 50 meta-analyses investigated an anesthesiology intervention, only the 12% (95% CI, 5%–25%) had a power $\geq 80\%$ (Imberger, Gluud et al. 2015). The same research group performed a further analysis where they showed that meta-analyses that surpassed their optimal information size had sufficient protection against overestimation of intervention effects (Thorlund, Imberger et al. 2011). Furthermore, in a sample of 22 cardiovascular meta-analyses reported to be conclusive, the 55% (n=12) were found to contain insufficient data to detect a 25% risk reduction (AlBalawi, McAlister et al. 2013).

Consequently, as in a single randomized controlled trial a sample size calculation is needed to guarantee a sufficient number of patients in order to have reliable results in detecting an effect that is closer to the ‘true’ one, a similar goal is needed for a meta-analysis.

Random error

If meta-analyses are small, they are likely to be updated in the future. The Cochrane Collaboration requires a regular updating of analyses in order to reflect the most currently research (Garner, Hopewell et al. 2016). This frequent updating lead to overestimate or underestimate the results since the type I error increase without being handled. The risk of random error increases more than 5% if accumulated data are analyzed during multiple up-dates (19). The more statistical tests are performed due to the accumulation of additional data, the higher is the probability of having a false positive or false negative result. This phenomenon is commonly identified as ‘*multiplicity due to repeated significance testing*’(Thorlund K, Wetterslev J et al. 2011). This increased error is analogous to the risk of error present when interim analyses are done in a single trial. In a single trial, adjustments are required for the increased random error caused by sparse data and repetitive testing. Monitoring boundaries are also commonly used to control the risk of random error at desired levels and to allow us to make inferential conclusions in the interim analysis (Bassler, Montori et al. 2008). As well, in a meta-analysis such procedure should be performed. As a matter of fact, several studies have confirmed that meta-analyses have false negative or false positive results. In cardiovascular meta-analyses, for instance, *Albalawi et al.* found that 17% of the statistically significant meta-analyses were false positives and 64% of those non statistical significant meta-analyses were potential false negatives (AlBalawi, McAlister et al. 2013). As well, *Brok et al.* declared that 19 out of 39 apparent conclusive Cochrane neonatal meta-analyses become inconclusive when the statistical analyses take into account the risk of random error due to repetitive testing (Brok, Thorlund et al. 2009).

Despite the existing literature on this topic, recommendations about how taking into account the random error are not included neither in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement nor in the Cochrane Users Handbook (AlBalawi, McAlister et al. 2013). To handle the random error a technique has been suggested by Gluud C. et colleagues at the Copenhagen Trial Unit: The Trial Sequential Analysis (TSA) (Thorlund K, Wetterslev J et al. 2011). This frequentistic method aim to reduce the uncertainty intrinsic in meta-analyses results,

providing a technique able to handle the type I and II error, combining the calculation of the required information size with the adjustment of the statistical threshold.

The Trial Sequential Analysis – method

The Trial Sequential Analysis (TSA) is a statistical technique developed to handle the random error and prevent premature statements of superiority of the experimental or control intervention or avoid probably falsely declarations of absence of effect in case of having too few data. TSA has two goals: (a) measure and account for the strength of the available evidence and (b) control the risk of type I error. The strength of the evidence is measured by defining a required information size while the control of the type I error is reached altering the way in which statistical significance is measured. Thus, TSA introduces the concept of the required information size (RIS) for a meta-analysis with adjusting the threshold for declaring the statistical significance. This technique can be viewed as more conservative than the conventional meta-analysis and it has been recently used in several systematic reviews (Bjelakovic, Gluud et al. 2014, Lee, Chan et al. 2015, Wetterslev, Meyhoff et al. 2015).

The required information size for a conclusive meta-analysis

The calculation of the required information size is similar to the sample size calculation in a single trial, which typically requires the expected proportion of patient with the outcome in the control group (or a standard deviation in case of continuous outcome), the expected relative risk reduction of the experimental intervention or the anticipated intervention effect (the minimum effect worthwhile for the patient), and the desired maximum risk of both type I error and type II error. The formula for the required information size is the following:

$$IS_{Patients} = 2 \cdot (Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot 2 \cdot \sigma^2 / \delta^2$$

Where δ is the *a priori* estimate of the mean difference among the two groups and σ^2 is the associated variance. Whereas, the value of Z referred to the desired level of type I and II error.

Heterogeneity adjustment

A meta-analysis includes likely heterogeneity due to the included trial populations, interventions, and methods therefore, it is not realistic to consider homogeneous the included trials in a meta-analysis

(Thorlund K, Wetterslev J et al. 2011, Thorlund, Imberger et al. 2012). A meta-analysis has usually more variation in population and interventions, compared with a single trial. Therefore, since increased variation can decrease the precision of results, information size must incorporate all sources of that variation, including heterogeneity. It seems so reasonable that meta-analysis' sample size needs to be adjusted in order to allow for the variance introduced by this heterogeneity: this is achieved increasing the size of the sample. One approach for incorporating heterogeneity in information size is to adopt a heterogeneity -adjustment factor and multiply the information size by it.

In the fixed effect model is assumed that all included trials can be view as duplicates of the same trial. On the other hand, in the random effect model it is assumed that all included trials come from a distribution of several possible trials. The variance in a random effects model is always higher and greater than that in the fixed model. Therefore, the heterogeneity-adjustment factor must account for the increase in variation that a meta-analysis incurs from going from the fixed-effect assumption to the random-effects assumption. An accurate adjustment can be achieved by making the heterogeneity adjustment factor equal to the ratio of the total variance in a random-effects model meta-analysis and the total variance in a fixed-effect model. The heterogeneity-adjustment factor is therefore always equal to or greater than 1.

In order to adjusted the size of the sample for the amount of heterogeneity expressed by I^2 , the RIS obtained through the formula above is multiplied for $1/(1-I^2)$.

However, when the trial weights are not equal, using I^2 will lead to an underestimation of the adjustment factor, and so, an underestimation of the required information size (Wetterslev, Thorlund et al. 2009).

In this case, we can define a measure of diversity (D^2) as the quantity which satisfies the equation:

$$D^2 = \frac{v_R - v_F}{v_R} = \frac{\tau^2}{\tau^2 + \sigma_D^2}$$

Where v_R and v_F represent the variances (inverse of trials' weights) in the random and fixed models respectively, τ^2 is the between trial variance and σ_D^2 is the typical moment based sampling error within the trials. D^2 can be defined as the percentage of the total variance (sum of between trial variance and sampling error) in a random effects model, contributed by the between trial variance. Using D^2 in calculating the required information size in any random effects model meta-analyses seems less

biased than the I^2 in simulation study (Wetterslev, Thorlund et al. 2009). Therefore, the required information size is adjusted taking into consideration this measure of diversity.

The alpha-spending function and trial sequential monitoring boundaries

The standards for testing statistical significance in meta-analyses should be, at least, equal to those of an RCT. Updating meta-analyses is similar to interim analyses of an RCT. In an RCT is mandatory to perform an interim analysis when it could be unethical to keep recruiting patients if one of the groups investigated (experimental or the control group) is superior to the others. When calculating the sample size, it should be decided the level of statistical significance we want to use as threshold to test when sample size has been reached. At this point, only when the size of the sample has been fulfilled, a two-sided p-value of less than 0.05 (test statistic z-value of ± 1.96) can be accepted as statistical level. There is a general consensus about the fact that the sequential testing as in an interim analysis increase the risk of type I error more than 5% if not handled, resulting even in a type I error from 10% to 30%. Therefore, the statistical level should be more conservative and restricted (less than the nominal p value of 0.05) before the a priori estimated sample size has been reached. Such adjustments can be performed through the use of sequential monitoring boundaries that function as a threshold for the employed test statistic. Sequential monitoring boundaries demand a conservative interpretation when data are sparse in a single trial, but become increasingly tolerant as more data accumulate and get closer to the RIS. Similar boundaries are used for cumulative meta-analysis updates and they are referred as *trial sequential monitoring boundaries*.

The methods for adjusting the significance threshold in TSA are based on the Lan and DeMets approach (Thorlund K, Wetterslev J et al. 2011). Their method referred to the alpha spending function, which is a monotonically increasing function that can assign a maximum type I error risk at each significance testing according to the accrued information. As the accrued information increases, the size of the acceptable type I error also increases. The alpha spending function is defined from 0 to 1, where 0 represents no patients randomized and 1 where the required information size has been reached. At any point between 0 and 1, the alpha spending function is equal to how much type of I error has been acceptable. This results in conservative boundaries when limited amount of information has been accumulated, i.e. early stages, and more lenient boundaries when more and more data are gathered. The alpha – spending function Lan DeMets provides strongly conservative p value closer to 0.01 at early stages of accumulated data and more lenient p values (closer to 0.05)

when the sample size gets closer to the required information size. Therefore, each z value of the monitoring boundaries corresponds to the maximal allowed cumulative type I error for that number of participants.

In conclusion, the trial sequential monitoring boundaries applied on the TSA have the aim to maintain the overall risk of type I error $\leq 5\%$ independently of how many times the hypothesis is repeated. The complex calculations for trial sequential monitoring boundaries are all automated via easy-to-use free software (2011), and can be performed by clinicians and clinical researchers with basic training in statistics.

The cumulative test statistic (Z-curve)

Once the information size is calculated, the studies are added one at a time in a chronological order and the results are summarized as a new study is added. At each stage the conclusiveness of the evidence is analyzed and provided as a z score producing a curve which is commonly named as the Z-curve. This statistic is given by the log of the pooled intervention effect divided by its standard error, and is commonly referred to as the Z-statistic or the Z-value. Under the assumption that the two investigated interventions do not differ (the null hypothesis) the Z-value will approximately follow a standard normal distribution (a normal distribution with mean 0 and standard deviation 1). The larger the absolute value of an observed Z-value, the stronger is the statistical evidence that the two investigated interventions do differ and so the less is the probability that the data come from population where the null hypothesis is true. The cumulative z-score estimates the random error and is interrelated with the trial sequential monitoring boundaries.

Figure 1 showed how TSA is displayed:

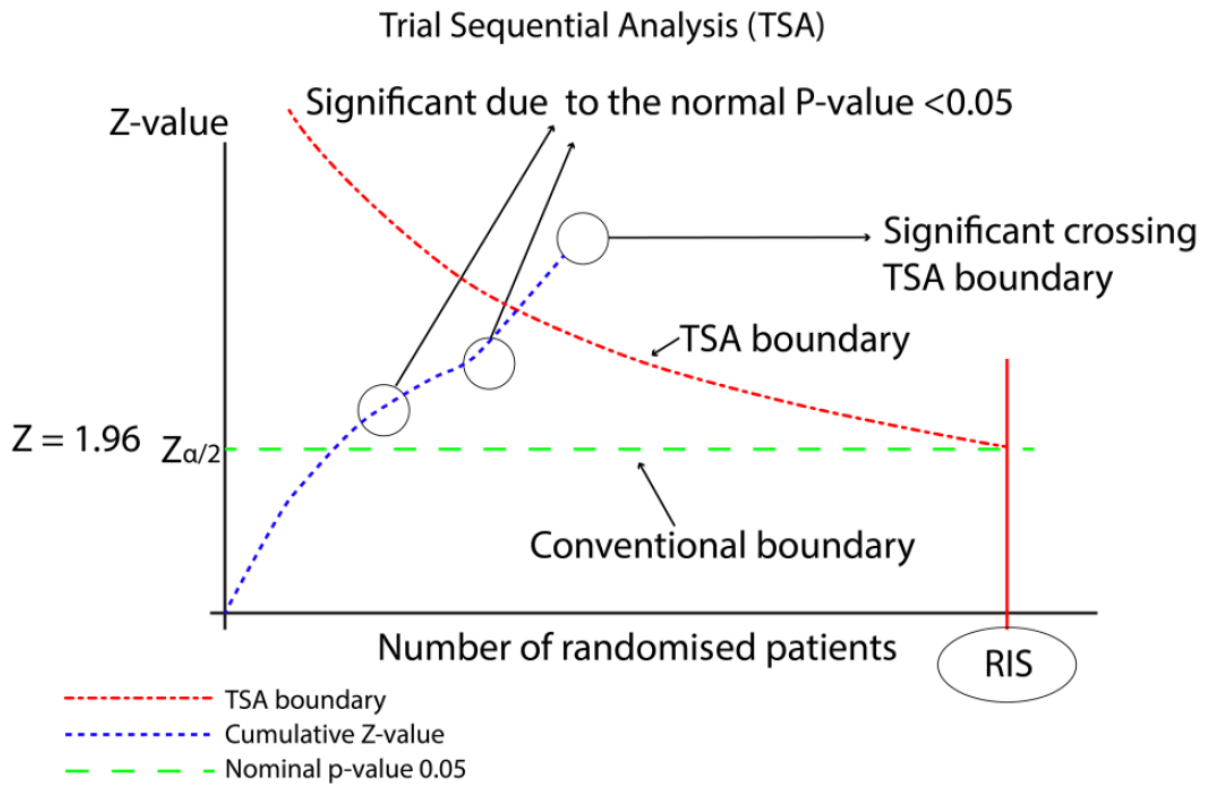
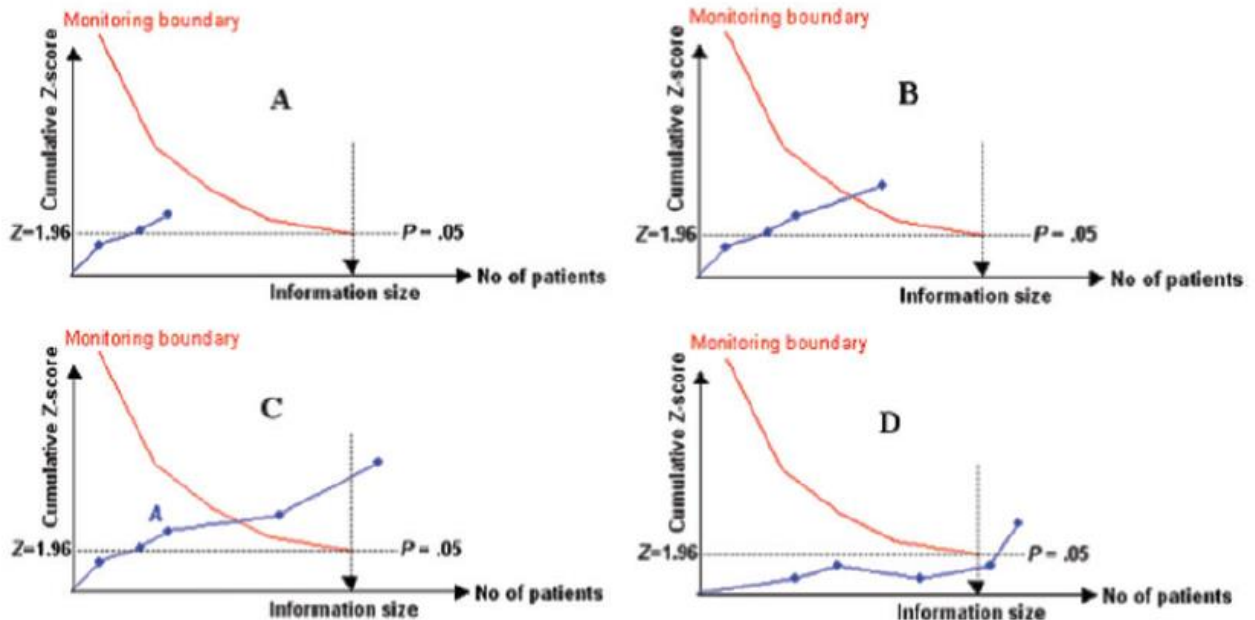


Figure 1. Trial Sequential Analysis (taken by Imberger et al. 2016, (Imberger, Thorlund et al. 2016))

To better clarify how a TSA can result, four examples with the relative interpretation are extracted by *Brok et. al* (Brok, Thorlund et al. 2009):



The red line corresponds to the trial sequential monitoring boundaries, the blue line is the cumulative z-score over the time, the dotted black horizontal line is the conventional statistic of $p = 0.05$ and the dotted black vertical line represents the required information size. The trial sequential monitoring boundaries (red) needs to be crossed by the cumulative z-curve in order to obtain reliable evidence.

Following the explanation of the four scenarios:

- A. Non conclusive evidence: the number of participants has not reached the required information size yet and the z-cumulative curve has not crossed the trial sequential monitoring boundaries.
- B. Evidence for at least 25% relative risk reduction: the number of participants has not reached the information size, but the cumulative Z-curve crossed the monitoring boundary.
- C. Evidence for at least 25% relative risk reduction: the number of participants reached the information size and the cumulative Z-curve crossed the monitoring boundary.
- D. Evidence of less than 25% relative risk reduction: The cumulative Z-curve has not crossed the monitoring boundary before reaching the information size.

In summary, when the cumulative Z-curve (the series of Z-statistics after each consecutive trial) crosses the trial sequential monitoring boundary for benefit and the RIS, a sufficient level of evidence has been reached and no further trials may be needed to demonstrate the superiority of the intervention. However, if the RIS has not been crossed yet even though the monitoring boundary for benefit is crossed, there is sufficient certainty that the effect is beneficial for the patients but further trials are needed to amplify the confidence in the conclusion. On the contrary, if the cumulative Z-curve does not cross any of the trial sequential monitoring boundaries, insufficient evidence to reach a conclusion can be declared and additional trials may be needed.

Adjusted confidence interval following trial sequential analysis

Having considering the adjustment of the overall type I error in significance test, some considerations should be done also on the construction of confidence intervals. If a meta-analysis is subjected to repeated statistical evaluation, it produces also a series of confidence intervals over the time and the probability that all of these confidence intervals will contain the ‘true’ overall effect is certainly less than 95%. Therefore, when a meta-analysis is subjected to repeated statistical evaluation, there is an exaggerated risk that the ‘naïve’ confidence intervals will yield spurious inferences.

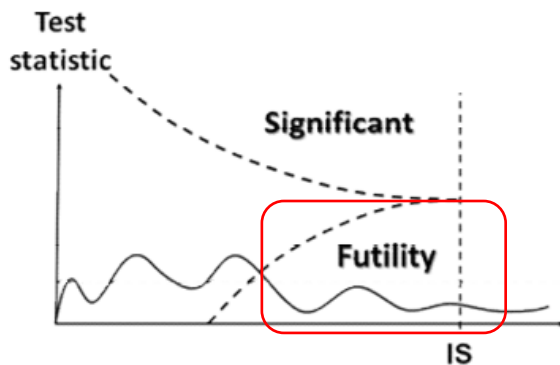
The confidence interval can be adjusted as the statistical significance testing by replacing the threshold of the conventional statistical significance with the value obtained with the statistical monitoring boundaries.

The futility test or beta-spending function

The TSA software allows us to even assess when an intervention unlikely have an effect as large as the anticipated one (Thorlund K, Wetterslev J et al. 2011, Jakobsen, Wetterslev et al. 2014). In other words, when an intervention has an effect that is lower in effect than those considered minimally important or worthwhile for the patients. When a no effect results is found, researchers need to know if this is related to lack of power until an adequate information size is achieved or if the intervention do not have the effect as large as the anticipated. The socio-economical resources can be not indeed wasted in unnecessary further trials.

The software can build the ‘futility boundaries’ as a no effect threshold, which were originally developed in the interim analysis in RCT. They are calculated using a power function analogous to

the alpha-spending function for constructing the superiority and inferiority boundaries with application of numerical integration. This leads to early conclusion of no effect of an intervention. Below the futility threshold, the likelihood of a statistically significant effect is so low that becomes irrelevant. Therefore, the randomisation of new patients is not recommended since the intervention does not possess the effect wanted to be detected and further trials might be superfluous.



Limitation

Although the TSA is an easy tool to apply and it can be useful in handling the random error, it has its limitations. The TSA program it has been increasingly applied on dichotomous outcome however, less is its application on continuous outcomes. The reason for that is due to the fact that continuous outcomes are more complex to deal with (Guyatt, Thorlund et al. 2013).

In the presence of different tools for the same continuous outcome, effect measures are expressed in standardized mean difference. The TSA program does not comprehend meta-analysis of SMDs. The effect size expressed in standard deviation units is not easy and clear to interpret to most of clinicians and is therefore prone to produce unrealistic information size requirements. As a consequence, the use of effect measures expressed in mean difference is the only acceptable and plausible method in the light of the most useful clinical interpretation.

Another limitation is related to the right setting for type I and II errors, which is always debatable. The choice of a more stringent alpha level or a different level of power could vary the scenario and release different size of required sample. This is however true not only for the TSA approach but even in the process of the sample size calculation in single trials.

The dependence of the required information size on estimates of the proportion of events in the control group and on the anticipated intervention effect size can be seen limiting. Indeed, a low proportion of the event in the control group or small effect size would result in adjusted threshold for statistical significance highly conservative or less conservative in presence of larger effect size to detect. Moreover, adjusted for heterogeneity might be sometimes a limitation since is difficult to guess and estimate its value when only few trials are available.

One other concern is the challenge of the choice of the adequate anticipated intervention effect, worthwhile for the patients. It should be an adequate threshold to balance intervention's harm and benefits. The role of TSA is to help clinicians to create the most plausible, reliable and conservative scenario given a clinical question, limiting misleading interpretations. TSA can be trustful as a meta-analysis might be if only the variables are supported by a rationale: authors should consider if the anticipated intervention effect is the most plausible for the patients. Moreover, minimal anticipated intervention effect might be not similar across included trials. Inference on conclusiveness of a meta-analysis can only be valid and generalized to the patients' population for which the minimal important difference is referred. Therefore, TSA results are just valid under the assumption given by the clinical question.

TSA can handle the random error however, other sources of bias can affect the meta-analysis results. Systematic biases are important variables that must be considered when assessing harms and benefits of an intervention. The intrinsic risk related to the flow in methodology conduction cannot be eliminated so far through he TSA. This analysis can be a valuable tool to prevent false negative or false positive results but does not in any way solve the issue of systematic bias. This is however valid in each situation where a recommendation has to be taken starting from primary studies.

Alternative methods to the TSA – Examples

The TSA we have explained so far is just one of the several techniques that exists in literature and that can control the risk of random error in the context of sparse data and repeated updates in cumulative meta-analysis. I am going to briefly report some examples of alternative techniques. However, a consensus about the necessity to use these techniques has not been reached yet: some authors claimed that the meta-analysis does not have usually the control of the generation of new evidence so that stopping rules and sequential methods are consequently not applicable. Others researchers support the view that a meta-analysis is useful to achieve recommendations on benefit

and harms of intervention so, sequential methods are instead important as for individual primary studies (Higgins, Whitehead et al. 2011). Methods to avoid the random error beside the frequentistic Trial Sequential Analysis include the sequential meta-analysis, the law of the iterated logarithm, and Bayesian analyses (i.e. semi-Bayesian analyses; full Bayesian analyses). Nevertheless, consensus on which method is the more reliable and adequate has not been reached yet (Bender, Bunce et al. 2008, Higgins, Whitehead et al. 2011).

- Law of iterated logarithm

In this approach, the standardized test statistic is ‘penalized’ with a factor λ to account for the multiple testing and for the estimation of heterogeneity between RCTs when applied to a random effect model (Hu, Cappelleri et al. 2007). The adjustment factor λ is determined through simulation with the aim to control the significance level under different scenarios. The value of λ factor, for example, is equal to 1.5 when controlling the overall type I error at the desired level for a maximum of number of inspection in a cumulative meta-analysis up to 25. This approach however does not control the type II error. As a matter of fact, this method let researchers control the type I error for a broad range of situations simultaneously but it may have less power since it is not a method calibrated for a specific situation.

- Sequential meta-analysis

Sequential meta-analysis (SMA) following Whitehead’s boundaries approach (Higgins, Whitehead et al. 2011). In a SMA each randomized control trial contributes with two values: z , as the measure for the effect size in that RCT and v as the amount of information in the RCT which is proportional to the number of patients included. After each RCT, the total amount of information cumulates in $Z=\Sigma z$ and $V= \Sigma v$. Just as with the alpha-spending and beta spending based boundaries, the sequential method for monitoring the amount of information produce superiority, inferiority, and futility boundaries. These boundaries are built through a test, called triangular test, to yield the minimum possible risk of committing an error (balancing the type I error and II error). In the context of medical research, conventional theory does not support this balance; prevention of alpha error has always been considered more important.

The advantages of this technique are: no prior estimate for total information size is necessary; stopping a cumulative meta-analysis for futility is an option; the power can be quantified; point and

interval estimates are adjusted for the multiple testing. Nevertheless, this technique does not consider the heterogeneity between trial and does not provide an information size.

- Semi- Bayesian procedure

The semi-Bayesian procedure allows for adding the concept of heterogeneity to the frequentistic sequential meta-analysis (SMA). The Bayesian theorem is proposed to make inference about the value of the heterogeneity: an informative prior distribution might contribute to produce a realistic estimate in the early stages of a sequential meta-analysis. A simulation study showed that a semi-Bayesian approach can address the underestimation of the heterogeneity in the sequential meta-analysis, which may lead to smaller confidence intervals (Higgins, Whitehead et al. 2011).

- Bayesian approach

The Bayesian approach is based on a different philosophy: it expresses results in terms of probability and incorporates the external evidence. Initial uncertainty is expressed through *a prior distribution* about the quantities of interest. Current data and assumptions concerning how they were generated are summarized in the *likelihood*. The *posterior distribution* for the quantities of interest can then be obtained by combining the prior distribution and the likelihood. The posterior distribution may be summarized by point estimates and credible intervals, which look much like classical estimates and confidence intervals. The prior distribution may be an expression of subjective belief about the size of the effect (it can be based even on information from non randomized studies) whereas the likelihood represents both the data from the studies included in the meta-analysis and the analytic model (i.e. fixed or random effects) (Higgins JPT and Green S 2011).

In **table 1** the comparison of the rationale, advantages and disadvantages of each method.

METHOD	RATIONALE	ADVANTAGES	DISADVANTAGES
Trial Sequential Analysis (TSA)	TSA is a technique that combine the required information size (RIS), a conventional threshold for a statistically significant treatment effect and the alpha-spending monitoring boundaries with adjusted statistical thresholds.	This approach can control both the type I and II errors. It can provide an information size adjusted for the heterogeneity between trials. It can declare early the ‘futility’ of an intervention. Easy-to-use free software	Continuous outcome: impossible deal with SMD. Challenge of defining the plausible parameters required for information size calculation. The process must be prospective not retrospective.
Law of iterated logarithm	The standardized test statistic is ‘penalized’ with a factor λ to account for the multiple testing and for the estimation of heterogeneity between RCTs when applied to a random effect model	This approach can control the type I error for a broad range of situations simultaneously.	This approach does not control the type II error: it may have less power since it is not a method calibrated for a specific situation. Dependency of the value λ on the number of inspections
Sequential meta-analysis (SMA)	Each randomized control trial contributes with two values: z , as the measure for the effect size in that RCT and v as the amount of information in it. The sequential method for monitoring the amount of information produce superiority, inferiority, and futility boundaries.	This approach does not calculate an information size but it can quantify the power. Balance between type I and II errors.	Conventional theory does not support this balance: prevention of alpha error has always been considered more important. Poor estimation of the heterogeneity assumed as in a conventional meta-analysis.
Semi-Bayesian procedure	A sequential meta-analysis which incorporates the heterogeneity. The Bayesian theorem is used to add a priori distribution to the frequentistic sequential approach.	Add the concept of updating the heterogeneity through the Bayesian thinking.	Not inform about a required information size. Challenge in defining the priori distribution for heterogeneity.
Bayesian approach	The Bayesian approach incorporate multiple prior distributions with different anticipated distribution of intervention effects (i.e. a skeptical, a realistic and an optimistic prior distributions)	Results are communicated as probabilities. It can incorporate external evidence.	Challenge of defining a specific alternative hypothesis, the variation in the prior distributions of the intervention effect and the heterogeneity. Require deep knowledge of statistics.

Case analysis 1: TSA in low back pain rehabilitation

Method

Case study

We re-analysed a systematic review published by Saragiotto and colleagues in 2016 on The Cochrane Library which addressed the efficacy of motor control exercise as rehabilitative intervention for low back pain in adults (Saragiotto, Maher et al. 2016). Among low back pain rehabilitation interventions, exercise has been showed having a moderate effect in reducing low back pain and motor control exercise is one of the most common treatment used by clinicians. This intervention consists on specific exercises focused on the activation of the deep trunk muscle with the aim to restore the control and the coordination of these muscles. The treatment is often 1-1 supervised and the program range from 20 days to 12 weeks, with a median of 12 sessions, from one to five per week (Saragiotto, Maher et al. 2016).

Our attention was limited to pain intensity, a continuous outcome, found to be the most reported in low back pain clinical trials, followed by disability, range of motion, and quality of life (Castellini, Gianola et al. 2015). Therefore, we selected this Cochrane review as pain intensity is the primary outcome and effect estimates in its meta-analyses were expressed in mean difference with naïve 95% confidence intervals. The first meta-analysis in the review was selected.

Extraction data

We extracted the following data from the meta-analysis: type of control intervention, time at follow up, population, overall meta-analysed effect and its confidence interval, heterogeneity between the results of the trials, and meta-analytic technique used. We searched in the review's method section for the minimal important difference the authors considered as threshold for the clinically judgment of findings.

TSA scenarios

We conducted the analyses using individual trial data from the selected meta-analysis (Saragiotto, Maher et al. 2016). We considered as a standard model the Trial Sequential Analysis based on the

following parameters: the anticipated intervention effect defined by the authors of the review (Saragiotto, Maher et al. 2016); the standard deviation of the mean difference between intervention groups (assumed taking into consideration the pooled meta-analysis estimates of all trials regardless the bias risk); an alpha level of 5%, a beta of 20%; the estimated diversity of the meta-analysis and the meta-analytic model (random-effects model) as those used in the original review.

We aimed to offer four scenarios in which a different value for each of the parameters (anticipated intervention effect, alpha level or type I error, beta level or type II error and the heterogeneity) would be tested to calculate the DARIS. We changed each parameter at a time until providing the last scenario as the most conservative one. Following are the amendments:

1. The first scenario presented DARIS calculations on the basis of a smaller anticipated interventions effect. We chose the upper limit of the confidence interval closer to the null effect as the least likely possible effect and so the most conservative one. This conservative approach allows to comprehend whether an intervention can produce an effect, or at least, the least likely magnitude even though not relevant. If the null effect is included in the TSA-adjusted confidence interval, there will be very low probability that the assessed intervention would have actually an effect.
2. In the second scenario, the DARIS calculation changed according to the choice of lower alpha level of 1%.
3. The third scenario showed the changes induced if the power was increased to 90%.
4. The last scenario considered the variation of the diversity (D^2) among trials. We hypothesized the worst estimated diversity based on the 95% confidence interval of I^2 in the meta-analysis.

Results

Characteristics of the meta-analysis

The meta-analysis included 13 randomised clinical trials with a total number of patients of 872 suffering from chronic low back pain (Figure 1). The main intervention was motor control exercise while the control intervention included other exercises (i.e., general or conventional exercises, stretching, McKenzie). The follow up was at 3 months from randomisation. Pain intensity was measured with different tools, however, the review authors converted the measurements into a common 0 to 100 scale. Thus, the meta-analysed effect was expressed in mean difference (MD). The

meta-analysed effect size was -7.53 points favoring motor control exercise with a 95% confidence interval from -10.54 to -4.52 points. Heterogeneity was moderate, $I^2 = 43\%$.

The authors defined an effect clinically important when “*the magnitude of the effect size was at least medium (>10% of the scale)*”. Accordingly, they concluded that “*for the outcome pain, there is low quality evidence (downgraded due to risk of bias and publication bias) that there is a small, but not clinically important, effect of motor control exercise (MCE) for reducing pain at short term (mean difference (MD) -7.53 ; 95% confidence interval (CI) -10.54 to -4.52 ; P value < 0.001 , 13 trials) compared with other exercises*”.

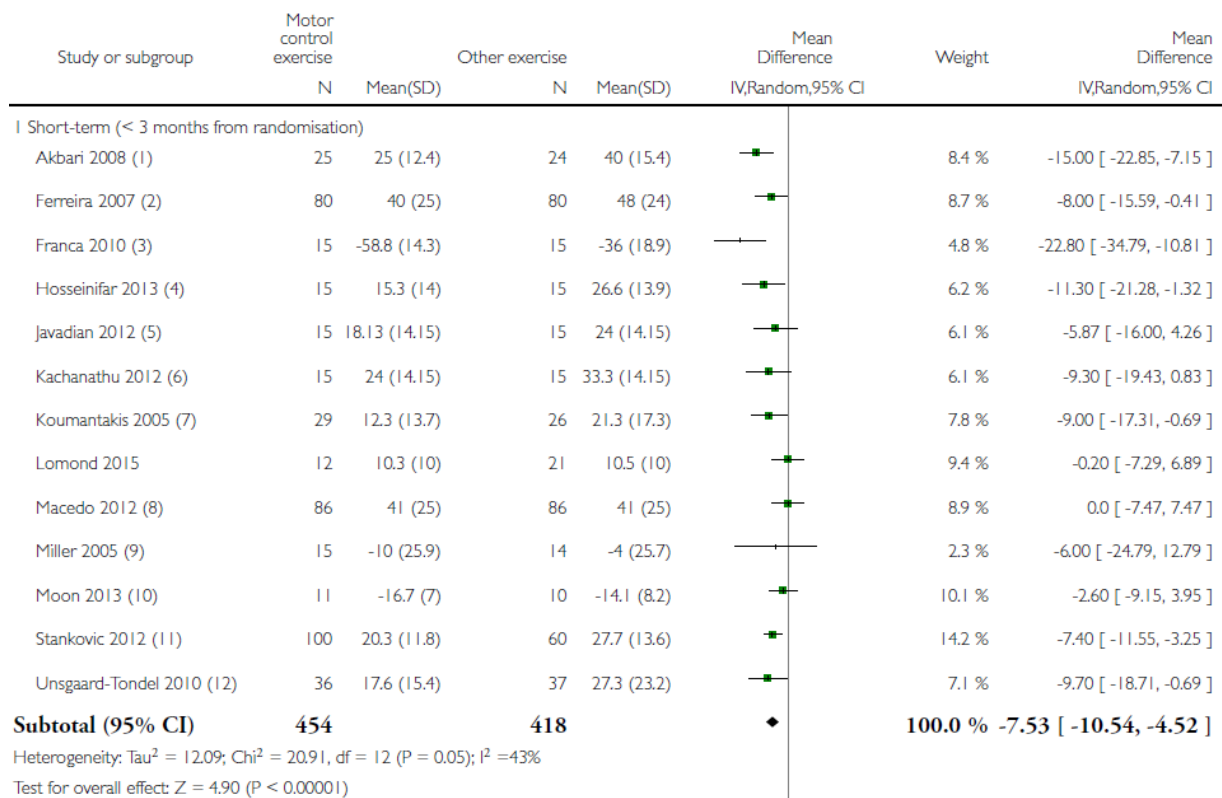


Figure 1. Meta-analysis on the efficacy of the motor control exercise compared to other exercises on pain intensity as outcome.

Results of the TSA scenarios

The first scenario, Figure 2, was obtained imputing the parameters adopted in the review (Saragiotto, Maher et al. 2016). The required information size was calculated based on an anticipated intervention effect of 10 point on a visual analogue scale (VAS) 0–100 (10% of the scale). The review' authors identified a clinically relevant effect when an effect size >10% of the total points of the VAS tool measurement is obtained. Generally, a minimal important change of at least 15 points or 30% on the VAS as outcome measure is suggested as a meaningful effect. Nevertheless, a universal consensus has not been reached yet since a paper found that an analgesic effect of 10 mm on a 100 mm visual analogue scale can represent a 'minimal effect'(van Tulder, Malmivaara et al. 2007). Actually, different clinical settings might reflect different meaningful effect and so different minimal important differences (Zanoli 2005). Despite the contrast in the literature, we would select 10 points out of 100 for our measurements.

The required information size was 323 patients. The cumulative Z-curve crossed the monitoring boundaries during the fourth trial and surpassed the required information size with the fifth trial. The fifth trial ended up with a reliable conclusion that the experimental intervention is able to statistically significantly detect a clinical difference of 10 point on a VAS scale compared with the control intervention (other exercises). All trials reported after the fifth are unjustified as the required information size to detect such a difference has been already reached, along with the statistical significance.

- Scenario 1. Anticipated intervention effect

We set the anticipated intervention effect of 4.52 points as being the lower limit of the 95% confidence interval of the overall effect size in the meta-analysis. All the other parameters remained unchanged. The RIS increased up to 1580 patients since the effect of the intervention that we would like to observe was smaller. The cumulative z-curve crossed the monitoring boundaries for benefit with the fifth study: the intervention yielded an effect that is both statistically and clinically significant. From the 6th to the 13th study the results were confirmed given a narrower confidence interval. Since the required information size has not been reached, the CI 95% had to be adjusted for the alpha spending function and it ranged from –11.98 to –3.07. Looking at the diversity (49%), we could conclude that the intervention does have an effect but there is a moderate level of heterogeneity across studies. Further trials are needed in order to amplify the confidence in the results. Figure 3.

- Scenario 2. Alpha level

In figure 4, we run a TSA setting a lower alpha level of 1%. The scenario slightly varied. Graphically, the monitoring boundaries for benefit and futility had the contact point on the vertical line of RIS at $Z=2.58$. The required information size increased up to 2352 and the monitoring boundaries for benefit were crossed after the tenth trial. The diversity-adjusted confidence interval widened from -13.62 to -1.44 . The null effect is still excluded therefore it confirmed that the intervention can provide an effect at least as the least likely effect of 4.52 points. Nevertheless, further trials are needed to reach a conclusion with the 95% of certainty.

- Scenario 3. The beta level

Using the beta level of 0.10 (corresponding to a 90% of power ($1-\beta$)), the diversity adjusted required information size increased to 2996 patients. The lower the type II error chosen, the larger is the sample size. In this scenario, the monitoring boundaries for benefit were surpassed with the last trial but the DARIS has not been reached yet. The diversity-adjusted confidence interval become wider, from -12.28 to -2.77 . Figure 5.

- Scenario 4. The diversity

In figure 6, we assume a plausible value of D^2 of 59%. This percentage was selected observing the 95% confidence interval of the heterogeneity in the meta-analysis, ranging from 20% to 59% and assuming its upper limit as the highest acceptable value of diversity in our analysis. The DARIS now became 3685 participants. The cumulative Z-curve touched only the conventional naïve statistical threshold but not the monitoring boundaries for benefit. Indeed, the diversity adjusted monitoring boundaries included the null effect ranging from -19.81 to 4.76.

Figure 2. Trial sequential analysis of motor control exercise versus other exercises. Outcome: pain. Original scenario. Diversity-adjusted required information size (DARIS): minimal important difference (MID) 10 points; standard deviation of the mean difference of 22.77; a 5%, b 20%, and diversity 49.58%.

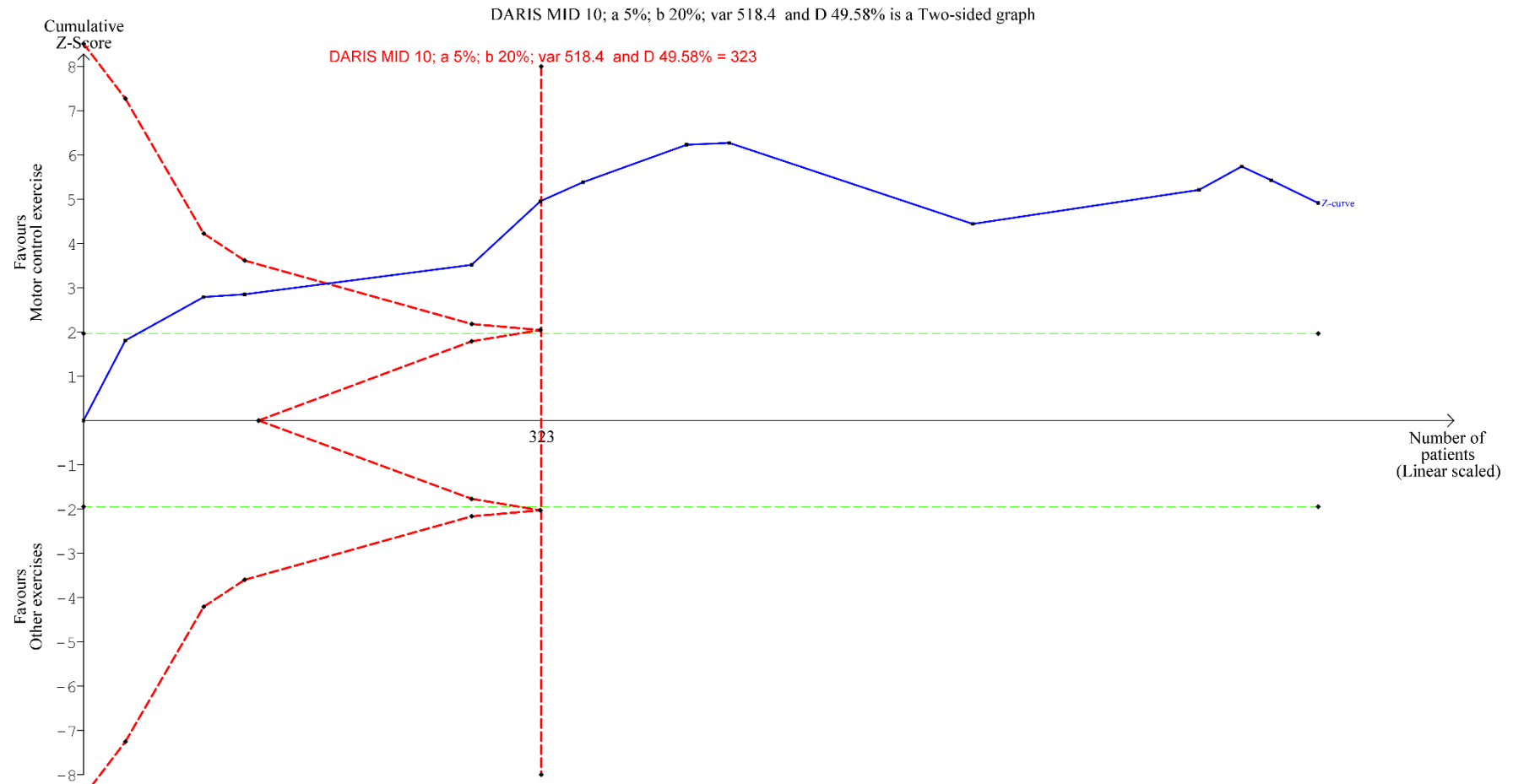


Figure 3. Trial sequential analysis of motor control exercise versus other exercises. Outcome: pain. Scenario 1. Diversity-adjusted required information size (DARIS): minimal important difference (MID) 4.52 points; standard deviation of the mean difference of 22.77; a 5%, b 20%, and diversity of 49.58%.

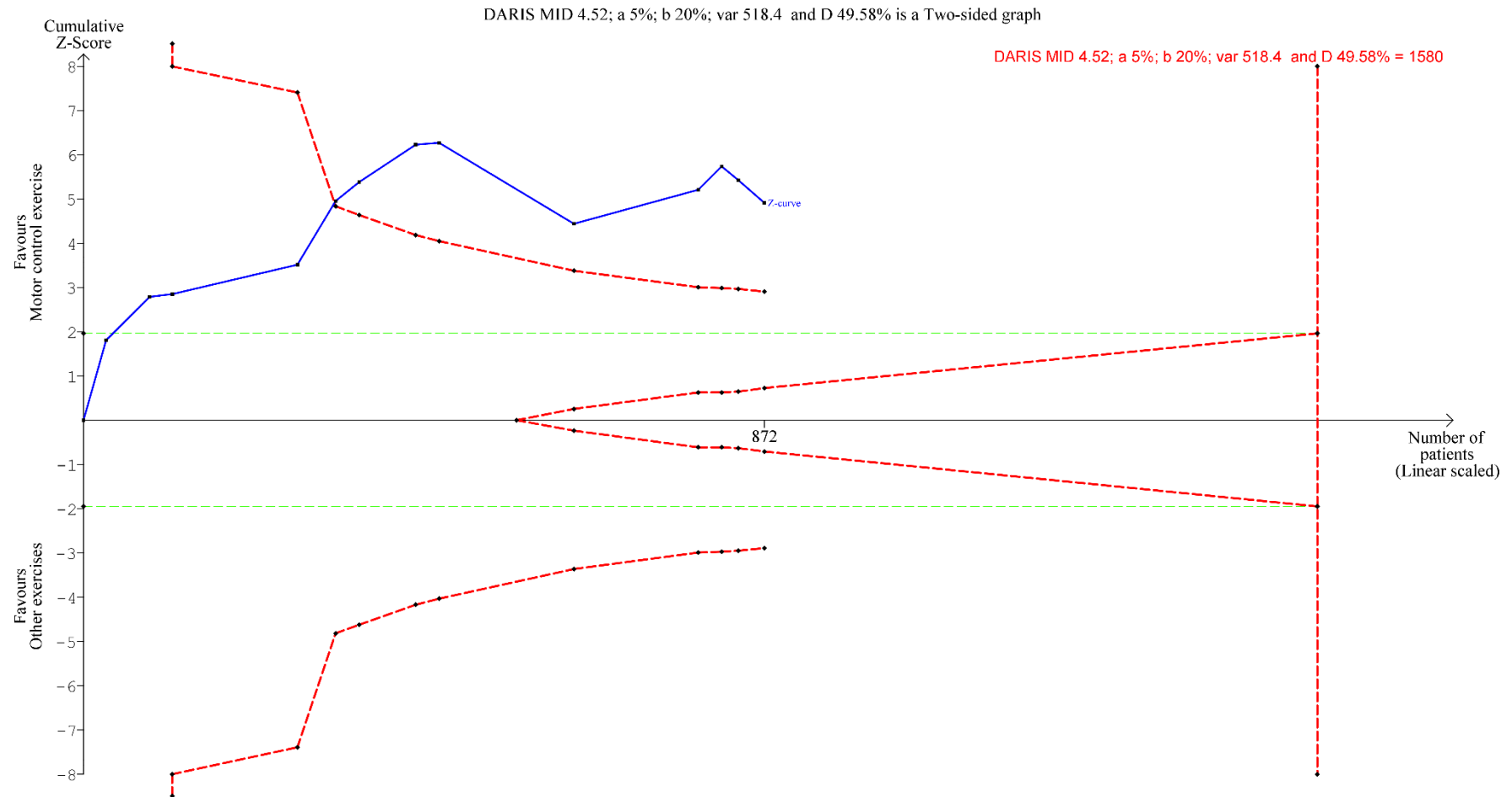


Figure 4. Trial sequential analysis of motor control exercise versus other exercises. Outcome: pain. Scenario 2. Diversity-adjusted required information size (DARIS): minimal important difference (MID) 4.52 points; standard deviation of the mean difference of 22.77; a 1%, b 20%, and diversity of 49.58%.

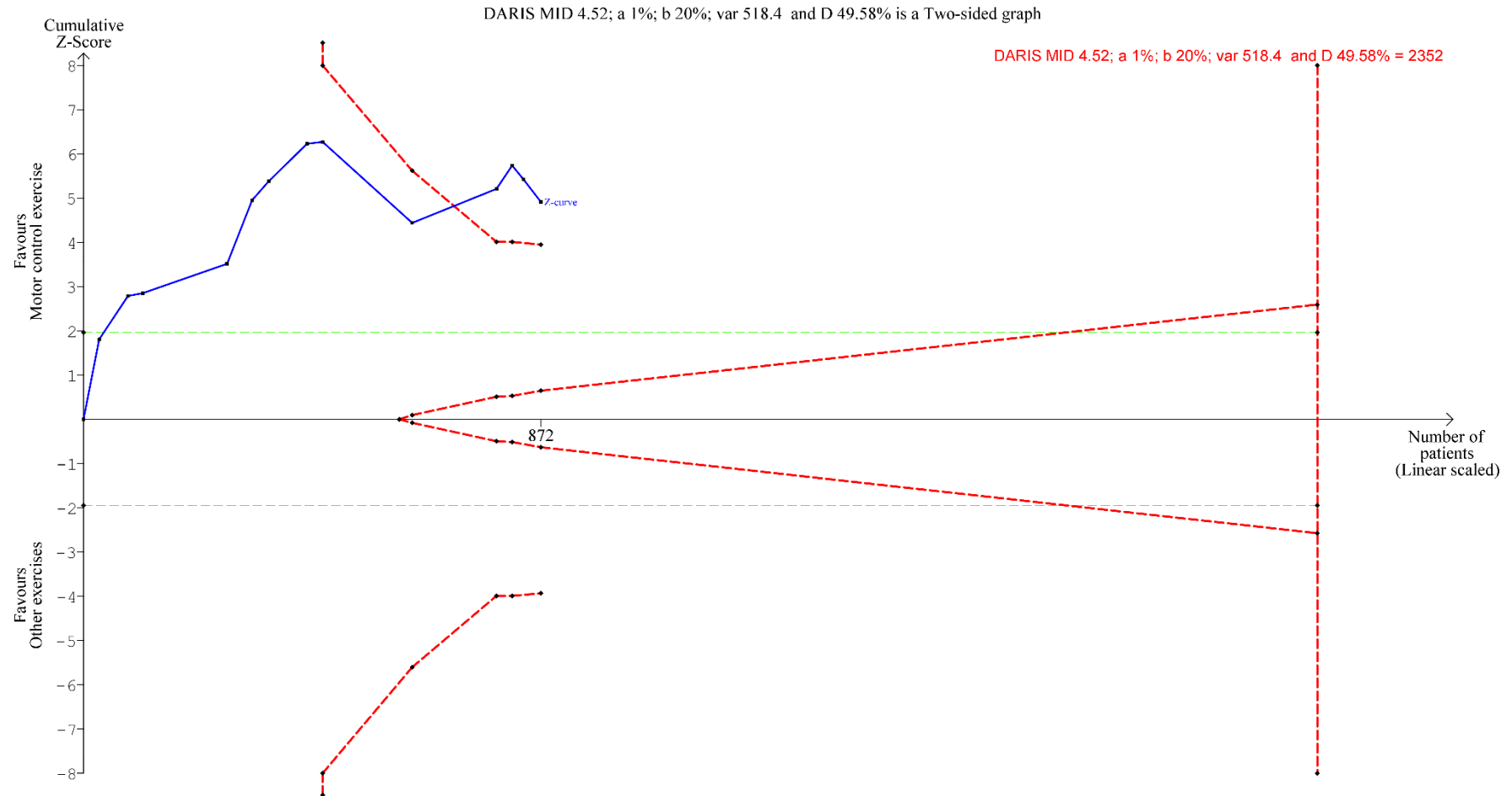


Figure 5. Trial sequential analysis of motor control exercise versus other exercises. Outcome: pain. Scenario 3. Diversity-adjusted required information size (DARIS): minimal important difference (MID) 4.52 points; standard deviation of the mean difference of 22.77; a 1%, b 10%, and diversity of 49.58%.

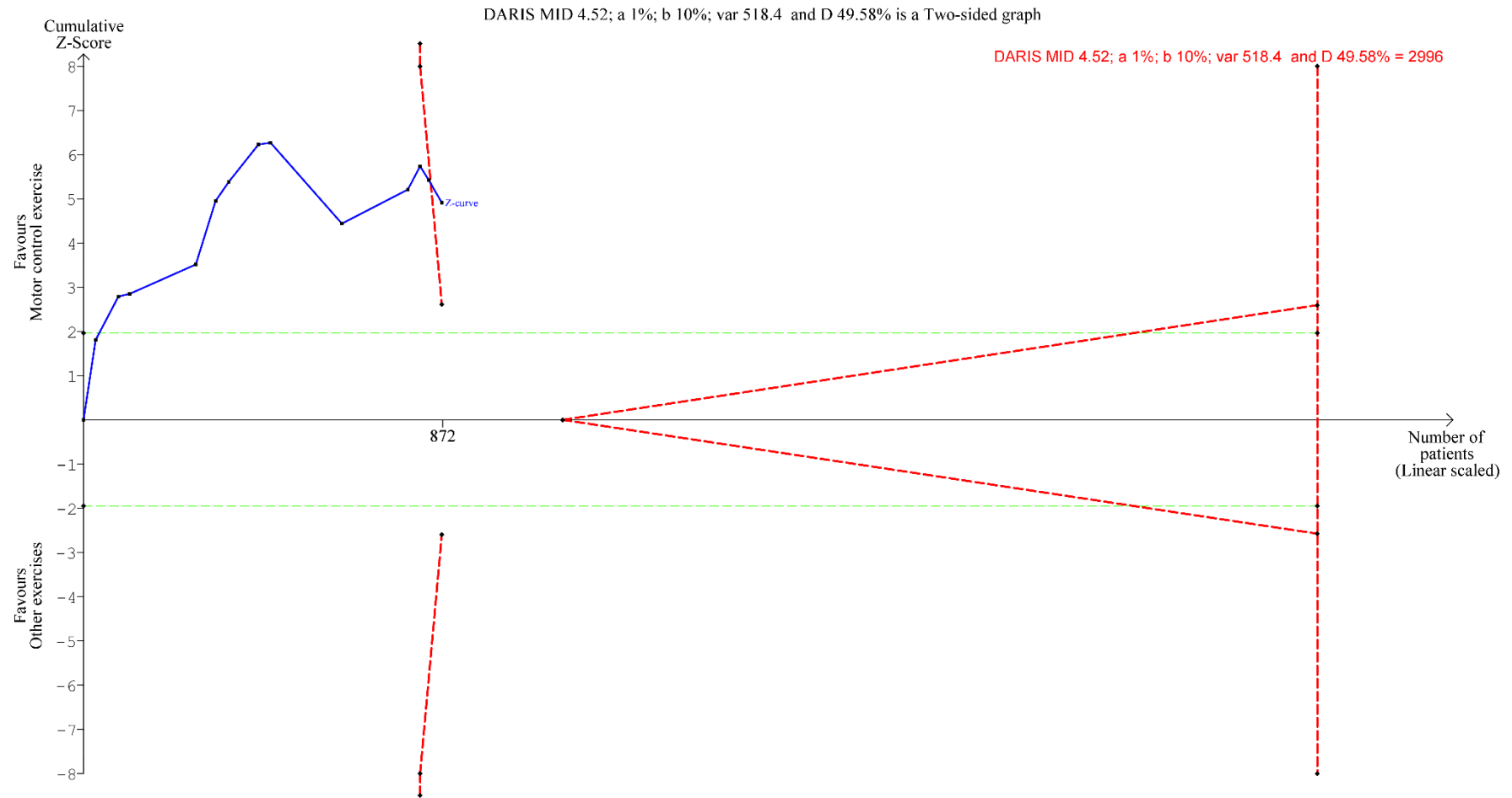
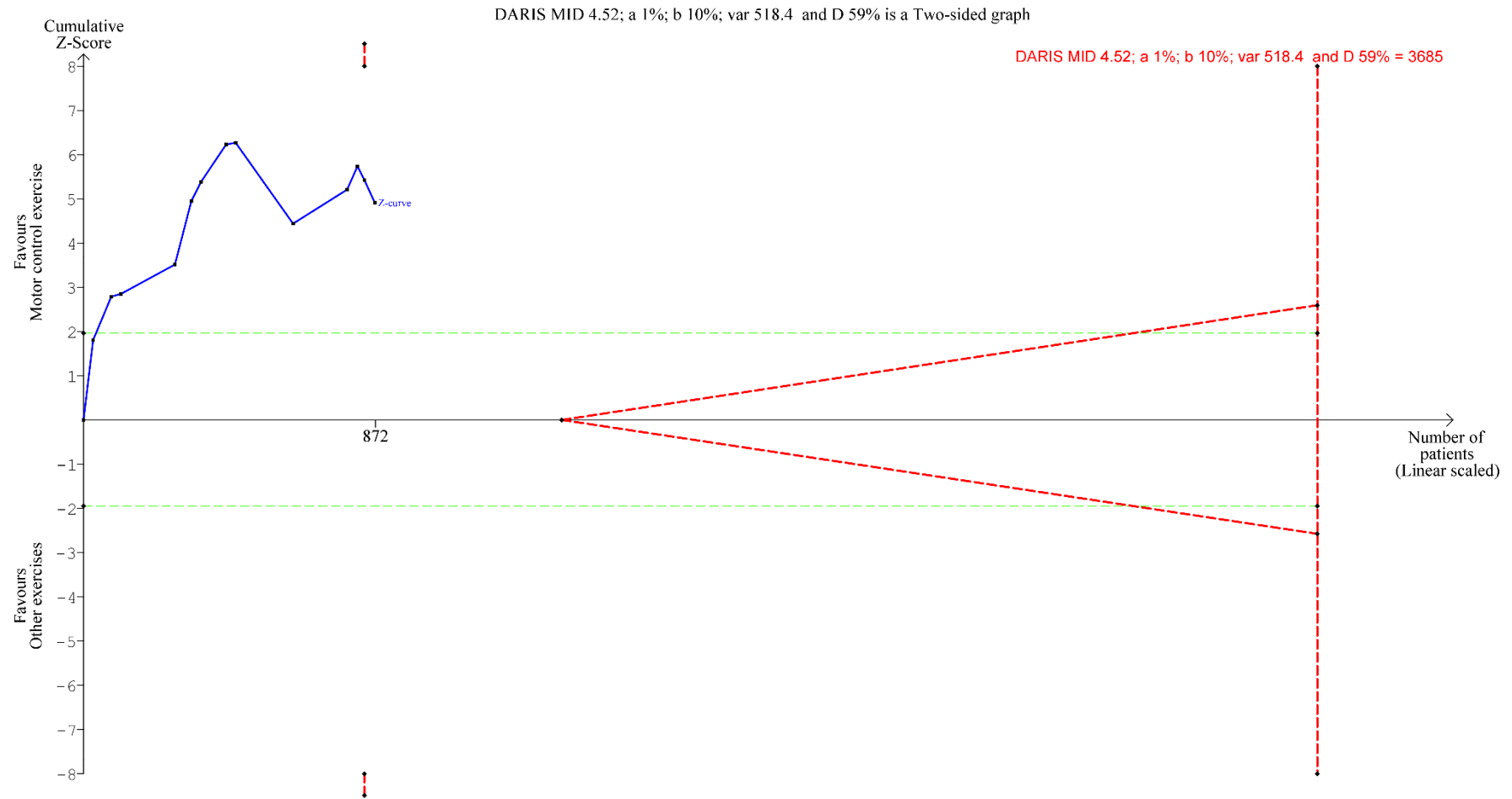


Figure 5. Trial sequential analysis of motor control exercise versus other exercises. Outcome: pain. Scenario 4. Diversity-adjusted required information size (DARIS): minimal important difference (MID) 4.52 points; standard deviation of the mean difference of 22.77; a 1%, b 10%, and diversity of 59%.



Discussion

We examined different TSAs scenarios on continuous outcome and demonstrated how a range of possible required information size for a meta-analysis can be yielded according to plausible treatment effects, different degrees of heterogeneity of the effect or different levels of type I and II errors. The selected meta-analysis found a statistically significant but not clinically relevant effect of the experimental intervention. Nevertheless, the TSA performed on the review author' assumption yielded a slightly different interpretation: the evidence achieved is enough to accept a statistical and clinical effect of 10% of the pain scale tool. Review' authors did not take into consideration the range of the 95% confidence interval which included an effect size of at least 10 but just the point estimate.

Overall, the amount of certainty and the size of RIS fluctuate as the range of possible desired effect and assumptions changes. The RIS ranged from a number of 323 in the less conservative scenario to 3685 patients in the most conservative one. As well, the confidence interval widened when the cumulative Z-curve has not reached the RIS until including the null effect when a higher amount of diversity is considered. Our first four scenarios all suggested benefit of the intervention but for very small and likely irrelevant intervention effects seen from a clinical point of view.

Anticipated a priori all the elements would avoid to fall in misleading results and lead to create the best plausible scenario for patient' needs. Therefore, what it should matter is that all the parameters needed to calculate the RIS must be predicted during the process of a systematic review protocol. A prospective and a priori application should be more in line with detecting false positive or false negative results.

The choice of the most plausible anticipated intervention effect, worthwhile for the patients, might be a challenge for clinicians and researchers since it should represent an adequate threshold to balance intervention's harm and benefits (Guyatt, Oxman et al. 2011). The smaller is the size of the anticipated intervention effect to be detect, the higher is the demanded RIS. In our example, although a minimal important change of at least 15 points or 30% on the visual analogue scale (VAS) is usually suggested as a meaningful effect (Ostelo, Deyo et al. 2008), a smaller effect may be beneficial for a population in a chronic painful stage. Our choice of using the least likely effect of 4.52 aimed to show whether the intervention can be beneficial even of a smaller intervention effect and, if so, if a wider effect can be plausible.

Choosing a lower value of statistical threshold as 0.01 allows for a smaller risk of false positive results, however a concern has been arisen regarding the problem of multiplicity in systematic reviews (Bender, Bunce et al. 2008). As a matter of fact, the statistical threshold is rarely adjusted for the number of the primary outcome used, rising concerns in results interpretation. Lately, some authors have suggested to address this issue adjusting a priori the statistical threshold for the number of primary outcome assessed (Jakobsen, Wetterslev et al. 2014).

Moreover, increase the level of the power to 90% is recommended in the view of the top of the hierarchy of the evidence (Garattini, Jakobsen et al. 2016).

A degree of heterogeneity has also to be defined. It is natural to expect an additional variation in meta-analysis due to the wide span of patients and interventions included compared with a single trial. However, confidence interval of a traditional meta-analysis does not point to the heterogeneity and its clinical implications. It has been emphasized this concept not only through the TSA application but also through a new approach based on predictive confidence intervals been introduced recently. This reinforces the need of providing interval that includes the heterogeneity, being able to illustrate range of true effects expected in future settings.

A further concern is represented by the choice of the control group: compare the experimental intervention towards other intervention rather than placebo or no treatment might in fact yield obfuscating results. Studies should compare the experimental intervention towards placebo or no treatment in order to be aware that the effect detected would be actually linked to the intervention itself, even though it seems that placebo can have a modest not meaningful effect on patient reported outcome as pain (Hrobjartsson and Gotzsche 2010). The effect on pain varied, even among trials with low risk of bias, from negligible to clinically important.

Other sources of bias can still remain. The systematic error can add uncertainty but results must be always interpreted in the view of the quality of primary studies. In rehabilitation field the risk of bias is still an important issue. A recent overview of reviews showed that RCTs in stroke rehabilitation lack of adequate quality in randomization, allocation concealment and blinding (patient, therapist, and assessor), which is often perceived as impossible to obtain (Santaguida, Oremus et al. 2012). In all scenarios we disregarded the risk of bias since the majority of the included trials presented high risk of bias in order not to add confounders in TSA interpretation. However, the intrinsic risk related to the flaws in methodology conduction cannot be eliminated so far: TSA can be a valuable tool to prevent false negative or false positive results but does not in any way solve the issue of bias. Disregarding the systematic error, our results provided

a useful example on how handle the risk of random error through the TSA: this methodology is a dynamic process which changes according to the amount of evidence already collected and the effect of intervention desired to detect. The scenarios obtained cannot be considered wrong or correct but just be different and interpreted in the light of the clinical question, taking into consideration the population assessed, the interventions compared and the point of view of the patients. Thus, TSA can be so viewed as a '*moving target*': adding a new trial, a new scenario would be created because the heterogeneity between trials can be different. What is more, a fluctuation in the scenario is more acceptable if the conclusion would be the most plausible and conservative one instead of support conclusion that can be unreliable due to not adequately adjusted statistical significance.

Case analysis 2: TSA in cardiovascular diseases

Published as: Comment on: “Cell Therapy for Heart Disease: Trial Sequential Analyses of Two Cochrane Reviews”. Castellini G, Nielsen EE, Glud C.
Clin Pharmacol Ther. 2017 Jul;102(1):21-24.

A second case analysis was made analysing the use of TSA on cell therapy for heart diseases by Fisher and colleagues (Fisher, Doree et al. 2016). The present article discusses the usefulness of Trial Sequential Analysis and its dependence on the choice of the parameters for calculation of the required information size and the adjacent monitoring boundaries, and comments on the approach by Fisher et al..

Comment

The usefulness of TSA is closely related to the choice of the assumed parameters for the calculation of the required information size and the adjacent monitoring boundaries. The calculation of the required information size for a dichotomous outcome is comparable to the sample size calculation in a single trial and requires assumptions of the expected proportion of patients with the outcome in the control group, the assumed risk ratio reduction of the experimental intervention, the desired maximum risks of both type I error and type II error, and the degree of heterogeneity.

Fisher et al. conducted Trial Sequential Analysis (TSA) on two Cochrane systematic reviews assessing bone marrow-derived cells for patients with acute myocardial infarction or heart failure (Fisher, Doree et al. 2016). The review authors highlighted that the effects of cell therapy as resulted in their systematic reviews could have been inflated by the underpowered meta-analyses (Fisher, Doree et al. 2016). In order to reduce the risk of random errors, Fisher et al. applied the TSA method on two dichotomous outcomes: all-cause mortality for both conditions; and rehospitalisation of the patients with heart failure.

We endorse the use of TSA method in systematic reviews, but we have some concerns regarding the parameters chosen by Fisher et al.. Below, we discuss each assumed parameter used to perform TSA on the outcome all-cause mortality for heart failure patients, but our considerations apply to all Fisher et al.’s performed TSAs on dichotomous outcomes.

The chosen proportion of death for the control group is 15%. This matches well with the observed proportion of deaths in Fisher et al.'s meta-analysis (15.9%; 50/315 patients in the control group). Thus, we do not contest their choice.

Fisher et al. anticipated an intervention effect of 35% risk ratio reduction (RRR) and justified this choice providing a reference to an equivalent RRR, associated with percutaneous coronary intervention for acute myocardial infarction (Hartwell, Colquitt et al. 2005). However, there are mismatches between this publication and the meta-analysis by Fisher et al. in terms of the patients and the interventions assessed. Hence, we do not know the real effect of the intervention. Therefore, it is important to choose the utmost possible realistic or plausible intervention effect in order not to 'falsely' lower the required information size which in turn may overestimate the 'proof' of the observed RRR. It is likely that the vast majority of interventions (over 95% of all interventions assessed) do not seem to have any positive intervention effects at all (Ioannidis 2005). Hence, a RRR of 10%, 15%, or 20% might have been a better choice. A 35% RRR seems unrealistically large and not plausible for the outcome death. Furthermore, a RRR of 25%, used by the authors for a sensitivity TSA, also seems unrealistic.

Fisher et al. selected an alpha level of 0.05 (type I error). The choice of a type I error ought to reflect the risks of multiplicity in systematic reviews following the number of the primary outcomes (Bender, Bunce et al. 2008). The direct consequence of multiplicity due to multiple outcomes is erroneous yield of conclusions in favour of tested interventions if any of the primary outcomes reaches statistical significance by chance alone. Referring to Fisher et al.'s most recent meta-analysis published in 2015, we assumed that only two primary outcomes (mortality and rehospitalisation) were considered (Fisher, Doree et al. 2015). This is why, by adjusting the threshold for significance according to the defined a priori number of primary outcomes, we can deal with multiplicity (Jakobsen, Wetterslev et al. 2014). This approach consists of dividing the alpha level for the value half away between 1 (no adjustment) and the number of the outcomes. So, for the review in question, this would be $0.05/1.5$ resulting in a type I error of 0.033.

The beta level or type II error that Fisher et al. used was 20%, equal to a power of 80%. When we are looking at the top of the hierarchy of the evidence (Fisher, Brunskill et al. 2014), a power of 90% (or more) seems better in securing an observation of the postulated effect if

present. This is a precaution against the risk of discharging a valuable treatment based on too little evidence.

Heterogeneity is another issue which needs to be taken into consideration. Fisher et al. found an I^2 equal to 0% in their meta-analysis and used this parameter for calculating the heterogeneity-adjusted RIS in their TSA. Nevertheless, an I^2 estimated among trials in an early meta-analysis can be defective (Wetterslev, Thorlund et al. 2009). The I^2 estimate might be unreliable due to lack of power and precision. Therefore, it can fluctuate over time. This is why we suggest assuming a higher heterogeneity in a TSA than the observed. This issue was not addressed by Fisher et al. Moreover, diversity (D^2) seems a more rational approach as it provides more valid results before interventions are turned into treatments (Wetterslev, Thorlund et al. 2009).

We replicated the TSA on all-cause mortality for heart failure patients done by Fisher et al. (Fisher, Doree et al. 2016). We calculated the diversity-adjusted required information size (DARIS) using a control event proportion of 15%, an anticipated risk relative reduction of 35%, a type I error of 0.05, a type II error of 0.20, and a D^2 of 0%. The DARIS was 1,236 patients, i.e., much higher than the already accrued number of only 759 patients (Figure 1 – Scenario A). The cumulative Z-curve surpassed the traditional monitoring boundary during the fifth trial and touched the trial sequential monitoring boundary for benefit at the eleventh trial. Fisher et al. concluded that the TSA results showed evidence of a reduction of the risk of mortality when cell therapy was administered in heart failure patients.

We conducted four TSAs scenarios changing one parameter needed to calculate the DARIS at a time (anticipated intervention effect, alpha and beta levels, and heterogeneity). We consider our values to be more plausible.

Figure 1, scenario B. Using a risk relative reduction of 20% and with all the other parameters unchanged from scenario A, we obtained a DARIS of 4,074 patients. The cumulative Z-curve surpassed the conventional statistical threshold at the fifth trial, but the last trial did not cross the monitoring boundary for benefit. The RRR is 0.42 and the TSA-adjusted confidence interval is between 0.16 and 1.09. The assumption of a smaller RRR alone leads to a conclusion less favorable towards cell therapy.

Figure 1, scenario C. Adjusting the alpha level for multiplicity to 0.033, we obtained a DARIS of 4,590 patients. The TSA-adjusted confidence interval widens to 0.07 to 2.34. We do not have enough information to provide a reliable conclusion of the effect of the intervention.

Figure 1, scenario D. Setting the power to 90%, we obtained a DARIS of 6,048 patients. The cumulative Z-curve is far from any monitoring boundary for benefit.

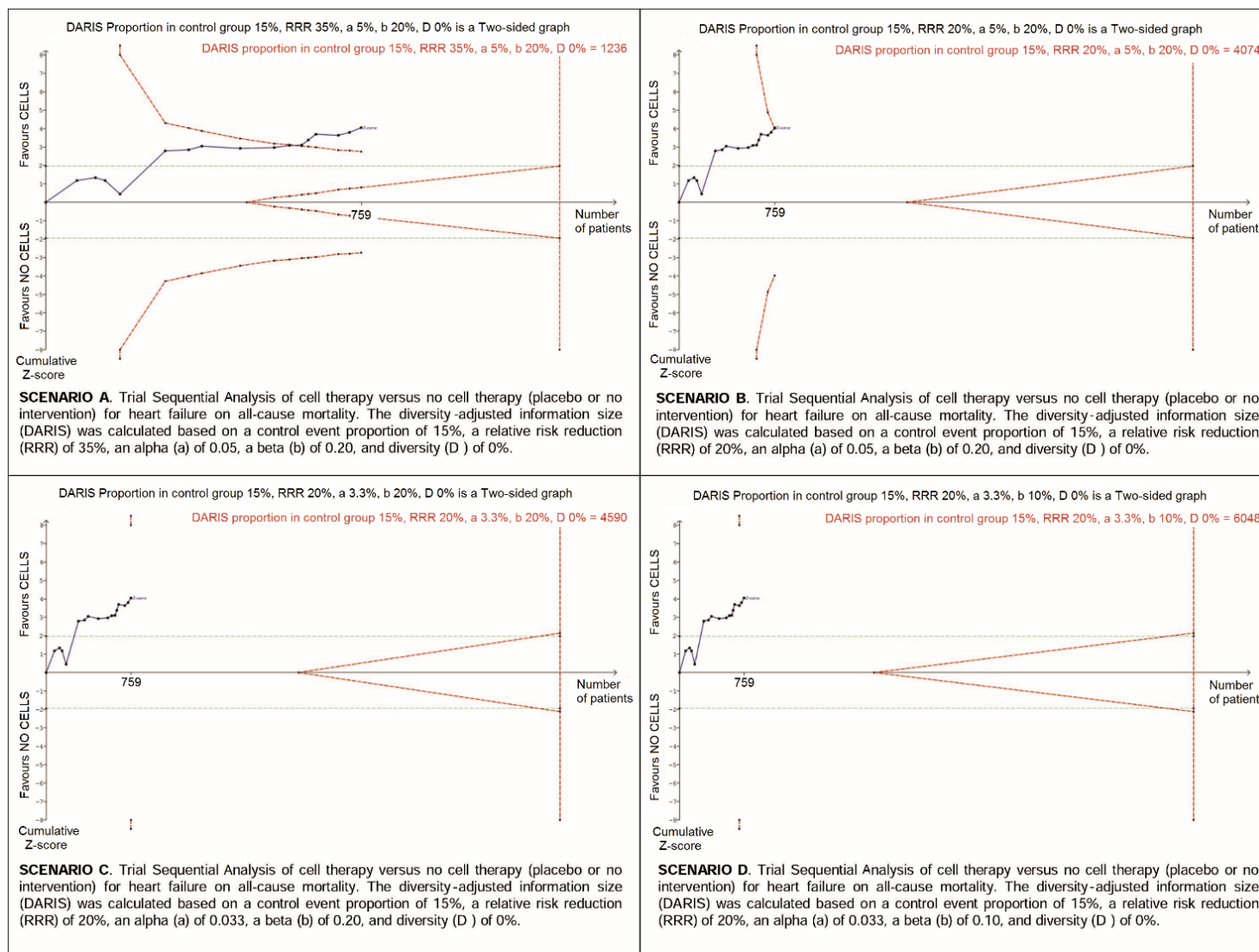


Figure 1. Trial Sequential Analysis of cell therapy versus no cell therapy (placebo or no intervention) for heart failure on all-cause mortality. Scenario A -D.

Figure 2. We assumed a plausible value of D^2 of 31%. This percentage was selected, observing the 95% confidence interval of the heterogeneity in the meta-analysis, ranging from 0% to 31% and assuming its upper limit as value in our analysis. We obtained a DARIS of 8,765 participants. The cumulative Z-curve crossed the conventional statistical threshold, but it was far from the monitoring boundaries for benefit, futility, and harm. The RRR of 0.42 does not change and the TSA-adjusted confidence interval remains wide, from 0.07 to 2.34.

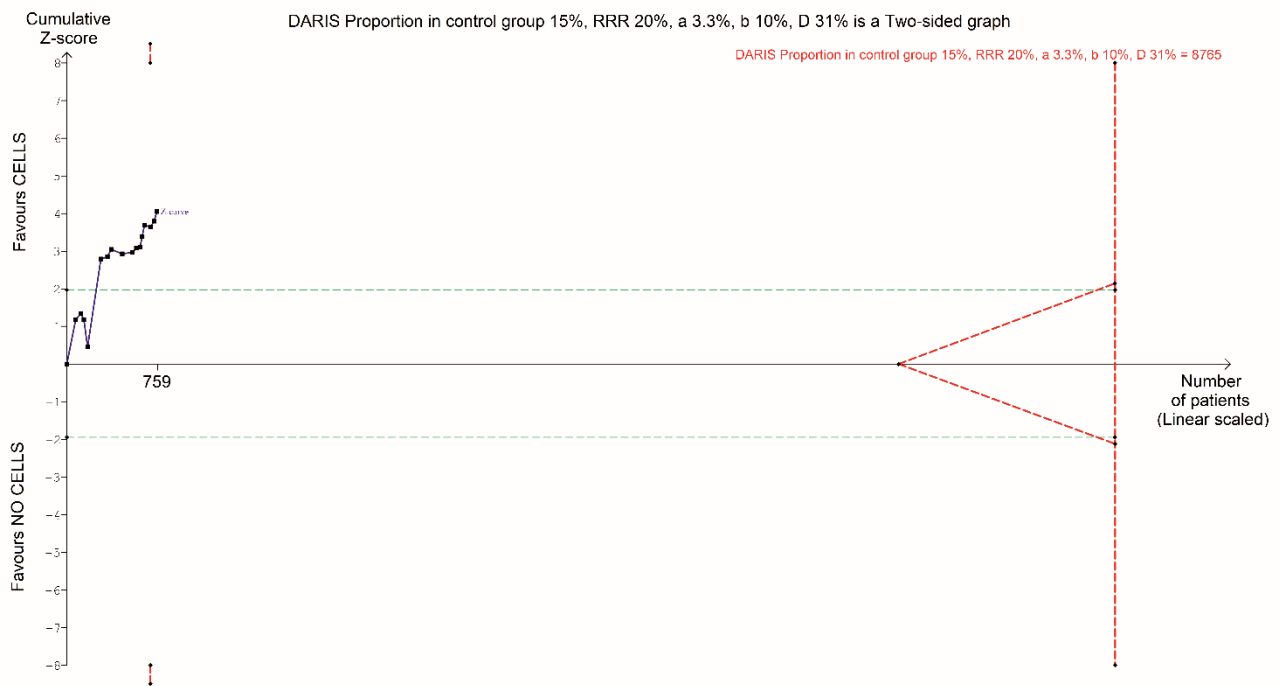


Figure 2. Trial Sequential Analysis of cell therapy versus no cell therapy (placebo or no intervention) for heart failure on all-cause mortality. The diversity adjusted information size (DARIS) was calculated based on a control event proportion of 15%, a relative risk reduction (RRR) of 20% an alpha (a) of 0.033, a beta (b) of 0.10, and diversity (D) of 31%.

In addition, Fisher et al. declared that “no adjustment for risk of bias” was performed. We should underline that TSA cannot adjust for risk of bias at all; it can only offer a method to control the risks of random errors. When the individual trials are at high risk of bias, the influence of systematic errors should be considered when interpreting the results. This makes the conclusions drawn by Fisher et al. on their TSA even less plausible.

In conclusion, the TSA is a powerful method which can inform readers and researches whether a sufficient amount of information has been accrued for a conclusion to be made, and it can roughly estimate how much more information is needed to accept or discard an intervention

effect (Kulinskaya and Wood 2014). TSA is not able to wash away systematic errors. Moreover, to provide the most realistic scenario for clinicians and patients, we strongly recommend deciding the assumptions and parameters for the calculation of the DARIS and its adjacent monitoring boundaries a priori. We did not do so in the present analyses, and critics may argue that the parameters we chose were too conservative. However, due to both risks of random errors and systematic errors in the conducted randomised clinical trials on stem cells for patients with heart failure, we need further randomised clinical trials conducted in accordance with the SPIRIT Statement.

General conclusion

Trial Sequential Analysis (TSA) is a statistical technique developed to control the risks of random errors and prevent premature conclusions of statistical superiority, no effect, or inferiority of experimental interventions in meta-analyses (Thorlund K, Wetterslev J et al. 2011). The TSA is a powerful technique which can inform readers and researches if a relevant amount of information has been accrued to make a conclusion. It can estimate how much more information is needed to accept or discard an intervention effect but cautiousness in defining the parameters should be taken (Kulinskaya and Wood 2014). Indeed, the usefulness of TSA is closely related to the choice of the assumed parameters for calculation of the DARIS and its adjacent monitoring boundaries. Therefore, to provide the most realistic scenario for clinicians and patients, deciding the assumptions and parameters of the TSA a priori is strongly recommended. What is to remember, however, is that TSA is not able to wash away systematic errors.

For each meta-analysis answered a clinical question, there should be parameters which best fit for it: this chapter offers a more didactic way to figure out how specific parameters can change the interpretation of findings according to what we want to detect. TSA can be trustful as a meta-analysis might be if only the variables are supported by a rational, given a clinical question: authors should consider if the anticipated intervention effect is the most plausible for the patients assessed and which is the degree of certainty one wants to detect.

Chapter 2. Precision of results: a focus on the GRADE system

Introduction

Since 2008, the GRADE approach (Grading of Recommendations Assessment, Development and Evaluation) has gained its momentum as an international standard to assess the strength of body of the evidence, informing transparently and explicitly the confidence that researchers have on the results. (Balslem, Helfand et al. 2011).

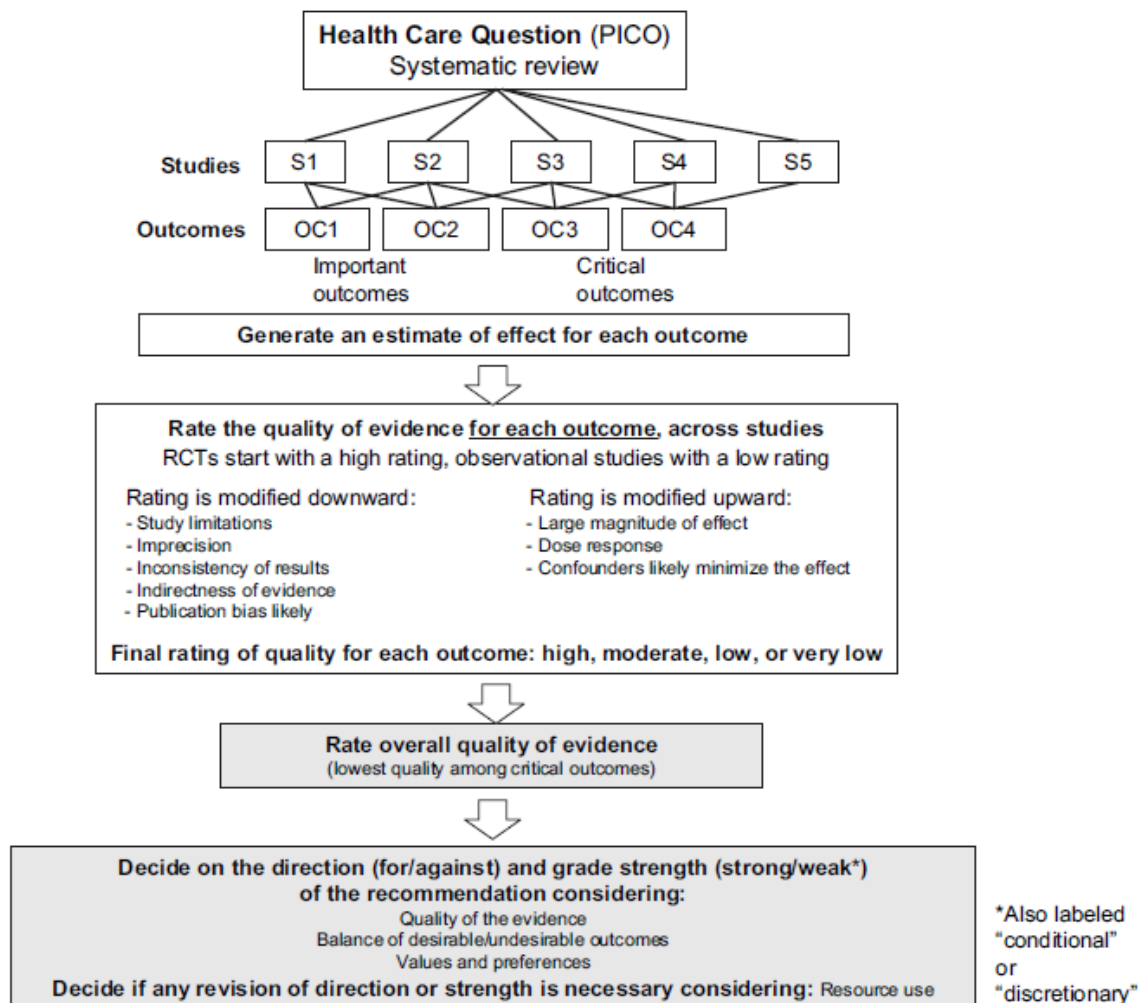
GRADE is a method for rating the quality of evidence in systematic reviews and grading strength of recommendations in guidelines. The m can be applied to several clinical questions ranging from diagnosis, prevention, therapy and to public health and health system questions. The add value of this system is that it offers a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines in health care and for carrying out the steps involved in developing recommendations.

The GRADE introduces the concept of the “quality of the evidence” which is a judgment about the extent to which we can be confident that the estimates of the effect are correct. Using the GRADE system, this judgment is carried out for each critical or important outcome for a patient and are based on: the study design (randomized trial vs observational study), the risk of bias, the precision of the overall estimate across studies, the consistency of the results across studies, the indirectness of results and the publication bias.

The GRADE approach is based on two steps:

1. Assessment of the body of the evidence, i.e. systematic reviews;
2. Formulation of recommendations.

Following, the schematic view of the GRADE process of grading the quality of the evidence and the strength of recommendations (Guyatt, Oxman et al. 2011):



Once systematic review or guideline authors decide the questions in terms of population, outcome, comparison and intervention (PICO), the process needs to define the critical or important outcomes for the patients. Then, the data from the eligible studies are used to generate the best estimate of the effect on each patient important outcome.

GRADE approach in systematic reviews

GRADE's approach begins with the study design. The randomized controlled trial design starts as high quality evidence whereas the observational study as low quality evidence.

The GRADE working group has selected five factors that might lower the quality of the evidence and three factors that might increase the quality. GRADE’s approach to rating confidence in effect estimates (quality of evidence) is shown in the following figure:

1. Establish initial level of confidence		2. Consider lowering or raising level of confidence		3. Final level of confidence rating
Study design	Initial confidence in an estimate of effect	Reasons for considering lowering or raising confidence		Confidence in an estimate of effect across those considerations
		↓ Lower if	↑ Higher if	
Randomized trials →	High confidence	Risk of Bias Inconsistency Indirectness Imprecision Publication bias	Large effect Dose response All plausible residual confounding and bias • Would reduce a demonstrated effect or • Would suggest a spurious effect if no effect was observed	High ⊕⊕⊕⊕
Observational studies →	Low confidence			Moderate ⊕⊕⊕○
				Low ⊕⊕○○
				Very low ⊕○○○

In the process of the quality of the evidence’ assessment, authors use this approach to rate the quality for each outcome across studies in a systematic review. GRADE is “outcome centric”, this means that each single important outcome is rated for quality and the quality might be different from one outcome to another. Before assessing the quality of the evidence, systematic reviewers and guideline developers should identify all potential patient important outcomes, including benefits, harms, and costs. Reviewers will then assess the quality of evidence for each important outcome.

As we can see in the figure above, the GRADE approach results in an assessment of the quality of a body of evidence as high, moderate, low, or very low which have the following meanings (Balshem, Helfand et al. 2011):

- High: we are very confident that the true effect lies close to that of estimate of the effect
- Moderate: we are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect but there is possibility that it is substantially different.
- Low: our confidence in the effect estimate is limited. The true effect may be substantially different from the estimate of the effect
- Very low: We have very little confidence in the effect estimate. The true effect is likely to be substantially different from the estimate of the effect.

The 5 domains which can lower the quality

1. Risk of bias

This domain concerns with the internal validity of study results. Risk of bias refers to systematic deviation of the results which is caused by the way the study is designed or conducted. In clinical trial, results can be inflated by particular source of bias and vary in direction: bias due to a particular design flaw (e.g. lack of blinding) may underestimate an effect in one study but overestimate it in another study. Systematic bias in randomized trials include allocation concealment, blinding (participants, personnel or outcome assessor), loss to follow-up and intention to-treat principle in results analyses. Moreover, limitations might include stopping early for apparent benefit and selective reporting of outcomes according to the results. Regarding observational studies, results might be inflated due to inappropriate controls and or inadequate adjustment for prognostic imbalance (Guyatt, Oxman et al. 2011).

2. Inconsistency

Inconsistency refers to differences of the magnitude of intervention effects across studies. In systematic reviews, authors should investigate the reason of these differences: explanations can lie in the population (e.g. age, severity of disease), interventions (e.g. frequency, doses), outcomes (e.g. in follow up) or study methods (e.g. risk of bias). If one or more categories provide an explanation, different estimates should be discussed across patients, outcomes or interventions. If large variability (often referred to as heterogeneity) in magnitude of effect remains unexplained, the quality of evidence decreases. Judgment of the extent of heterogeneity is based on similarity of point estimates, extent of overlap of confidence intervals, and statistical criteria including tests of heterogeneity and I^2 (Guyatt, Oxman et al. 2011).

3. Indirectness - Directness of Evidence generalizability, transferability, applicability

Direct evidence is when research directly compares the interventions in which we are interested on the population in which we are interested and measures the outcomes important to patients. When population, interventions, outcomes are different from those we are interested in we have indirectness (Guyatt, Oxman et al. 2011).

Indirectness can be about:

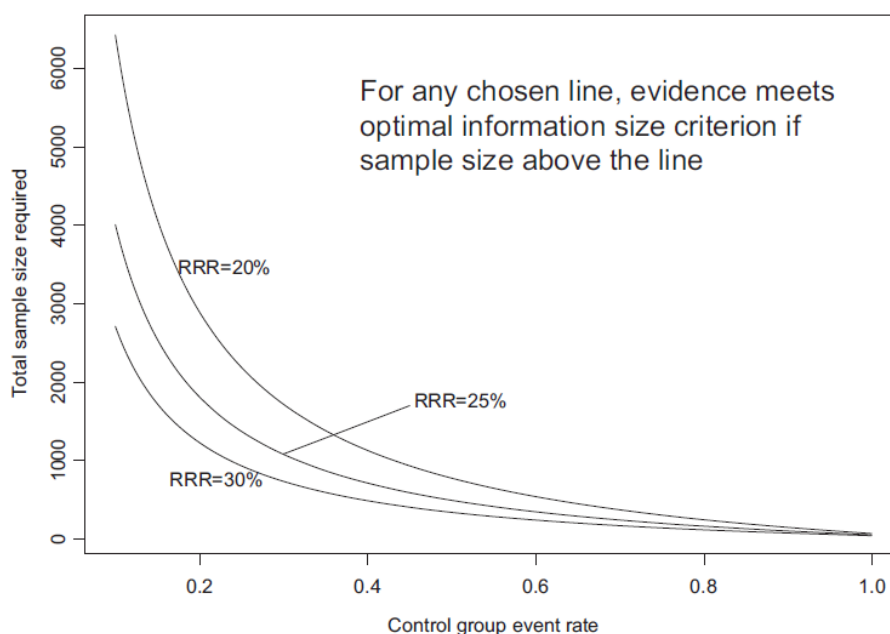
1. populations/patients (e.g. interested in children found adults population)
2. interventions (e.g. interested in high dosage, found low dosage, interested in long treatment, found short)

3. outcomes (e.g hip fracture vs bone density; interested in long term but found short term results).
4. interventions that have not been tested in head-to-head comparisons.

4. Imprecision

In a systematic review, imprecision refers to the confidence in the estimates of effect, whose extent is mostly captured by the confidence interval. Conceptually, the confidence interval (typically 95%) is the range in which the truth probably lies. In other words, it is the range of results which would include the true underlying value if an experiment is repeated numerous times and the confidence interval recalculated for each experiment.

When the confidence interval around the estimate of treatment effect is not sufficiently narrow, the quality of the evidence is rated down by one level whereas if the confidence interval is very wide, rated down by two levels. However, the confidence interval alone is not sufficient to judge imprecision because trials with small numbers of events may get the results fragile. Therefore, other two elements are to consider: (1) a clinical threshold of benefit or harm; (2) the calculation of the optimal information size. GRADE working group authors suggest: “if the total number of patients included in a systematic review is less than the number of patients generated by a conventional sample size calculation for a single adequately powered trial, consider the rating down for imprecision.” To calculate the optimal information size, Guyatt et al. suggest to use online calculators (i.e. <http://www.stat.ubc.ca/~rollin/stats/ssize/b2.html>.) or, as an alternative, they provide a figure (reported on the right) to consult in order to determine the optimal information size for risk relative reductions of 20%, 25% and 30% across varying control event rates.



Generally, this is the rule suggested in Guyatt et al. (Guyatt, Oxman et al. 2011) for dichotomous outcomes:

- If the optimal information size criterion is not met, rate down for imprecision, unless the sample size is very large (at least 200, and perhaps 4000 patients).
- If the optimal information size criterion is met and the 95% confidence interval excludes no effect (i.e. confidence interval around risk ratio excludes 1.0) precision is adequate.
- If the optimal information size criterion is met, and confidence interval overlaps no effect (i.e. confidence interval includes risk ratio of 1.0), rate down if confidence interval fails to exclude important benefit or important harm.

5. *Publication bias*

Studies suggesting a beneficial intervention effect or a larger effect size might have a higher chance to be published, while negative intervention effects might remain unpublished. In this situation, a systematic review of the published studies could identify a spurious beneficial intervention effect, or miss an important adverse effect of an intervention. Publication bias refers to the possibility that results are inflated by only positive studies, industrial sponsored trials or early studies which can overestimate effects when “negative” studies face delayed publication. Usually, publication bias is assessed through a funnel plot showing asymmetry of the included studies: smaller studies might be not symmetrically distributed around either the point estimate (dominated by the larger trials) or the results of the larger trials themselves (Guyatt, Oxman et al. 2011).

The three domains which can raise the quality in observational studies are:

1. Large magnitude of effect (RRR 50%/RR 2)
2. Dose response relation
3. Residual confounding

The endpoint of the GRADE approach: Evidence Profile and Summary of Findings Table

The Evidence Profile table and the Summary of Findings table are two approaches through which the GRADE working group intends to present the quality of the available evidence and the judgments

related to the quality rating. These two approaches have different purposes and are directed to different audiences:

Evidence profile table is a summary of the evidence for a given question with a detailed quality assessment and an explicit judgment of each factor which determines the quality. It is always used by guideline producers to agree about the judgments underlying the quality assessment.

Summary of Findings (SoF) tables presents, for each relevant comparison of alternative management strategies, the quality rating for each outcome, the best estimate of the magnitude of effect in relative terms, and the absolute effect that one might see across subgroups of patients with varying baseline or control group risks. The SoF does not include the detailed judgment and it is prepared within a systematic review.

GRADE and TSA in systematic reviews and meta-analysis

A recent publication by Jakobsen and colleagues suggest to include the TSA as a supplement of imprecision assessment of the GRADE approach (Jakobsen, Wetterslev et al. 2014). Both TSA and GRADE take into consideration the imprecision domain in systematic reviews. Consequently, to investigate how TSA can yield a different interpretation of this domain in meta-analysis results compared with those obtained by the GRADE approach could be captivating.

Imprecision, along with risk of bias, are the most common domains associated with GRADE downgrading of overall evidence quality or certainty (Pandis, Fleming et al. 2015). Systematic reviews employ multiple parameters to evaluate imprecision: accrued sample size, required or optimal information size (OIS) (meta-analytic ‘sample size’), alpha, beta, confidence intervals of the overall effect, and specified critical margins of ‘no effect’, ‘important benefit’ or ‘important harm’ (Guyatt, Oxman et al. 2011). GRADE combines all components in a simple rule: “*If the optimal information size criterion is not met, rate down for imprecision, unless the sample size is very large (at least 2000, and perhaps 4000 patients); if the optimal information size criterion is met and the 95% confidence interval (CI) excludes no effect, do not rate down for imprecision; if the optimal information size criterion is met, and the 95% CI overlaps no effect, rate down for imprecision if the CI fails to exclude important benefit or important harm.*” (Schünemann H 2013, Schunemann 2016)

The GRADE rules of thumb are based on broad assumptions and generalities across medical fields. The most relevant advantage is facilitating the trustworthiness of recommendations, enabling users to reflect on the sample as a basis for recommendations. However, rating imprecision in isolation,

without a formal evaluation of accrued sample and magnitude of effects (e.g. benefits or harms), would be hazardous (Anttila, Persson et al. 2016, Schunemann 2016). Because random errors are a frequent cause of erroneous estimation of treatment effect, often in small meta-analyses (Imberger, Thorlund et al. 2016), several authors have highlighted the need to adjust the statistical threshold and calculate a required information size in meta-analyses to increase the validity and reliability of its conclusions (Thorlund, Imberger et al. 2011, Jakobsen, Wetterslev et al. 2014). Among the techniques that can control for the risk of random error in the context of sparse data (Higgins, Whitehead et al. 2011), Trial Sequential Analysis (TSA) is often used to control for spurious findings (Simmonds, Salanti et al. 2017, Wetterslev, Jakobsen et al. 2017) and is currently suggested as a potential supplement for a more throughout assessment of imprecision when using the GRADE system (Jakobsen, Gluud et al. 2014).

The abstract was presented as oral communication at the 8th International Conference of EBHC Teachers & Developers hosted by GIMBE Foundation in October 2017 in Taormina (Italy) and at the 25th Cochrane Colloquium in Edinburgh (UK) in September 2018.

Imprecision Assessment: A Comparison between the GRADE System and the Trial Sequential Analysis

Published as: Castellini G, Bruschetti M, Gianola S, Gluud C, Moja L. Assessing imprecision in Cochrane systematic reviews: a comparison of GRADE and Trial Sequential Analysis.

Syst Rev. 2018;7(1):110. Published 2018 Jul 28. doi:10.1186/s13643-018-0770-1

Abstract

Background The evaluation of imprecision is a key dimension of the grading of the confidence in the estimate. GRADE gives recommendations on how to downgrade evidence for imprecision, but authors vary in their use. Trial Sequential Analysis (TSA) has been advocated for a more reliable assessment of imprecision. We aimed to evaluate reporting of and adherence to GRADE and to compare the assessment of imprecision of intervention effects assessed by GRADE and TSA in Cochrane systematic reviews.

Methods Cross-sectional study. We included 100 Cochrane reviews irrespective of type of intervention with a key dichotomous outcome meta-analyzed and assessed by GRADE. Methods and results sections of each review were assessed for adequacy of imprecision evaluation. We reanalysed imprecision following the GRADE Handbook and the TSA Manual.

Results Overall, only 13.0% of reviews stated the criteria they applied to assess imprecision. The most common dimensions were the 95% width of the confidence intervals and the optimal information size. Review authors downgraded 48.0% of key outcomes due to imprecision. When imprecision was reanalysed following the GRADE Handbook, 64% of outcomes were downgraded. Agreement between review authors' assessment and assessment by the authors of this study was moderate (kappa 0.43, 95% confidence interval [CI] 0.23 to 0.58). TSA downgraded 69.0% outcomes due to imprecision. Agreement between review authors' GRADE assessment and TSA, irrespective of downgrading levels, was moderate (kappa 0.43, 95% CI 0.21 to 0.57). Agreement between our GRADE assessment following the Handbook and TSA was substantial (kappa 0.66, 95% CI 0.49 to 0.79).

Conclusions In a sample of Cochrane reviews, methods for assessing imprecision were rarely reported. GRADE according to Handbook guidelines and TSA led to more severe judgement of imprecision rather than GRADE adopted by reviews' authors. Cochrane initiatives to improve adherence to GRADE Handbook are warranted. TSA may transparently assist in such development.

Keywords: Review, Meta-analysis, Bias, Confidence Intervals, Epidemiologic methods

Aim

We conducted an empirical assessment of a sample of Cochrane systematic reviews (SRs) in which the focus was imprecision as a threat to validity. We investigated the reporting of and the adherence to GRADE in assessing imprecision, the expected and observed downgrading of evidence, and the reasons for downgrading. Moreover, after having estimated the Cochrane authors' handling of GRADE and imprecision, we applied ourselves GRADE assessment of imprecision following the GRADE Handbook guidelines (Schünemann H 2013) and TSA following the TSA Manual (Thorlund K, Wetterslev J et al. 2011) to independently replicate the assessments of imprecision.

Methods

Search strategy

For this cross-sectional study, Cochrane systematic reviews were sampled from the Cochrane Database of Systematic Reviews (Cochrane). We purposively retrieved 100 reviews in reverse chronological order starting at the time of our search, 23 February 2017. The most current reviews, i.e. the latest published, were selected to ensure inclusion of the most recent publications following the introduction of GRADE (Guyatt, Oxman et al. 2008) and its detailed guidance (Guyatt, Oxman et al. 2011, Guyatt, Oxman et al. 2011, Schünemann H 2013, Schunemann 2016). The nature of this study was explorative and no sample size was calculated.

Eligibility criteria and study selection

Titles and abstracts were screened for eligibility in their chronological order of publication. Full texts were retrieved and evaluated against our inclusion criteria by one investigator. A second investigator checked all eligible records, and a final list was agreed. Cochrane systematic reviews were considered eligible for inclusion if: (1) they were reviews of interventions; (2) they meta-analyzed at least two randomized controlled trials for a dichotomous outcome; 3) the dichotomous outcome was listed in the summary of findings (SoF) table.

We excluded diagnosis/prognosis reviews, studies on health service organisation, overviews of systematic reviews and network meta-analyses, and meta-analyses with only uninformative trials (i.e. with no events). The unit of analysis was one outcome for each review: either the primary outcome

or the first outcome meta-analyzed and listed in the SoF. We reasoned a priori that this outcome would most likely provide the basis for calculating sample size and orient clinical decision-making.

Data collection

General characteristics and reporting of GRADE in Cochrane reviews

Two investigators independently extracted data from all selected reviews. A third investigator resolved disagreements. We used a standardized ad hoc data collection form that we piloted on the first 5 Cochrane reviews and then revised according to problems identified. For each review we sought general information (e.g. author, contact author country, Cochrane review group name, new or updated review, type of intervention - pharmacological or non-pharmacological).

We evaluated what authors reported in the review methods section for assessment of imprecision. In particular, we wanted to determine whether the authors stated they had assessed imprecision and how (e.g. required or optimal information size, benefit/harm thresholds, width of 95% confidence intervals (CI), use of TSA) and in what way they planned to use imprecision assessment (e.g. evidence is downgraded when the required or optimal information size is not reached). We then recorded the grading of imprecision of the outcome selected from the SoF table and the reasons for downgrading. In some cases, we searched other sections of the full-text article for additional information.

Adherence to GRADE

We judged whether the review authors adhered to GRADE guidance for downgrading or non-downgrading evidence for imprecision. To determine whether the imprecision evaluation was appropriate (e.g. expected and observed downgrading of evidence), we consulted the instructions for downgrading for imprecision and re-assessed the optimal information size as suggested by the GRADE Handbook guidelines (Chapter 5.2.4.2 Imprecision in in systematic reviews in Schünemann H). For each review, we calculated the optimal information size in which we assumed an alpha of 0.05, a beta of 0.20, an *a priori* anticipated intervention effect – e.g., risk ratio reduction (RRR) or improvement – defined using the clinically relevant threshold reported by the review authors or, when not stated, the RRR suggested by GRADE authors, as a default threshold of 25% (Guyatt, Oxman et al. 2011), and the control event proportion of the meta-analysis. We used the normogram for events proposed by Guyatt et al. to determine the expected optimal information size (Guyatt, Oxman et al.

2011). Finally, we determined whether the reviewers incorporated and reported the imprecision assessment in their evaluation of evidence quality.

Agreement between GRADE assessment of imprecision and TSA

We evaluated agreement between downgrading of evidence as proposed in the original reviews and those performed by the authors of this article, and that resulting from TSA. For each review, we performed overarching TSAs: for each outcome, we re-analysed all trials that had been originally included in the meta-analysis. Data were synthesized using the same effect size with its 95% confidence interval (CI) and the same meta-analytic technique (i.e., random effects or fixed effect models), applying the reported statistical heterogeneity (I^2 value). Each trial was sequentially added in the TSA by publication year; this created a series of time points that formed the basis of the cumulative analysis. All TSAs were performed using TSA software (v 0.9.5.5 Beta) (2011). For each review, we calculated the diversity-adjusted required information size (DARIS) using, again, an alpha of 0.05, a beta of 0.20, a RRR as defined by the review authors or the default threshold of 25%, and the control event proportion. When a random effects model was chosen, between-trial variability was taken into account by adjusting the required information size with the diversity (D^2) originating from the meta-analysis of trials (Wetterslev, Thorlund et al. 2009). The Lan-DeMets trial sequential monitoring boundaries based on O'Brien-Fleming α -spending function was used (Wetterslev, Thorlund et al. 2008, Wetterslev, Jakobsen et al. 2017). The cumulative Z-curve (the series of Z-statistics after each consecutive trial) was calculated and plotted against these monitoring boundaries. In our primary analysis, we assumed for TSA minimal important or realistic anticipated intervention effects for all outcomes. If, following TSA methods (Jakobsen, Wetterslev et al. 2014), none of the sequential boundaries for benefit, harm or futility were crossed, imprecision was downgraded by two levels (Additional file 1 reports TSA judgement).

Data analysis

The data were summarized with descriptive statistics. Absolute and relative frequencies for categorical items and median and interquartile range (IQR) for continuous items were used. We reported adherence to GRADE through figures and agreement between the assessments involving the use of GRADE and TSA in contingency tables, with calculation of Cohen's kappa (Watson and Petrie 2010).

Agreement between GRADE imprecision assessment performed by the review authors and TSA was rated on the ordinal scale as: 0, not downgraded; 1, downgrade by 1 level; 2, downgrade by 2 levels. Moreover, we dichotomized the ordinal scale into ‘downgraded’ and ‘not downgraded’ for imprecision to evaluate the agreement irrespective of level of downgrading. Interpretation of agreement strength (k-values) was made according to the scale devised by Landis and Koch: <0.00 poor, 0–0.20 slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, 0.81-1.00 almost perfect (Landis and Koch 1977).

Univariable logistic regression was performed to investigate the impact of variables on downgrading for imprecision (dependent variable): Cochrane Group, the country of the contact author, type of intervention, number of patients included in the meta-analysis, heterogeneity among trials, and meta-analysis technique (random effects or fixed effect models). Each method, GRADE assessment of imprecision by the review authors, GRADE performed by the authors of this article, and TSA, was separately assessed and one variable evaluated at a time. The impact of these three variables on the agreement between the methods was then tested.

For hypothesis testing, a probability value of <.05 was considered statistically significant. All statistical tests were 2-sided. Stata statistical software was used for all statistical analyses (StataCorp.2003.).

Sensitivity analysis

TSAAs were replicated using RRRs of (A) 20% and (B) 30%, keeping all other assumptions the same. The concordance between the GRADE judgement on imprecision by the review authors and the TSA assessment was calculated irrespective of the levels of downgrading to determine whether the choice of the anticipated intervention effect affected the agreement.

Results

Characteristics of Cochrane reviews

We included 100 out of 216 potentially eligible Cochrane systematic reviews published in 2017 (issues 1 and 2) and 2016 (issues 12 and 11) of *The Cochrane Library* (Fig. 1), involving 36 (67.9%) out of 53 different Cochrane groups. Figure 1 shows the flow chart of reviews’ selection with reasons for exclusion. Additional file 2 reports included reviews and their main characteristics.

The three most active review groups were Pregnancy and Childbirth (n = 13 reviews), Neonatal (n = 11 reviews), and Heart (n = 8 reviews). The corresponding authors were based in the United Kingdom (n = 28 reviews), Australia (n = 15 reviews), or Canada (n=11 reviews). Sixty-one Cochrane reviews were updates of previous reviews. Table 1 presents the general characteristics of the reviews.

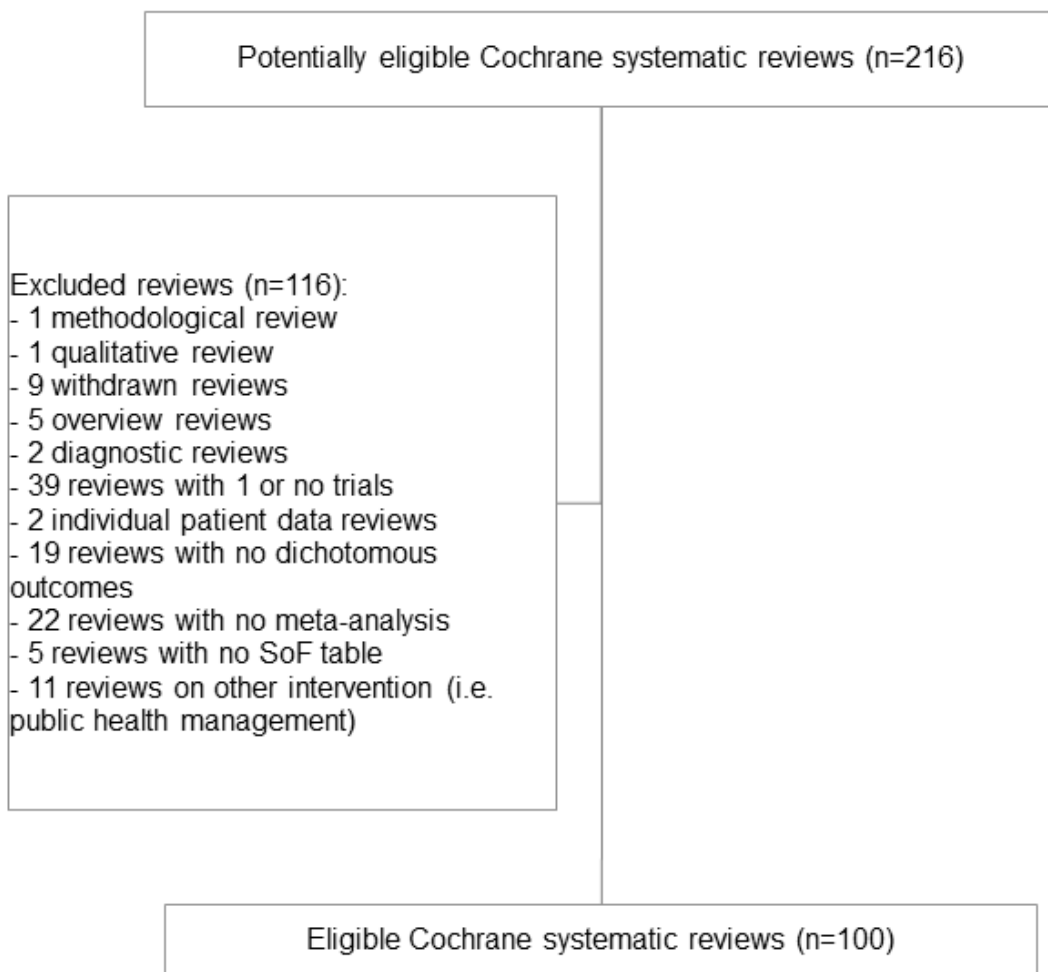


Figure 1. Flow chart of systematic review.

Characteristics	Value (no. of reviews)
No. of countries (no. total)	22
Top five countries	
- United Kingdom	28
- Australia	15
- Canada	11
- China	7
- Italy	6
No. of Cochrane groups (no. total)	36
Top five Cochrane groups	
- Cochrane Pregnancy and Childbirth Group	13
- Cochrane Neonatal Group	11
- Cochrane Heart Group	8
- Cochrane Airways Group	6
- Cochrane Gynecology and Fertility Group	6
Status of systematic reviews (no./out of 100)	
- Updated	61
- New	39
Type of intervention (no./out of 100)	
- Pharmacological	54
- Non-pharmacological	46

Table 1. General characteristics of the 100 Cochrane systematic reviews.

Overall, meta-analyses on the selected outcome were performed with a median of 5 randomized clinical trials (RCTs) (IQR, 2 to 9 RCTs; range, 2 to 30 RCTs). Most meta-analyses (81.0%) reported an effect measure expressed as risk ratio, 17.0% used the odds ratio, and only 2.0% reported the risk difference. Half of the SRs (51.0%) achieved statistically significant results according to the naïve 95% CI. The median heterogeneity of the meta-analyses was 12.0% (IQR, 0.0% to 49.0%; range, 0.0% to 98.0%).

Reporting of GRADE in Cochrane reviews

Nearly all (96.0%) of the reviews referred to GRADE in their methods section (Table 2). Of the four reviews that did not mention GRADE but performed it, two presented some information in the discussion. Very few (13.0%) of the reviews that graded the evidence reported the criteria they applied to assess imprecision. The most common imprecision components were width of 95% CI

(8.0%) and optimal information size referred to participants (4.0%). Ten reviews combined at least two criteria to assess imprecision. Only two reviews reported a comprehensive list of reasons behind imprecision judgment, thus allowing for full replication. Two other reviews planned and conducted a TSA. Neither the publisher Cochrane Group nor the country of the contact author influenced the reporting of imprecision assessment (univariable logistic regressions, respectively: Cochrane Group, $p= 0.716$; country of contact author, $p= 0.782$).

	Reported (no. of reviews out of 100)
Methods section	
Reviewers carried out GRADE assessment	96
Criteria considered for assessing imprecision?	13
- <i>Width of 95% confidence interval</i>	8
- <i>Optimal information size – no. of participants</i>	4
- <i>Optimal information size – no. of events</i>	1
- <i>Threshold for benefit or harm</i>	1
- <i>Trial Sequential Analysis</i>	2
Results section	
Optimal information size – no. of events	2
Optimal information size – no. of participants	8
Thresholds for benefit or harm	15
Trial Sequential Analysis	2

Table 2 Summary of approaches to imprecision and formal quantitative analyses related to imprecision in Cochrane systematic reviews. Values are numbers.

The quality of the meta-analyzed dichotomous outcomes was often graded as low (41.0%), with few outcomes reaching high quality (9.0%). Few reviews clearly stated on which criteria their assessment of imprecision were based. However, lack of details on how imprecision was assessed did not prevent the systematic reviewers to evaluate it, completing the SoF. Overall, almost half of the outcomes (48.0%) were downgraded for imprecision, with only six reviews downgrading imprecision by two levels. The most frequent reasons for downgrading due to imprecision were low number of events or small sample size (26.0%) and wide 95% CIs (25.0%). Six outcomes were downgraded due to imprecision, but no reason was reported in the SoF tables or full-text.

Adherence to GRADE Handbook instructions

When the authors of this article followed the GRADE Handbook instructions on how to replicate assessment and evaluate adherence, 64 outcomes were downgraded due to imprecision. Sixty-six did not meet the OIS for events. Overall, in 30.0% of reviews, judgment of outcomes differed between the review authors and the authors of this article who followed the GRADE Handbook (Figs. 2 and 3). Cohen’s kappa coefficient between the grading of imprecision as proposed by the original authors and as reanalyzed by us following the GRADE Handbook was 0.43 (95% CI 0.23 to 0.58), which expressed moderate strength of agreement.

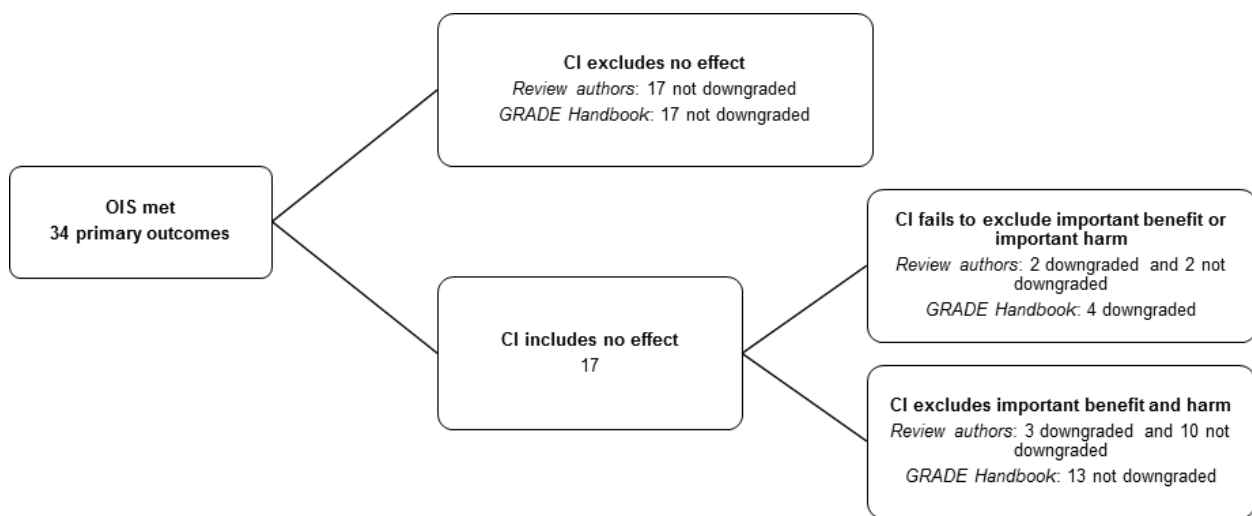


Figure 2. Primary outcomes that met the OIS – number of events: comparing GRADE assessment of imprecision carried out by review authors with GRADE carried out by the authors of this article following GRADE Handbook guidelines.

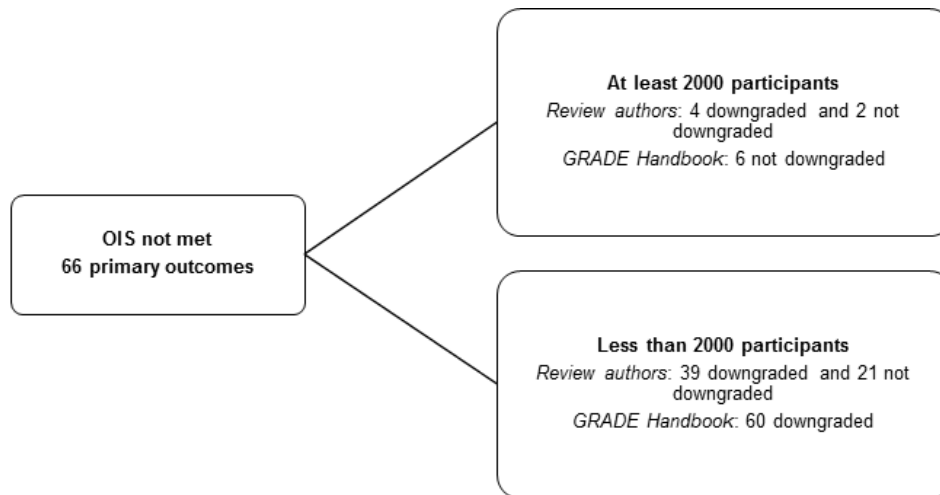


Figure 3. Primary outcomes that did not meet the OIS – number of events: comparing GRADE assessment of imprecision carried out by review authors with GRADE carried out by the authors of this article following GRADE Handbook guidelines.

Imprecision by TSA

The anticipated intervention effect was reported only by 12 reviews. For the other reviews, we adopted a 25% RRR to calculate the required information size. Overall, 69 outcomes were downgraded due to imprecision by applying TSA (downgrading by 2 levels since the anticipated intervention effect was assumed as being realistic). Indeed, five more outcomes were downgraded for imprecision by applying TSA as compared to using the GRADE Handbook instructions. The required information size was reached by 17 meta-analyses (17.0%). In the remaining 83.0%, the median number of participants needed to reach the required information size was 4 187 (IQR, 1 467 to 11 104 participants).

Agreement between GRADE by review authors and TSA

Weighted Cohen’s kappa coefficient showing agreement between GRADE performed by review authors and TSA was 0.20 (95% CI 0.11 to 0.30). The coefficient expressed slight agreement (Table 3).

	TSA			Total
	Not downgraded	Downgraded by 1 level	Downgraded by 2 levels	
GRADE by review authors				
<i>Not downgraded</i>	27	0	25	52
<i>Downgraded by 1 level</i>	3	0	39	42
<i>Downgraded by 2 levels</i>	1	0	5	6
<i>Total</i>	31	0	69	100

Table 3. Concordance in downgrading due to imprecision by 1 and 2 levels between GRADE carried out by review authors and TSA.

Considering only the outcomes downgraded or not downgraded due to imprecision, irrespective of levels, unweighted Cohen’s kappa coefficient was 0.43 (95% CI 0.21 to 0.57), expressing moderate strength of agreement (Table 4).

Agreement between GRADE by authors of this article and TSA

The imprecision evaluated by the authors of this article following the GRADE Handbook guidelines and by TSA was similar: unweighted Cohen’s kappa coefficient was 0.66 (95% CI 0.49 to 0.79), expressing substantial agreement (Table 4).

	TSA		
	Not downgraded	Downgraded	Total
GRADE by review authors			
<i>Not downgraded</i>	27	25	52
<i>Downgraded</i>	4	44	48
<i>Total</i>	31	69	100
GRADE by the authors of this article			
<i>Not downgraded</i>	26	10	36
<i>Downgraded</i>	5	59	64
<i>Total</i>	31	69	100

Table 4. Concordance in downgrading due to imprecision between GRADE carried out by the review authors and by the authors of this article and with TSA.

Results of logistic regression analyses

In the univariable logistic regression analysis, the type of the intervention (GRADE by review authors: $p=0.65$; TSA: $p=0.78$), the number of patients included in the meta-analysis (GRADE by the review authors: $p=0.08$; TSA: $p=0.07$), the heterogeneity (GRADE by review authors: $p=0.12$; TSA: $p=0.38$), and the meta-analysis technique (random effects or fixed effect models) (GRADE by review authors: $p=0.86$; TSA: $p=0.29$) were not associated with downgrading due to imprecision.

When GRADE assessment of imprecision carried out by the review authors was compared to TSA assessment and GRADE replicated by the authors of this article according to the Handbook guidelines, it seemed that the meta-analytic model (random or fixed effect) might influence the agreement in both cases ($p=0.04$ and $p=0.09$, respectively), whereas the number of patients might influence only agreement between GRADE by the authors of this article and TSA ($p=0.01$).

Sensitivity analysis

TSA with an anticipated intervention effect of 30% RRR revealed 60.0% of the SRs downgraded due to imprecision. When the anticipated intervention effect was lowered to 20% RRR, the percentage of SRs downgraded due to imprecision increased to 73.0%. Cohen's kappa coefficients, expressing downgrading or not due to imprecision, irrespectively of the downgrading levels (by 1 or 2), did not change much: coefficients 0.43 (95% CI, 0.22 to 0.58) and 0.43 (95% CI, 0.20 to 0.56), respectively.

Discussion

Given the implications of imprecision evaluation for recommending health care interventions as standard of care or cautionary noting that additional randomised clinical trials are warranted, it is expected that imprecision be transparently evaluated and reported. There remains ample room for improvement, however. Almost half of the outcomes were downgraded due to imprecision, but only about 1 in 10 reviews reported the criteria to downgrade imprecision. The width of 95% of confidence intervals and the number of study participants were the most common criteria to infer downgrading due to imprecision. One-third of the conclusions that did not downgrade the evidence for imprecision would have been contradicted based on GRADE assessment of imprecision following the GRADE Handbook or TSA if these methods had been applied. This was mainly because GRADE evaluation by the review authors was more lenient and also because the number of patients included in the meta-analyses was often insufficient to make any definitive conclusion.

The GRADE approach and TSA have different connotations that might influence the judgment process. While GRADE is defined as *"a semi quantitative approach that encompasses imprecision besides the other certainty domains"* and has intrinsic subjectivity (Schunemann 2016), TSA can be viewed as a purely quantitative and objective approach (Jakobsen, Wetterslev et al. 2014). Despite their differences, the agreement between the two approaches was substantial. Nonetheless, the two methods give different weight to the extent of imprecision. TSA tended toward more severe judgment, while GRADE seemed to be more easily applied by the review authors, resulting in overestimation of the certainty of evidence. When imprecision is rated as negligible (i.e. the true effect plausibly lies within the 95% CI), it is more likely that the effect estimates can be trusted and evidence quality rated highly. TSA will allow the reader to gauge the extent of confidence on imprecision of primary results of meta-analyses, though it might also be perceived as too radical by some. With a GRADE approach, reviewers' decisions on downgrading are more open to subjective decisions. To guide such decisions, further research is warranted on optimal information size and of different types of interventions under different circumstances.

Strengths and weaknesses of the study

This study has several limitations. We included only Cochrane reviews of health care interventions, a fairly homogeneous but partial sample of systematic reviews published in the medical literature. Reviews published in other medical journals might provide different results. However, since Cochrane strongly encourages use of the GRADE approach, including imprecision, it seems implausible that other journals would perform better. Our analyses are valid for pooled results and restricted to dichotomous outcomes, limiting the generalizability of our results. However, in the medical literature two third of primary outcomes reported by systematic reviews are dichotomous (Page, Shamseer et al. 2016). Besides, calculation and definition of a clinical threshold for continuous outcomes, i.e., the minimal important differences or minimal detectable changes, remain ambiguous and unclear (Copay, Subach et al. 2007, Armijo-Olivo, Warren et al. 2011). Furthermore, our logistic regression results may be under-powered to detect any significant factor, due to the limited number of studies included in the analysis.

While our analysis was protected against confounding by disease area and type of intervention because it is based on a sample of meta-analyses irrespective of interventions, it may still have been confounded by other meta-analysis characteristics. Given the wide diversity of the studies included, the quantitative results are suggestive and might change in the future when GRADE and Cochrane

propose methods and policies that strengthen imprecision assessment. Furthermore, imprecision assessment may vary between research areas (i.e. type of interventions), becoming speculative in fields where trials and reviews are too small to permit exploration of imprecision.

When we evaluated the conclusiveness of evidence by TSA in studies where the anticipated effect treatment effect was not reported, we could not directly assess the authors' assumptions on relevant intervention effects. We chose one arbitrary hypothesis when we recalculated the required information size based on a 25% relative risk reduction or improvement, which might be unrealistic for some outcomes. Nevertheless, the proportion of primary outcomes downgraded for imprecision did not change substantially when we applied different thresholds for benefits in our sensitivity analyses. Moreover, we only used a two level downgrading approach when our TSA did not show benefit, harm or futility. By e.g. employing Trial Sequential Analysis-adjusted confidence intervals, this simple approach could have been refined.

As well, our results are based on our subjective application of the GRADE approach following the GRADE Handbook, which, would then influence the assessments we made between the methods.

Despite these limitations, this project aspires to be a step towards optimized assessment and interpretation of the certainty of evidence regarding imprecision. Recently, evidence synthesis has evolved into a dynamic process. A new concept of systematic reviews, *living systematic reviews*, has been introduced (Elliott, Synnot et al. 2017). In this context, TSA has been suggested as a valid method to assess constantly updated evidence (Simmonds, Salanti et al. 2017) since it can offer constantly updating of optimal information size as new trials are added. This dynamicity could affect the imprecision domain, with assessment more closely related to what has been reached for a specific condition, outcome, and intervention at a certain time point.

Implications for systematic reviewers

As part of their mandate to identify potentially effective interventions, systematic reviewers should include precision as a key dimension to evaluate, particularly in situations where uncertainty about the ratio between benefits and harms is high and where new trial data may influence the summary judgment of the review (Riva, Puljak et al. 2017). More detailed guidance from PRISMA and PRISMA-P could facilitate the analysis and reporting of imprecision (Moher, Liberati et al. 2009, Moher, Shamseer et al. 2015). There is room for better standardisation of approaches and inclusion of quantitative methods, such as TSA, to formally evaluate imprecision.

Implications for clinicians

Imprecision assessment seems to be based on GRADEing criteria that vary considerably in meaning, value, or boundaries depending on context or conditions. Incomplete or vague imprecision assessment supports the natural tendency to simplify a review's findings about an intervention as positive or negative and to over-rely on P-values (Pocock and Stone 2016). Instead, clinicians using evidence to orient their practice should interpret imprecision as a dynamic and often uncertain dimension that requires thorough examination of all of the evidence, including size of the trials, number of participants with outcomes and heterogeneity (diversity) across trials. TSA offers the means to model heterogeneity and estimate optimal information size based on different heterogeneity thresholds.

Future research

Further research is needed to determine whether accurate assessment of imprecision might change the clinicians' perception about the definitive effectiveness of interventions. It would be beneficial to develop an international database of prospectively updated TSAs for all health interventions, where people can easily consult the progress of research. This might also help to diminish "the butterfly behavior of researchers" from moving onto the next research question before the previous quest has been fully exploited (Liberati 2004).

Conclusions

A significant lack of reporting of and adherence to GRADE was observed in Cochrane systematic reviews. Stricter adherence to GRADE Handbook guidelines and/or adoption of TSA would have led to more frequent downgrading of the quality of evidence. Our findings reiterate the need for a more reliable application of GRADE in Cochrane and non-Cochrane reviews for the assessment of imprecision. The reasons for downgrading should be defined following a more structured reporting system. Initiatives to improve adherence to GRADE indications are warranted to ensure high-quality systematic reviews. Cochrane groups, peer reviewers, and editors need to be more supportive during the GRADE assessment process as it is still difficult to apply and requires a structured reporting policy to ensure transparency and intelligibility in the process.

General conclusion

In an era in which clinicians, researcher and stakeholders need to be constantly updated with the evidence, it has been progressively important to reliably provide the extent of the confidence (the uncertainty) of the benefit or harm of the effect of an intervention. The effect estimate from the meta-analysis and its precision (or confidence interval) is one of the deciding factors in grading the existing evidence. Through this section, we showed how TSA and GRADE approach are integrated in the evidence synthesis in order to consider precision and trustworthiness of meta-analysis findings.

TSA has been recently introduced as statistical method to assist a modern version of the systematic review: the living systematic review, which might be particularly useful as basis for the creation of ‘living guidelines’ (Elliott, Synnot et al. 2017, Simmonds, Salanti et al. 2017). Indeed, TSA might help in controlling the statistical precision of the summary effects when continually or frequently update of meta-analyses for this living study design. When a living systematic review is being used to make decisions, then it is preferable to consider approaches to avoid inadvertent type I and II errors.

The GRADE approach has been more and more essential in the developing of guidelines and recommendations thanks to its comprehensive method to balance the certainty and magnitude of the evidence with benefits/harms and feasibility of interventions.

SECTION 3

Is it possible to simultaneously compare multiple interventions in a systematic review?

Background

A key limitation of a meta-analysis is that it can compare only two treatments at a time, yielding only partially information that clinicians, patients and policy-makers need. Since usually more than two treatment options are available in a real clinical setting for certain conditions, a meta-analysis can offer just a part of the whole picture and might be not supportive for an optimal clinical decision making.

In the last decade, a new meta-analytic technique called network meta-analysis (NMA) has been developed to assess the effectiveness of several interventions and offer the synthesis of the evidence across a network of randomized trials. The interest for this technique has been increased across researchers and over the time. Indeed, the PRISMA Statement for Reporting of Systematic Reviews published an extension incorporating the guideline for reporting Network Meta-analyses of Health Care Interventions (Hutton, Salanti et al. 2015).

In order to acquire the basis of this methodology, I attended a course at the Winter epidemiology school organized by the University of Bern in January 2017. Lectures were given by Prof. Georgia Salanti and Julian Higgins, the main experts of the network meta-analysis. Thereafter, a protocol has been developed to apply this methodology in low back pain field: a multi interventions comparison among rehabilitative treatment to reduce pain and disability in patients affected by acute low back pain will be performed.

Following, I have briefly reported the basis of NMA methodology.

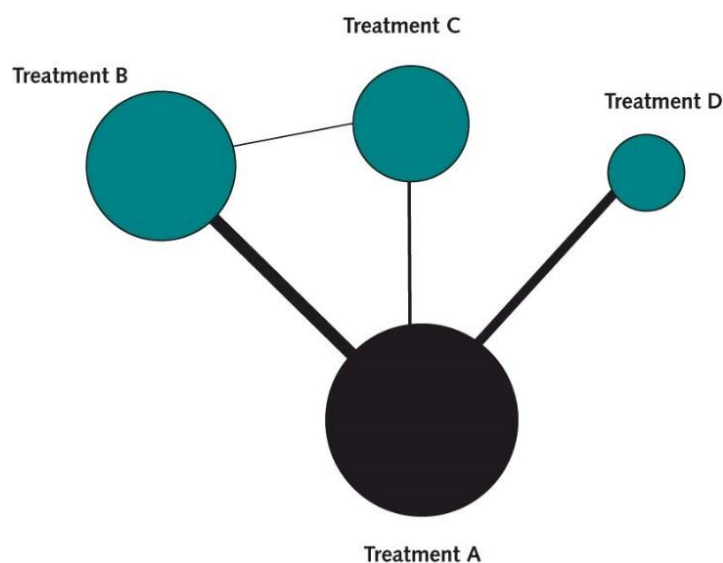
What is a network meta-analysis?

A network meta-analysis, in the context of a systematic review, is a meta-analysis in which multiple treatments (i.e. more than two) are compared using both direct comparisons and indirect comparisons of interventions. The direct evidence comes from RCTs, i.e. A vs C and B vs C, while

the indirect evidence comes from studies comparing treatment of interests with a common comparator, i.e. A vs B. The combination of the direct and indirect sources of the evidence builds the *network meta-analysis*, even called *mixed-treatment comparison* .

The NMA enables the rank of the interventions and the estimation of the probability that each intervention is the best for each outcome. The ranking presentation or the probabilities are very useful for clinical purpose since they are straightforward to understand and easy to read by clinicians who usually want to know the preferential order of treatments that could be prescribed to an average patient.

A systematic review with a network meta-analysis is generally more complex to conduct and mostly to interpret it. To present the resulting evidence, reviews with a NMA commonly include a graph of the network to summarize the numbers of studies that compared the different treatments and the numbers of patients who have been studied for each treatment:



This network graph shows:

- nodes: points representing the competing interventions;
- edges: adjoining lines between the nodes that show which interventions have been compared among the included studies.

The sizes of the nodes and the thicknesses of the edges typically represent the amounts of the existing evidence for specific nodes and comparisons. A network meta-analysis can include an unlimited number of treatments, trials and patients. However, in order to conduct a standard network

meta-analysis, the treatments should form a connected network, i.e. a path from each treatment to every other treatment in the network should exist.

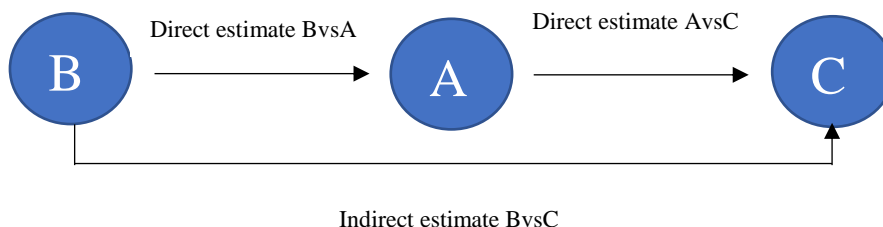
The validity of a network meta-analysis is dependent upon the assumption of transitivity and consistency, topics which are discussed below.

Indirect comparison: what it is?

When there is information on three or more competing interventions in a systematic review, it is possible to perform an indirect comparison. Suppose that there are “head-to-head” trials comparing direct evidence of interventions A versus B in one trial and in the other interventions A versus C. No trial, however, compares the interventions B versus C. We can extract an indirect comparison of the relative effect of B versus C by combining the summary estimates of AvsB and AvsC. This is possible considering each comparison as a vector that identifies the direction of the intervention effect. Therefore, the indirect comparison of B versus C may result as following:

$$\text{Indirect MD(BvsC)} = \text{direct MD(BvsA)} + \text{direct MD(AvsC)}$$

And graphically:



Using the relative intervention effects from each group of trial preserves the within trial randomization. Basically, the direct comparison is more precise and strong since it provides more information per randomized participant than the indirect comparison.

Transitivity

Clinical and methodological differences, “effect modifiers” (i.e. age, sex) are inevitable across studies in a systematic review. An important assumption which should be underlined in order to

allow an indirect comparison is that the common comparator A allows a transitive relationship between AB and AC effects. In other words, we can compare C and B via A. In order to do that, the comparator A should be similar in studies AB and AC. Transitivity requires similarity which is that the sets of studies used to obtain the indirect comparison are sufficiently similar in characteristics that moderate the intervention effect. Therefore, to undertake an indirect comparison is necessary to analyse and measure whether such differences are sufficiently large to cause intransitivity. In practice, the transitivity can be assessed by comparing the distribution of the effect modifiers across the different comparisons: if there is imbalance, the plausibility of a transitive relationship and so the validity of an indirect comparison would be threaten. The transitivity requires that all the competing intervention are “*jointly randomizable*” as if all the interventions can be compared simultaneously in a single multi-arm trial.

Mixed estimates

When both direct and indirect intervention effects are available, these can be combined into a single summary estimate of the intervention effect called *mixed estimate*. A mixed estimate is an inverse variance weighted average of the direct and indirect evidence.

Consistency

The network meta-analysis is subject to another assumption other than transitivity, the consistency. While the transitivity refers to clinical and methodological discrepancies across the different comparison, the consistency is the statistical manifestation of the transitivity. In other words, the consistency is the amount of agreement between direct and indirect intervention effects. The consistency is mathematically shown as following:

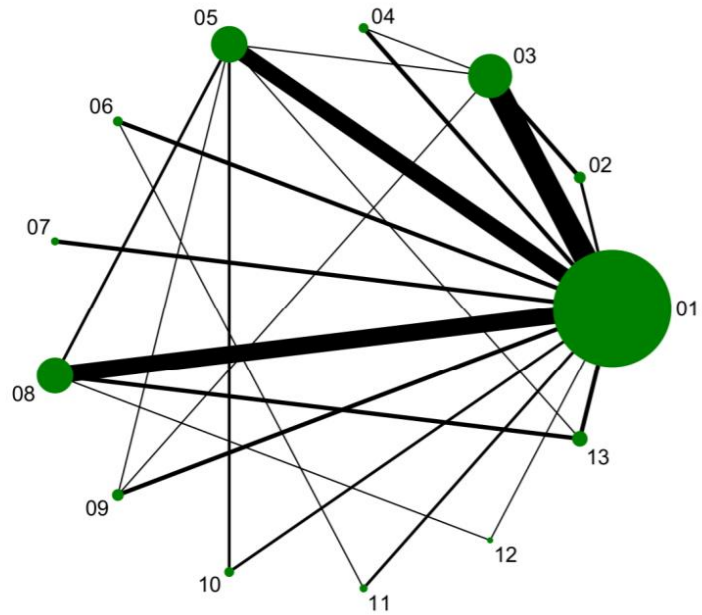
$$\text{“true” indirect MD}(B\text{vs}C) = \text{“true” direct MD}(A\text{vs}C) - \text{“true” direct MD}(A\text{vs}B)$$

We can consider the transitivity and the consistency as the clinical/methodological and statistical heterogeneity respectively in standard meta-analysis.

The graphical representation: network diagrams

The graphical representation of a network meta-analysis is called network diagram. Following there is an example extracted from a systematic review on the efficacy of exercise for lower limb osteoarthritis (Uthman, van der Windt et al. 2014).

1. No intervention control
2. Flexibility (F)
3. Strenghtening (S)
4. Aerobic (A)
5. Flexibility + strenghtening
6. Flexibility + aerobic
7. Strenghtening + Aerobic
8. Combined (F+S+A)
9. Acquatic strenghtening
10. Acquatic flexibilty + strenghtening
11. Acquatic combined (F+S+A)



In

In the figure above, the nodes are representing the competing interventions while the lines are showing the available direct evidence between pairs of interventions. Closed loops reveal the presence of a multi arm study.

It is important to underline that the assumption of transitivity and consistency should be hold in each single loop otherwise the validity of the results is vulnerable.

Ranking of the intervention effects

The network meta-analysis has the advantage to present the ranking of the competing interventions, using the ranking probabilities, which are the probabilities that an intervention is at a specific rank when compared to other interventions in the network for the selected outcome. These probabilities can be represented in several way, using for examples table or rankograms. For instance, it is possible to represent the rank through the area under the cumulative rankogram (SUCRA) which is a value between 0 and 1 and can be re-expressed as percentage. The larger the SUCRA, the higher the treatment in the hierarchy according to the outcome.

Case example of NMA: effectiveness of treatments for LBP interventions

Published as: Effectiveness of treatments for acute and sub-acute mechanical non-specific low back pain: protocol for a systematic review and network meta-analysis. Gianola S, Castellini G, Andreano A, Corbetta D, Frigerio P; Pecoraro V, Redaelli V, Tettamanti A, Turolla A, Moja L, Valsecchi MG

Systematic Reviews; 8; 196. July 2019.

The development and conduction of a network meta-analysis is challenging and the load of randomized controlled trials to analyse is remarkable therefore, a collaboration with other health professionals is fundamental to successfully develop it. For this reason, the project was in collaboration with other physiotherapists and researchers: this was a unique opportunity not only to develop a complex study design but even to develop ability to do networking and to share knowledge and have intellectual exchanges.

Methods

A systematic review protocol has been developed and registered on PROSPERO database (CRD42018102527, available at: <http://www.crd.york.ac.uk/>). This review protocol was prepared using the Preferred Reporting Items for Systematic Review and Meta-Analyses Protocol (PRISMA-P) guidelines and their recommendations (Moher, Shamseer et al. 2015, Moher, Stewart et al. 2016). Sections specific to NMA have been considered according to Chaimani et al.(Chaimani, Caldwell et al. 2017). We used the PRISMA-NMA extension statement to structure the contents of the actual systematic review and network meta-analysis (Hutton, Salanti et al. 2015).

Eligibility criteria

Types of studies

We only included randomized controlled trials (RCTs). RCTs were considered if authors explicitly stated that it is randomised (Higgins, Altman et al. 2011). Quasi-randomised trials and cross-over trials were excluded.

Participants

We included trials that involve participants older than 18 years, both males and females, experiencing pain for up to 12 weeks of non-specific LBP. We classified the population based on pain duration: acute (less than six weeks) or subacute (six to 12 weeks) (van Tulder, Becker et al. 2006). Accordingly, we selected trials for pain duration, regardless of the population definition reported for a study (e.g., chronic patients with pain for less than 12 weeks). When the duration of pain exceeds, for few weeks, the standard definition of subacute pain (i.e., recruitment from 8 to 16 weeks), we contacted the authors to obtain the data for our population of interest only, otherwise the study would be excluded. According to the definition of aspecific LBP, we excluded studies focusing on specific pathological entities (e.g., spondylolisthesis) and subgroups of patients (e.g., pregnant women). There was no restriction on the severity or stage of the symptoms. Studies focusing on both neck and back pain in which the two subgroups of patients could not be identified, or patients presenting with both conditions, were excluded.

Interventions

We considered all conservative rehabilitation or pharmacological treatments provided by health professionals, such as general medical practitioners or physiotherapists, aimed at relieving pain and/or reducing physical disability. We considered any modality (e.g. physical, pharmacological), treatment extent, frequency or intensity. We excluded RCTs or arms of RCTs including non-conservative treatments (e.g., surgical approaches), herbal medicine, homeopathy and all alternative treatments. We included acupuncture and dry needling. We set the following classification of interventions for potential nodes:

1. exercise (e.g., cognitive, back school)
2. manual therapy (e.g., spinal manipulation, mobilization, trigger point/myofascial therapy)
3. acupuncture (e.g., dry needling and acupuncture)
4. any physical therapy (e.g., low-laser therapy, diathermy, Transcutaneous Electrical Nerve Stimulation, ultrasound therapy, heat wrap)
5. taping (such as kinesiotaping)
6. usual care defined as treatment suggested by general medicine (minimal intervention: advice to stay active or to take drugs as needed; education)
7. paracetamol
8. non-Steroidal Anti-Inflammatory Drugs (NSAIDs), including COX-2 inhibitors

9. muscle relaxant drugs
10. opioid drugs
11. steroids
12. antidepressant drugs
13. inert treatment (e.g., placebo drug, sham therapy)
14. no treatment (no treatment, waiting list control)

Outcomes and study time-points

Primary outcomes were pain intensity (e.g., measured by numeric rating scale, visual analogue scale, McGill Pain Questionnaire or, box scale, other validated quantitative measures) and back-specific functional status (e.g., measured by the Oswestry disability questionnaire, Roland-Morris disability scale or other validated quantitative measures). If a trial reported more than one measure of pain intensity in different conditions (e.g., “night” or “at rest” or “at movement”), we would select “pain at rest” as a measure of generic pain. The secondary outcome was any adverse event.

All time points were abstracted. However, in the analyses we planned to summarize the immediate-term (closest to 1 week), short-term (closest to 1 month assessment), intermediate (closest to 3–6 months) and long-term effects (closest to 12 months).

Information sources

We searched the following electronic databases since the inception date up to November 30 2017: MEDLINE (PubMed), CENTRAL, EMBASE (Elsevier, EMBASE.com) using the appropriate Thesaurus and free-text terms. We contacted investigators and relevant trial authors, seeking information on unpublished data, if necessary.

We checked the reference lists of all the studies identified, and we examined the references of any systematic review or meta-analysis identified during the search process.

No restriction on language or publication period were applied. Non-English studies for which a translation cannot be obtained were classed as potentially eligible but were not be considered in the full review. A full electronic search strategy for Pubmed/Medline is presented in **Appendix 1**.

Study selection

Two of the authors of the present protocol independently screened the abstracts of all the publications obtained by the search strategy. These authors then independently assessed the full text of the potentially relevant studies for inclusion. We discarded all studies that did not fulfil the above inclusion criteria. We then obtained the full text of the remaining articles. We resolved disagreements through discussion and consult a third author if disagreements persist. Covidence software (COEVIDENCE) was used to manage the study selection phase.

Data extraction

We used a specifically designed and piloted data collection form using an Excel sheet (Microsoft Inc.). Two authors independently extracted characteristics and outcome data from the included studies. Disagreements were resolved through discussion or with assistance from a third author if necessary.

From each study included we extracted: name of the first author, year of publication, setting, number of centers, population definition (acute/subacute); number, gender and age of participants, dropouts; the interventions compared with their primary and secondary outcomes, time point follow-up, and duration of whole treatment.

All relevant arm-level data were extracted. We considered post-treatment assessments. When these were lacking, the post-treatment data were extrapolated by the difference between the baseline and mean change values and SDs were imputed using the average of the available SD for the same instrument. If any data from the baseline and post treatment were not available, the mean change and SD values were adopted as a last option (Higgins, Deeks et al. 2011).

We assumed that any patient meeting the inclusion criteria is, in principle, equally likely to be randomized to any of the eligible low back pain interventions.

Geometry of the network

We explicitly described the process leading to node grouping (James, Yavchitz et al. 2018, James, Yavchitz et al. 2018). The network of treatments was judged based on the characteristics of the available studies, presented and evaluated graphically. We evaluated: if the network was disconnected, if there was a sufficient number of comparisons in the network with available direct

data; if there was a high number of comparisons based on a single study; if any key treatment was missing. Next, the feasibility of a network meta-analysis was assessed.

All RCTs reporting only two arm comparisons between the same kind of intervention (e.g., exercise versus exercise) were excluded, whereas if they presented at least one third arm comparator, they would be included (e.g., exercise versus NSAIDs). We included both multi-arm trials comparing three or more interventions, and those comparing different dosages or regimens of an intervention to a different one. Intervention arms of different dosages and regimens of the same intervention across the RCTs were merged together for the global analysis of all outcomes. We did not consider all the comparisons in which an intervention presents multiple co-interventions for the experimental group (e.g., mixed treatment: laser therapy plus manipulation plus exercise versus waiting list controls) or for the control group (e.g., usual care: education, some physical exercise plus drugs taken as needed) to avoid inconsistencies across trials.

Risk of bias within individual studies

Two review authors independently assessed the risk of bias in the included studies. Disagreements was resolved through discussion or arbitration with a third review author when consensus cannot be reached. We assessed the risk of bias for each included study using the 'Risk of bias' (RoB) assessment tool recommended by The Cochrane Collaboration (Higgins, Altman et al. 2011). Specifically, we evaluated the following criteria: random sequence generation, allocation concealment, blinding of participants, providers and outcome assessment, incomplete outcome data, and selective outcome reporting. Each item was scored as 'high', 'low', or 'unclear' RoB if no sufficient information was reported. To summarize the overall RoB for a study, allocation concealment, blinding of outcome assessment, and incomplete outcome data were carefully considered in order to classify each study as: 'low risk of bias' when all three criteria are met; 'high risk of bias' when at least one criterion is unmet; and 'moderate risk of bias' in the remaining cases. Allocation concealment, blinding of outcome assessment, and incomplete outcome were chosen since expected to be of importance across the pain intensity outcomes in altering results..

Measures of treatment effect

Relative treatment effects

Through pairwise meta-analyses, we estimated the primary outcomes as continuous outcomes, using the mean difference (MD) or standardized mean difference (SMD) when different outcome measurements had been reported for each trial. The uncertainty of all estimates was expressed with its 95% confidence interval (CI).

Data summary

Methods for direct treatment comparisons

We performed conventional pairwise meta-analyses for each primary outcome using a random-effects model for each treatment comparison with at least two studies (DerSimonian and Laird 1986) using Stata software v. 15 and the command metan (Stata-IC 2017).

Methods for multiple comparisons

We performed the network meta-analyses within a frequentist setting, assuming equal heterogeneity across all treatment comparisons, and accounting for correlations induced by multi-arm studies (Salanti 2012, Miladinovic, Hozo et al. 2014). We used a multivariate normal model with random-effects (Higgins, Jackson et al. 2012, White, Barrett et al. 2012). We first fitted a design by treatment interaction model to assess the presence of inconsistency (global χ^2 test). If the null hypothesis of all inconsistency parameters being equal zero was not rejected, we fitted a consistency model. If a global significant inconsistency was found, we tried to interpret the significant inconsistency parameters, split nodes to possibly remove the problem, and try to model the inconsistency using meta-regression.

Relative treatment ranking

We estimated all ranking probabilities and cumulative ranking probabilities for each treatment and outcome. We then calculated the median rank with their 95% credible intervals, to assess the robustness of the finding. To determine a treatment hierarchy with a single number, we will calculate the surface under the cumulative ranking curve (SUCRA) and express it as a percentage (Salanti, Ades et al. 2011).

We performed network meta-analyses in Stata 15 (Stata-IC 2017) using the ‘network’ command and the ‘mvmeta’ command ((MTM). , White 2011, Higgins, Jackson et al. 2012, Chaimani, Higgins et al. 2013).

Assessment of statistical heterogeneity

In the standard pairwise comparisons, we assessed the statistical heterogeneity within each pairwise comparison using the I^2 statistic, where an I^2 value of 25% to 49% indicates a low degree of heterogeneity, 50% to 75% a moderate degree of heterogeneity, and more than 75% indicates a high degree of heterogeneity (Higgins, Thompson et al. 2003).

In the network meta-analyses, we assumed that the standard heterogeneity is constant across the different treatment comparisons. We estimated it including a random effect in the multivariate normal model, assuming a multivariate normal distribution with mean 0 and a variance-covariance matrix with diagonal elements τ^2 and off-diagonal elements equal to $\tau^2/2$, and discuss the magnitude of the estimated variance parameter.

Assessment of transitivity and statistical consistency in network meta-analyses

We assessed the assumption of transitivity (or similarity) by comparing the distribution of the potential effect modifiers across the various pairwise comparisons. If there were no multi-arm trials, we evaluated the inconsistency assumption in each closed loop of the network separately as the difference between direct and indirect estimates for a specific comparison (inconsistency factor). The magnitude of the inconsistency factors and their 95% CIs will be used to make an inference about the presence of inconsistency in each loop.

If multi-arm trials were present, as it is problematical to identify loop inconsistencies, we used the node-splitting approach to evaluate existing differences between direct and indirect estimates for each node (Dias, Welton et al. 2010).

To check the assumption of consistency in the entire network, we used the design-by-treatment model as described by Higgins (Higgins, Jackson et al. 2012). This method accounts both for loop and design (i.e. different sets of treatments compared in a trial) inconsistencies in multi-arm trials. Using this approach, we made an inference about the presence of inconsistency from any source in the entire network based on an χ^2 test. Inconsistency and heterogeneity are interwoven: to distinguish between

these two sources of variability, we will employ the I^2 statistic for inconsistency, as it measures the percentage of variability that cannot be attributed to random error or heterogeneity (within comparison variability).

Preliminary Results: back pain and specific functional status at 1 week of follow up

Study selection

After removing duplicates, the whole search strategy retrieved 6964 records. Reviewing the titles and abstracts, we discarded 6419 irrelevant citations. We examined the full text of the remaining 549 records of which 517 did not meet the inclusion criteria. Within this group, 249 records included a different population (e.g, chronic pain), 145 different interventions (e.g., mixed treatments) or comparisons (e.g., exercise versus exercise), 20 different outcomes (e.g., cost-effectiveness related to pain), 66 a study design different from RCT, 14 were further duplicates and 23 studies not in English or Italian that are still in awaiting assessment. In total, 22 authors were contacted, and 2 of the 5 who responded provided data usable for analysis. Finally, 36 studies were included (citations in **Appendix 2**). For a further description of screening process, see the study flow diagram in **Figure 1**.

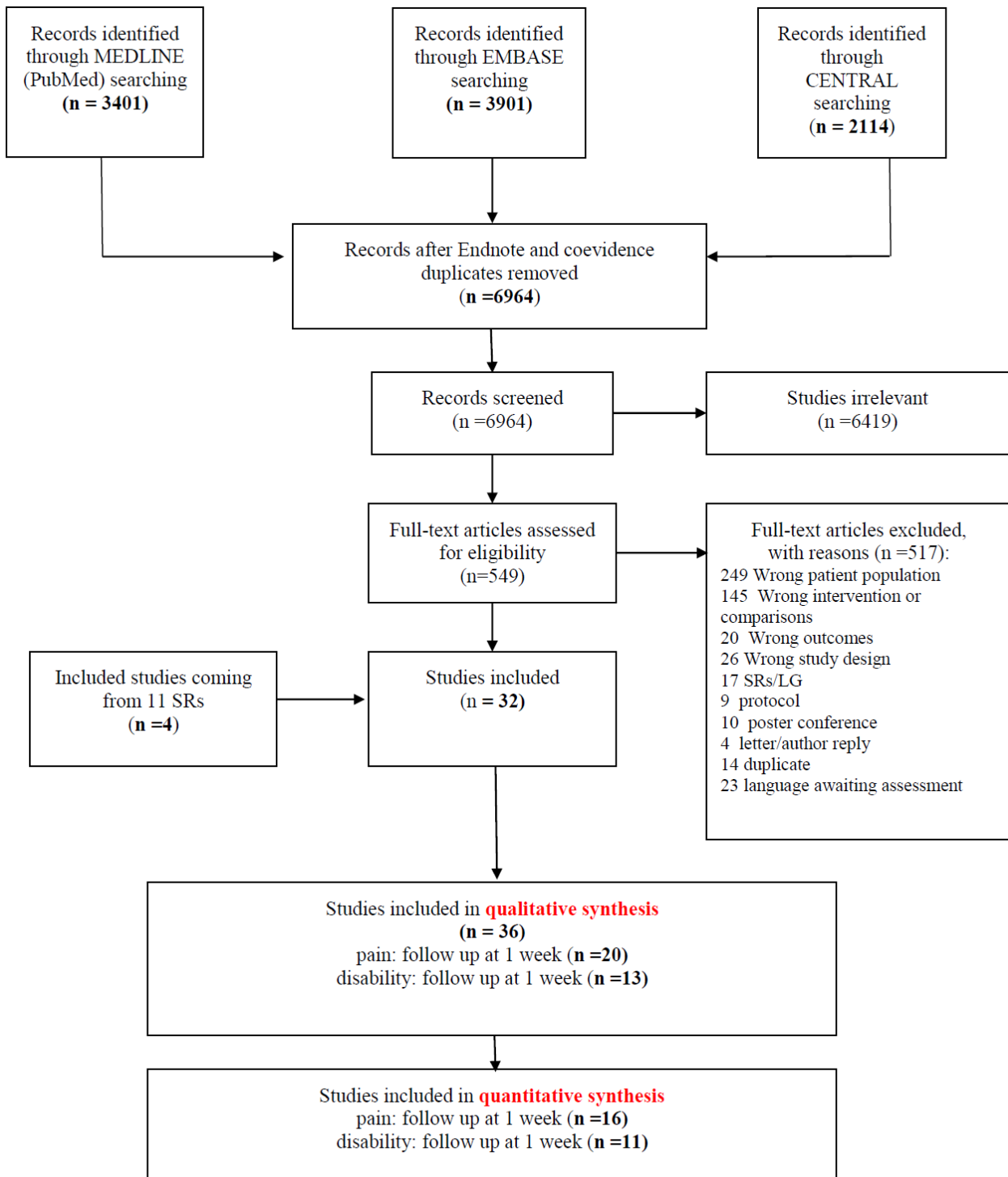


Figure 1. Flow Chart of study selection.

General Characteristics

A total of 7787 participants were included in 36 trials. The sample size varied between 92 and 227 participants (InterQuartile range -IQR) with a median of 126 participants. The majority of studies involved acute patients (n=19 trials). The median year of publication of the trials was 2004 (IQR 2000-2010). **Table 1** summarizes these general characteristics. Average age ranged from 38 to 44 and a median percentage male ranged from 41% to 53%.

General characteristics are summarized in **Appendix 3** including all studies and participants characteristics. No important concerns were raised regarding the violation of the transitivity assumption when the following variables were considered as potential effect modifiers: stage of LBP, length of treatment, age, sex. Similarity of trials characteristics was guaranteed in terms of clinical and methodological features.

We presented the NMA results for pain and disability compared with inert treatment (control treatment), and the SUCRA values at immediate term, 1 week of follow up. Therefore, we showed quantitative analysis for 27 RCTs included.

Study Characteristic	No. (%) of RCTs (N = 36)
Year of publication	
1971-1980	1 (2.8)
1981-1990	3 (8.3)
1991-2000	6 (16.7)
2001-2010	18 (50.0)
2011-2018	8 (22.2)
Interventions*	
Exercise	18 (50.0)
Manual therapy	7 (19.4)
Acupuncture	2 (5.6)
Usual care	17 (47.2)
Inert treatment	12 (33.3)
Heat wrap	3 (8.3)
Paracetamol	4 (11.1)
NSAIDs	8 (22.2)
Muscle relaxant drugs	9 (25.0)
Opioids	2 (5.6)
Length of treatment	
< 7 days	15 (41.7)
> 7 days	16 (44.4)
Not reported	5 (13.9)
Stage of LBP	
Acute LBP	19 (52.8)
Subacute LBP	9 (25.0)
Acute and Subacute LBP	8 (22.2)
Setting of center	
Multi-center	19 (52.8)
Single-center	17 (47.2)

*the tot number of interventions is higher due to multi-arms trials

Table 1. General characteristics

Risk of bias assessment

Figure 2 and 3 summarized the RoB assessments. Regarding the overall RoB across studies (n=36), only 8 trials were at low RoB (22%). We categorized 56% of the studies as at unclear RoB (n=20) and 22% at high RoB (n=8).

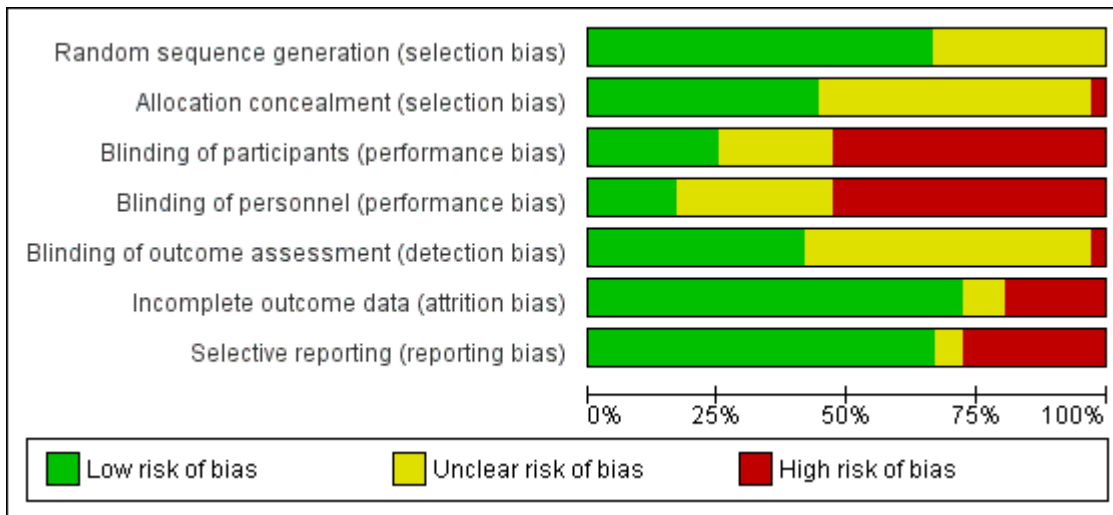


Figure 2. Risk of bias graph.

	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Blinding of participants (performance bias)	Blinding of personnel (performance bias)	Blinding of outcome assessment (detection bias)	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)
Berry 1988	?	?	?	?	?	+	+
Casale 1988	?	?	+	?	?	+	?
Dreiser 2003	+	+	+	+	?	+	+
Eken 2014	+	+	+	+	?	+	+
Hagen 2010	+	+	?	?	?	?	-
Hasagawa 2014	+	?	+	-	+	+	+
Hindle 1972	+	-	?	?	?	+	-
Hsiesh 2002	+	?	-	-	+	+	+
Jallama 2005	+	?	-	-	?	+	?
Ketenci 2005	?	?	+	?	?	+	+
Li 2008	+	+	+	+	?	+	+
Lindstrom 1995	?	?	-	?	?	+	-
Linton 2000	+	?	-	-	?	+	+
Little 2001	+	+	-	+	?	-	-
Machado 2010	+	+	-	-	+	+	+
Malmivaara 1995	+	+	-	-	+	+	+
Mayer 2005	+	?	-	-	?	+	-
Moffett 1999	+	+	?	-	?	+	+
Nadler 2002	?	?	-	-	+	+	+
Nadler 2003	?	?	-	-	?	-	+
Postacchini 1988	?	?	?	?	?	?	-
Pozo-Cruz 2012	?	?	-	-	+	+	-
Rabin 2014	+	+	-	-	+	-	+
Ralph 2008	?	?	?	?	?	+	-
Schneider 2015	+	+	-	-	+	+	-
Seferlis 1998	?	?	-	-	?	-	+
Serfer 2010	+	?	+	+	-	+	+
Shin 2013	+	+	-	-	+	+	+
Staal 2004	+	+	-	-	+	+	+
Storheim 2003	+	+	-	-	+	-	+
Szpalski 1994	?	?	?	?	?	+	+
Takamoto 2015	+	?	-	-	+	-	-
Tuzun 2003	+	+	+	?	+	+	+
Wand 2004	+	+	-	-	+	-	+
Whitfill 2010	?	?	?	?	?	?	+
Williams 2014	+	+	+	+	+	+	+

Figure 3. Risk of bias summary.

Effect of interventions on pain and disability at short term FU (1 week)

Pain was investigated by 20 studies at 1 week of FU: 16 out of 20 studies (N = 3760) provided data on 16 direct comparisons between 10 different treatment nodes (**Figure 4a**). **Figure 5a** shows the network according to level of bias in included trials.

Disability was investigated by 13 studies at 1 week of FU. The NMA on disability at 1 week (11 out of 13 studies [N =4266] provided data on 13 direct comparisons between 9 different treatment nodes, **Figure 4b**. **Figure 5b** shows the network according to level of bias in included trials.

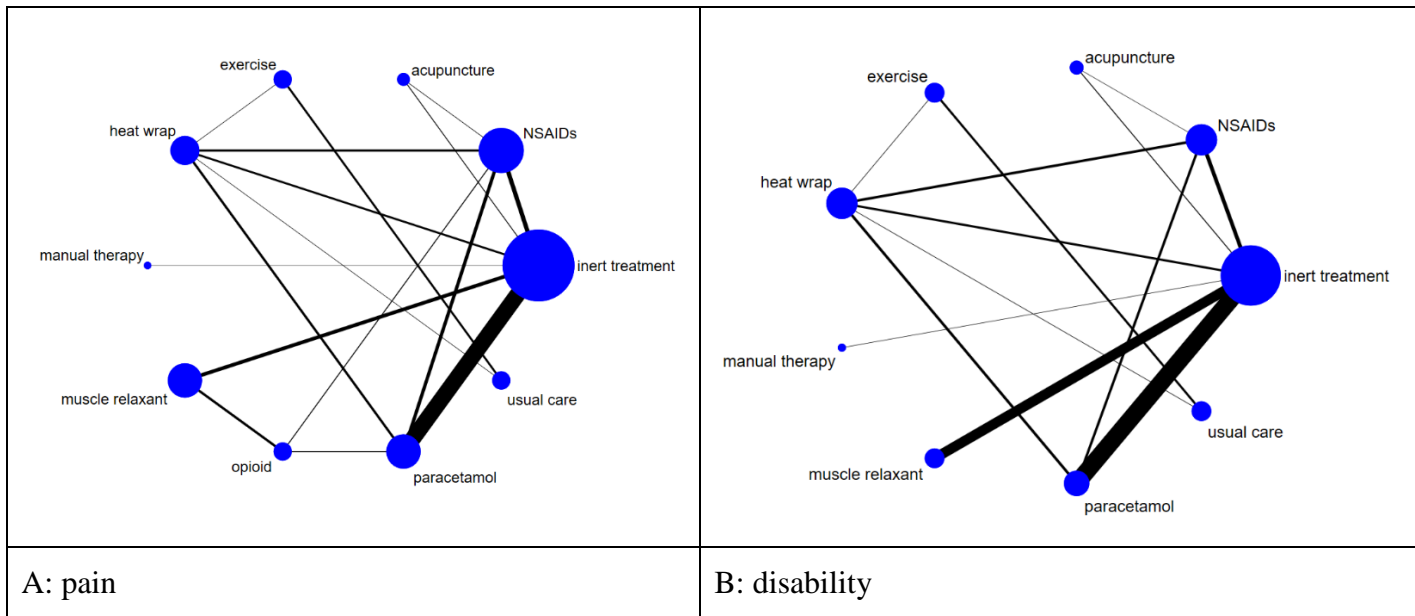


Figure 4. Network plot for pain and disability outcomes at short term of 1 week FU.

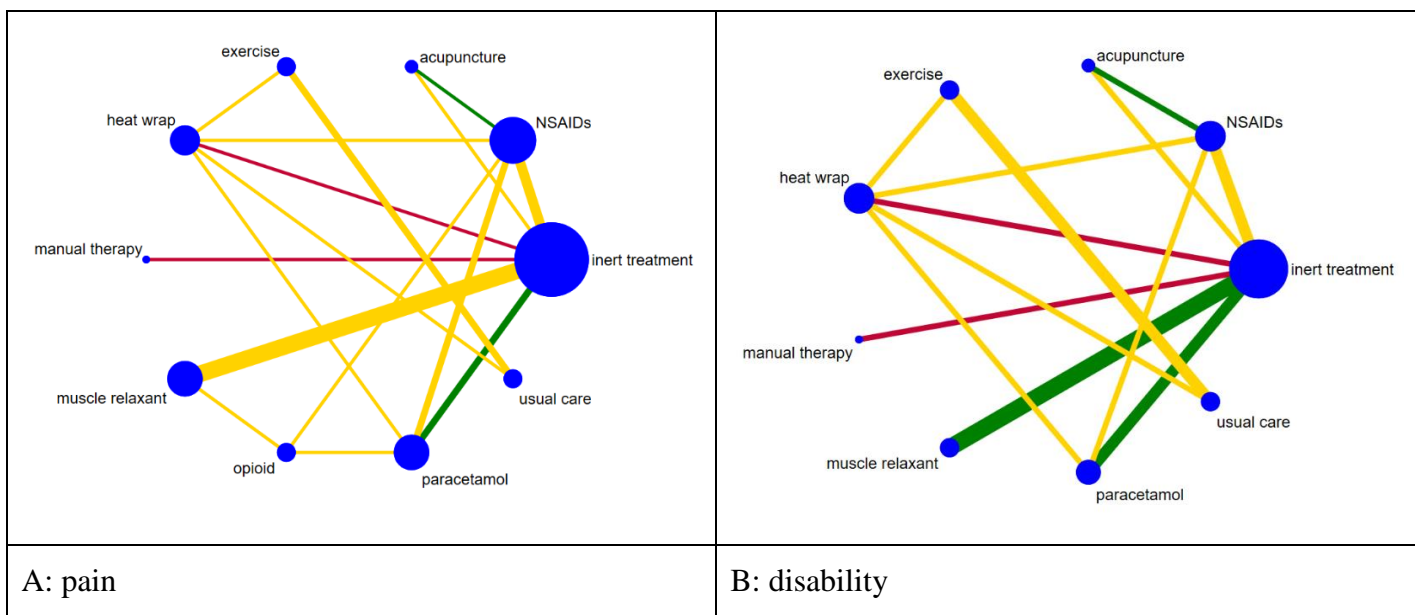


Figure 5. Network plot by risk of bias for pain and disability outcomes at short term of 1 week FU.

Pairwise comparisons

Treatment effects in pairwise meta-analyses are shown in Figures 6a for pain and 6b for disability.

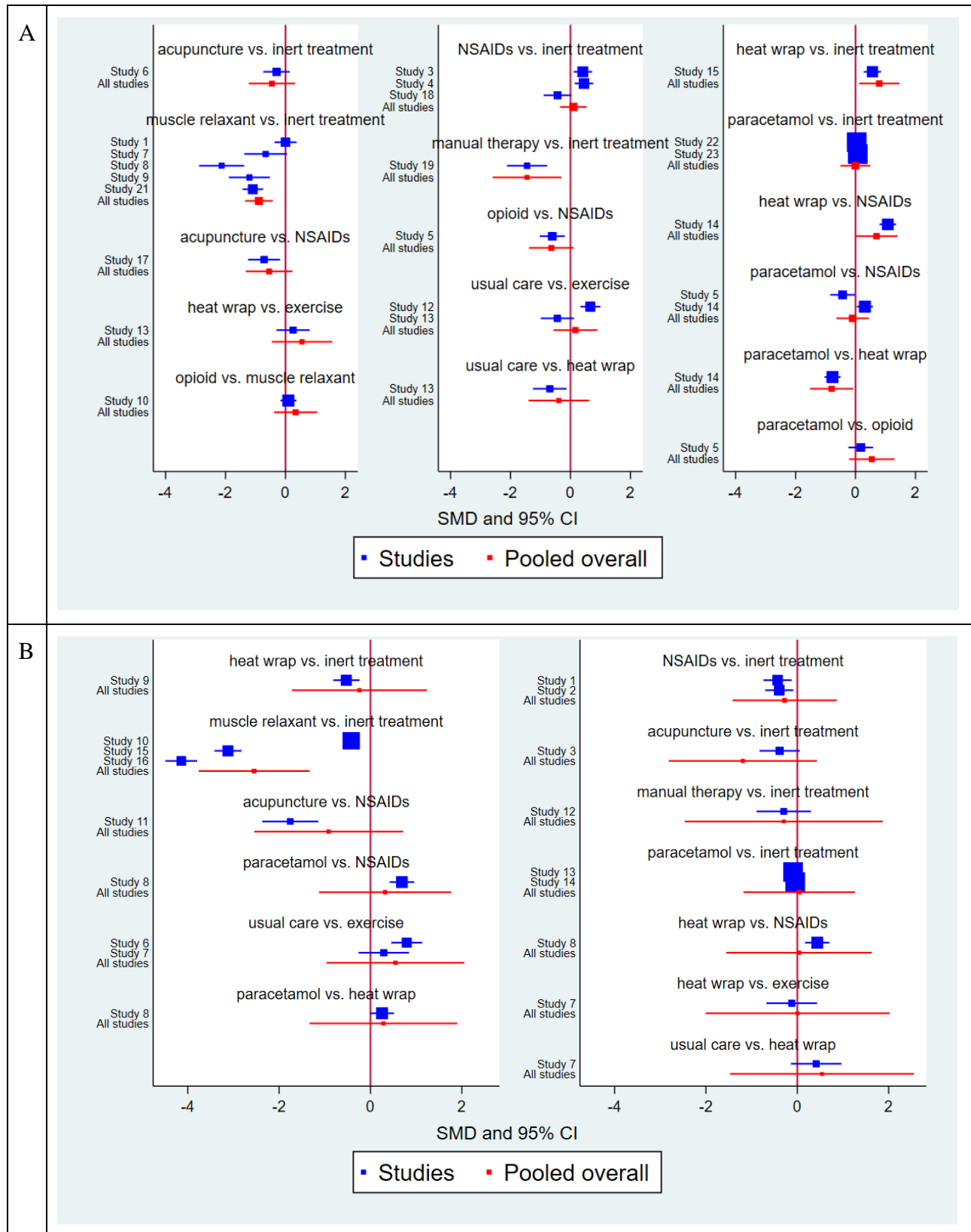


Figure 6a and 6b. Pairwise comparison for pain (A) and disability (B).

Inconsistency assessment

We tested inconsistency using the design-by-treatment interaction model (White 2011) approach. We first tested it globally and we found no evidence of inconsistency for both pain and disability (pain: $p=0.7639$, $\chi^2=3.35$ on 6 df; disability: $p=0.92$, $\chi^2=0.96$ on 4 df). We then examined inconsistency for each loop in each outcome analysis.

- **Pain**

We found 7 triangular and 2 quadratic loops. Three out of 9 loops were inconsistent since the 95% CI exclude the 0 that is, the direct estimate of the summary effect differs statistically from the indirect estimate (Figure 7a). However, we also examined all of the indirect sources of the evidence at once in order to compare direct with the indirect evidence from the whole of the rest of network, making use of all (indirect) loops that connect two interventions in the comparison, using the *network sidesplit* STATA command. Direct evidence estimates were consistent with indirect evidence in every comparison. We could not identify any important variable that differed across comparisons in those loops, but the number of included studies was very small in the inconsistent loops.

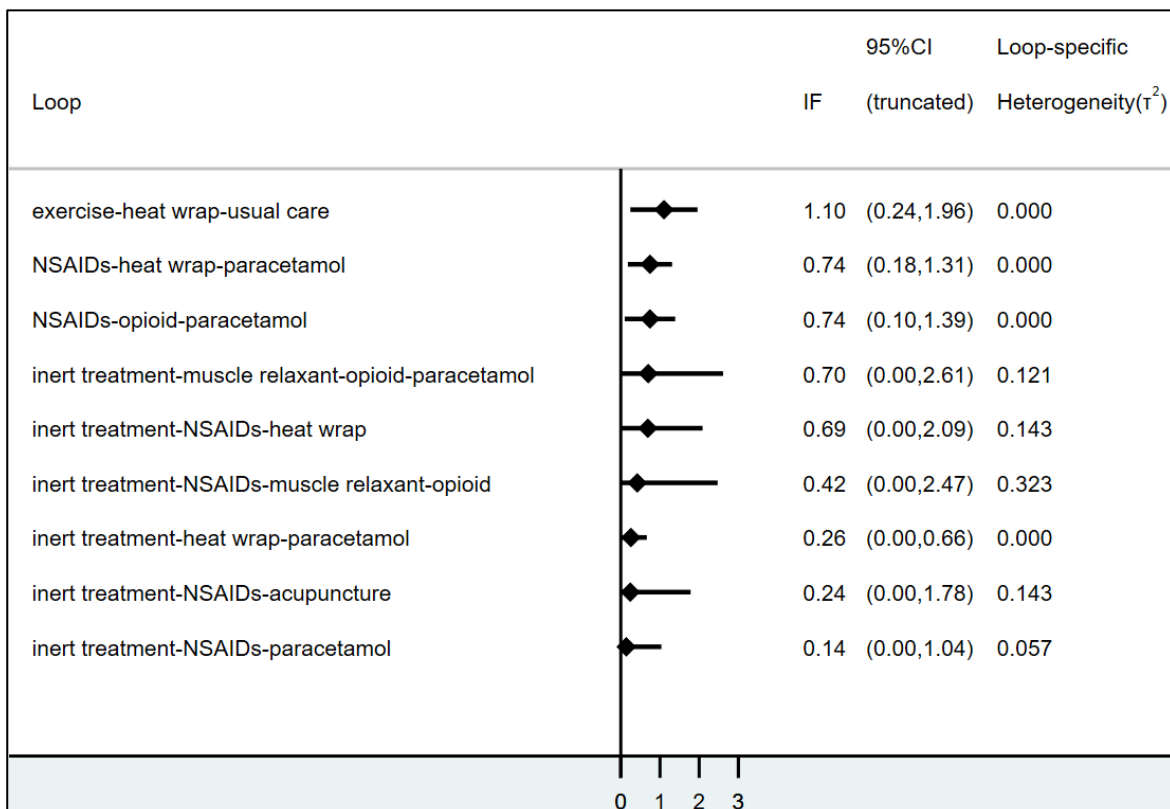


Figure 7a. Inconsistency plot for pain outcome.

- **Disability**

We found 5 triangular (Figure 7b). Two loops were inconsistent since the 95% CI exclude the 0 that is, the direct estimate of the summary effect differs statistically from the indirect estimate. As for pain outcome, we run the network *sidespilt* command with the results that direct evidence estimates were consistent with indirect evidence in every comparison. We could not identify any important variable that differed across comparisons in those loops, but the number of included studies was very small in the inconsistent loops.

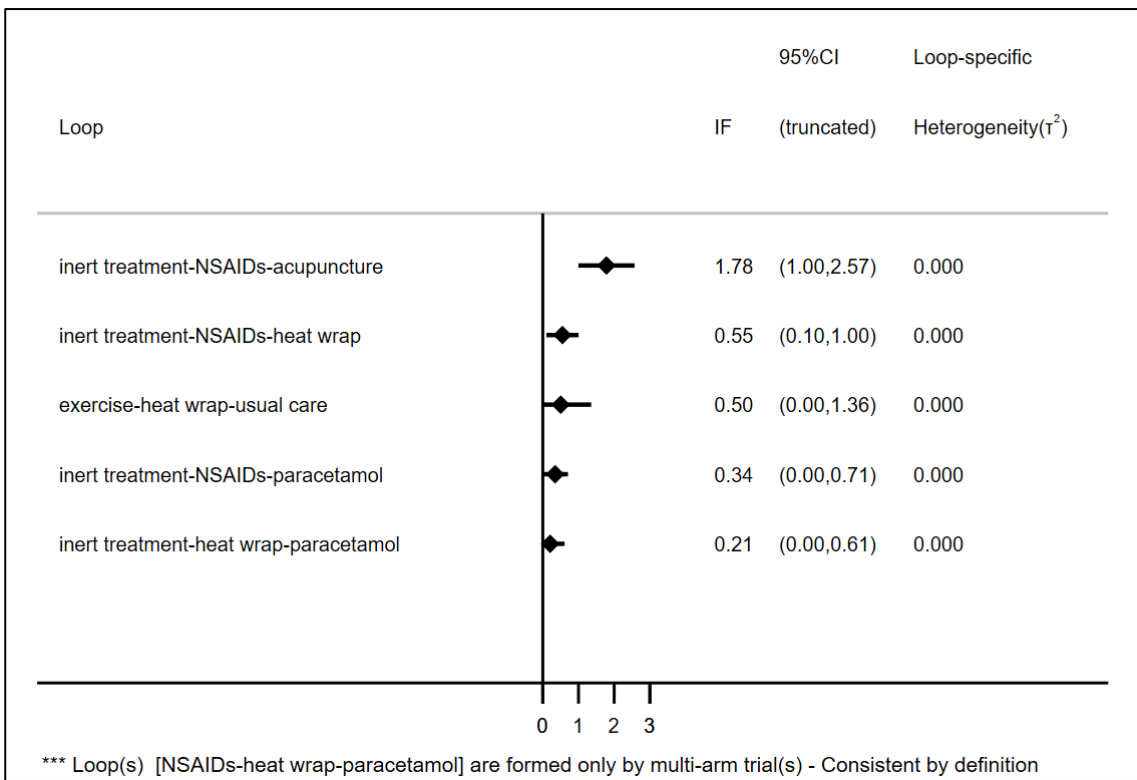


Figure 7b. Inconsistency plot for disability outcome.

Network meta-analysis results

NMA for pain showed that manual therapy (-1.45; 95%CI -2.60, -0.30) and muscle relaxants (-0.98; 95% CI -1.35, -0.42) significantly decreased pain compared to the inert treatment (**Figure 8a**).

NMA for disability showed that muscle relaxant drugs (-2.55; 95%CI -3.76, -1.33) is the only treatment that significantly decreased disability compared to the inert treatment (**Figure 8b**).

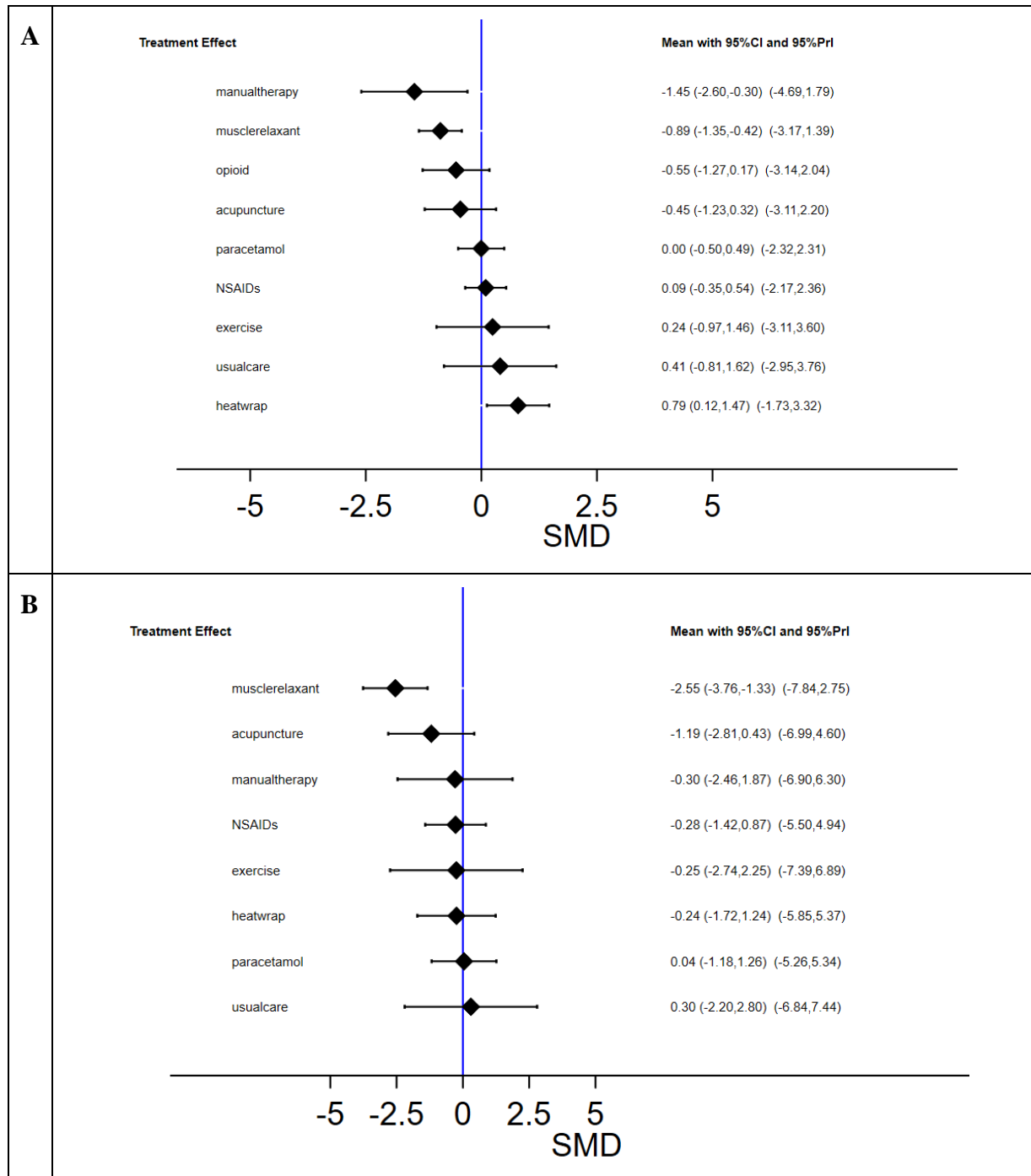


Figure 8a and 8b. NMA estimates and prediction intervals. Forest plots for pain (A) and disability (B), in acute and subacute LBP at immediate term vs inert treat.

Following, comparisons between treatments are presented in netleague tables. Results should be read from left to right and the estimate is in the cell in common between the column defining treatment and the row defining treatment.

inert treatment	0.41 (-0.81,1.62)	-0.00 (-0.50,0.49)	-0.55 (-1.27,0.17)	-0.89 (-1.35,-0.42)	-1.45 (-2.60,-0.30)	0.79 (0.12,1.47)	0.24 (-0.97,1.46)	-0.45 (-1.23,0.32)	0.09 (-0.35,0.54)
-0.41 (-1.62,0.81)	usual care	-0.41 (-1.65,0.83)	-0.95 (-2.33,0.42)	-1.29 (-2.58,-0.00)	-1.85 (-3.53,-0.18)	0.39 (-0.62,1.40)	-0.16 (-0.89,0.57)	-0.86 (-2.27,0.55)	-0.31 (-1.54,0.92)
0.00 (-0.49,0.50)	0.41 (-0.83,1.65)	paracetamol	-0.54 (-1.30,0.21)	-0.88 (-1.53,-0.23)	-1.44 (-2.70,-0.19)	0.80 (0.08,1.52)	0.25 (-0.99,1.49)	-0.45 (-1.33,0.43)	0.10 (-0.44,0.64)
0.55 (-0.17,1.27)	0.95 (-0.42,2.33)	0.54 (-0.21,1.30)	opioid	-0.34 (-1.06,0.38)	-0.90 (-2.26,0.46)	1.34 (0.41,2.28)	0.79 (-0.58,2.17)	0.09 (-0.93,1.11)	0.64 (-0.10,1.39)
0.89 (0.42,1.35)*	1.29 (0.00,2.58)*	0.88 (0.23,1.53)*	0.34 (-0.38,1.06)	muscle relaxant	-0.56 (-1.80,0.68)	1.68 (0.88,2.49)	1.13 (-0.16,2.42)	0.43 (-0.46,1.33)	0.98 (0.37,1.59)
1.45 (0.30,2.60)*	1.85 (0.18,3.53)*	1.44 (0.19,2.70)*	1.44 (0.19,2.70)*	0.90 (-0.46,2.26)	manual therapy	2.24 (0.91,3.57)	1.69 (0.02,3.36)	1.00 (-0.39,2.38)	1.54 (0.31,2.77)
-0.79 (-1.47,-0.12)*	-0.39 (-1.40,0.62)	-0.80 (-1.52,-0.08)*	-1.34 (-2.28,-0.41)*	-1.68 (-2.49,-0.88)*	-2.24 (-3.57,-0.91)*	heat wrap	-0.55 (-1.56,0.46)	-1.25 (-2.24,-0.26)	-0.70 (-1.40,0.00)
-0.24 (-1.46,0.97)	0.16 (-0.57,0.89)	-0.25 (-1.49,0.99)	-0.79 (-2.17,0.58)	-1.13 (-2.42,0.16)	-1.69 (-3.36,-0.02)*	0.55 (-0.46,1.56)	exercise	-0.70 (-2.11,0.72)	-0.15 (-1.38,1.08)
0.45 (-0.32,1.23)	0.86 (-0.55,2.27)	0.45 (-0.43,1.33)	-0.09 (-1.11,0.93)	-0.43 (-1.33,0.46)	-1.00 (-2.38,0.39)	1.25 (0.26,2.24)*	0.70 (-0.72,2.11)	acupuncture	0.55 (-0.23,1.32)
-0.09 (-0.54,0.35)	0.31 (-0.92,1.54)	-0.10 (-0.64,0.44)	-0.64 (-1.39,0.10)	-0.98 (-1.59,-0.37)*	-1.54 (-2.77,-0.31)*	0.70 (-0.00,1.40)	0.15 (-1.08,1.38)	-0.55 (-1.32,0.23)	NSAIDs

9a. Netleague table for pain. Effects of interventions are expressed in standardized mean differences. Legend: * indicates statistically significant differences.

inert treatment	0.30 (-2.20,2.80)	0.04 (-1.18,1.26)	-2.55 (-3.76,-1.33)	-0.30 (-2.46,1.87)	-0.24 (-1.72,1.24)	-0.25 (-2.74,2.25)	-1.19 (-2.81,0.43)	-0.28 (-1.42,0.87)
-0.30 (-2.80,2.20)	usual care	-0.26 (-2.84,2.32)	-2.85 (-5.63,-0.07)	-0.60 (-3.90,2.71)	-0.54 (-2.56,1.47)	-0.55 (-2.06,0.96)	-1.49 (-4.39,1.40)	-0.58 (-3.15,1.99)
-0.04 (-1.26,1.18)	0.26 (-2.32,2.84)	paracetamol	-2.59 (-4.31,-0.87)	-0.34 (-2.83,2.15)	-0.28 (-1.90,1.33)	-0.29 (-2.87,2.29)	-1.23 (-3.18,0.71)	-0.32 (-1.76,1.13)
2.55 (1.33,3.76)*	2.85 (0.07,5.63)*	2.59 (0.87,4.31)*	muscle relaxant	2.25 (-0.23,4.73)	2.31 (0.40,4.22)	2.30 (-0.48,5.08)	1.35 (-0.67,3.38)	2.27 (0.60,3.94)
0.30 (-1.87,2.46)	0.60 (-2.71,3.90)	0.34 (-2.15,2.83)	-2.25 (-4.73,0.23)	manual therapy	0.06 (-2.57,2.68)	0.05 (-3.26,3.36)	-0.90 (-3.60,1.81)	0.02 (-2.43,2.47)
0.24 (-1.24,1.72)	0.54 (-1.47,2.56)	0.28 (-1.33,1.90)	-2.31 (-4.22,-0.40)*	-0.06 (-2.68,2.57)	heat wrap	-0.01 (-2.02,2.01)	-0.95 (-3.04,1.13)	-0.04 (-1.63,1.56)
0.25 (-2.25,2.74)	0.55 (-0.96,2.06)	0.29 (-2.29,2.87)	-2.30 (-5.08,0.48)	-0.05 (-3.36,3.26)	0.01 (-2.01,2.02)	exercise	-0.95 (-3.84,1.95)	-0.03 (-2.60,2.54)
1.19 (-0.43,2.81)	1.49 (-1.40,4.39)	1.23 (-0.71,3.18)	-1.35 (-3.38,0.67)	0.90 (-1.81,3.60)	0.95 (-1.13,3.04)	0.95 (-1.95,3.84)	acupuncture	0.92 (-0.71,2.54)
0.28 (-0.87,1.42)	0.58 (-1.99,3.15)	0.32 (-1.13,1.76)	-2.27 (-3.94,-0.60)*	-0.02 (-2.47,2.43)	0.04 (-1.56,1.63)	0.03 (-2.54,2.60)	-0.92 (-2.54,0.71)	NSAIDs

9b. Netleague table for disability. Effects of interventions are expressed in standardized mean differences. Legend: * indicates statistically significant differences.

Ranking of treatments

- **Pain outcome**

Rank probability indicating the possibility of each intervention being the best (1) and then the worst (0) for pain are presented in table 2. In terms of efficacy, the most effective treatment was manual therapy (95.4%) and the last was the heat wrap (4.7%). Figure 10 shows the cumulative probability rank of the greatest likelihood of being the efficacious treatment for acute LBP.

Treatment	SUCRA	Probability of being best
Manual therapy	95.4	78.1
Muscle relaxants	86.3	14.2
Opioid	71.6	3.7
Acupuncture	66.8	3.1
Paracetamol	42.0	0.0
Inert treatment	41.3	0.0
NSAIDs	34.6	0.0
Exercise	33.1	0.7
Usual care	24.2	0.3
Heat wrap	4.7	0.0

Table 2. SUCRA for pain outcome.

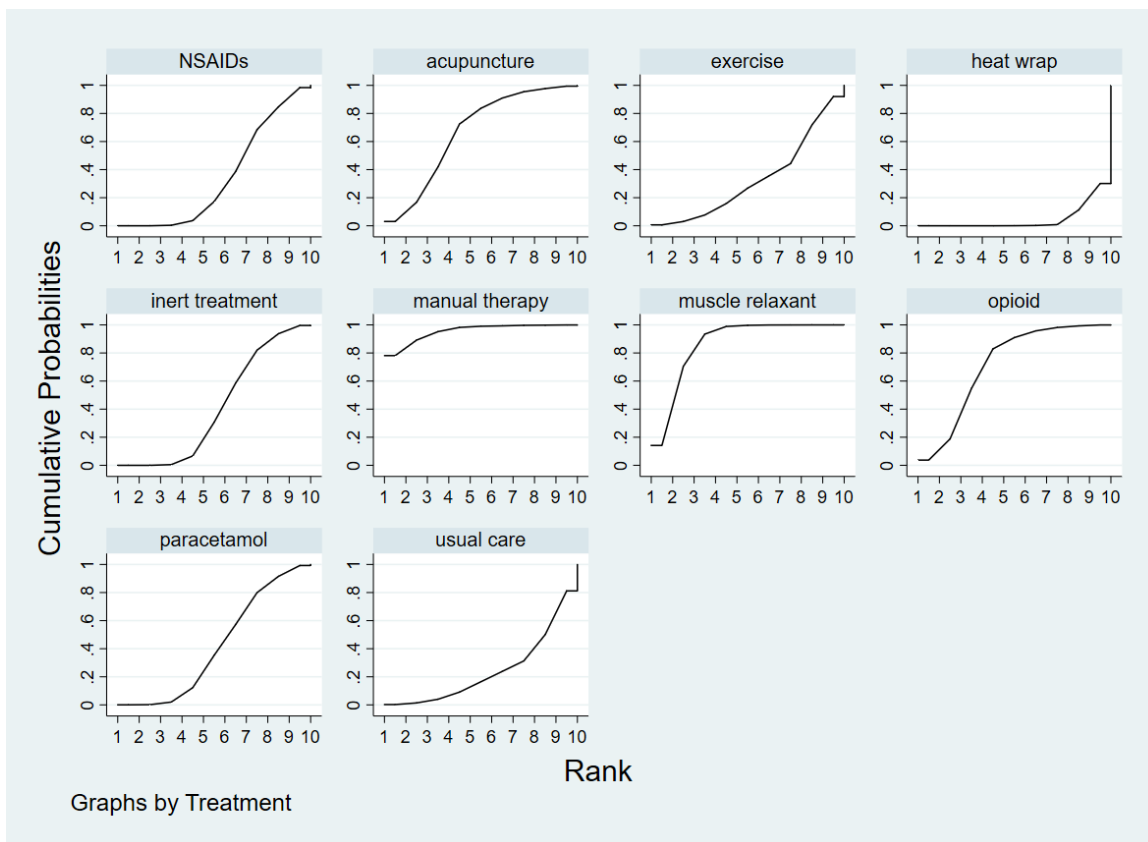


Figure 10. Cumulative SUCRA for pain.

- **Disability**

Rank probability indicating the possibility of each intervention being the best (1) and then the worst (9) for disability are presented in table 3. In terms of efficacy, the most effective treatment was muscle relaxant drugs (97.3%) and the less was the usual care (28.3%). Figure 11 shows the cumulative probability rank of the greatest likelihood of being the efficacious treatment for acute LBP.

Treatment	SUCRA	Probability of being best
Muscle relaxant	97.3	83.5
Acupuncture	74.0	8.3
Exercise	46.5	4.0
NSAIDs	45.9	0.1
Manual therapy	45.8	3.0
Heat wrap	45.5	0.3
Inert treatment	33.9	0.0
Paracetamol	32.8	0.1
Usual care	28.3	0.7

Table 3. Disability SUCRA.

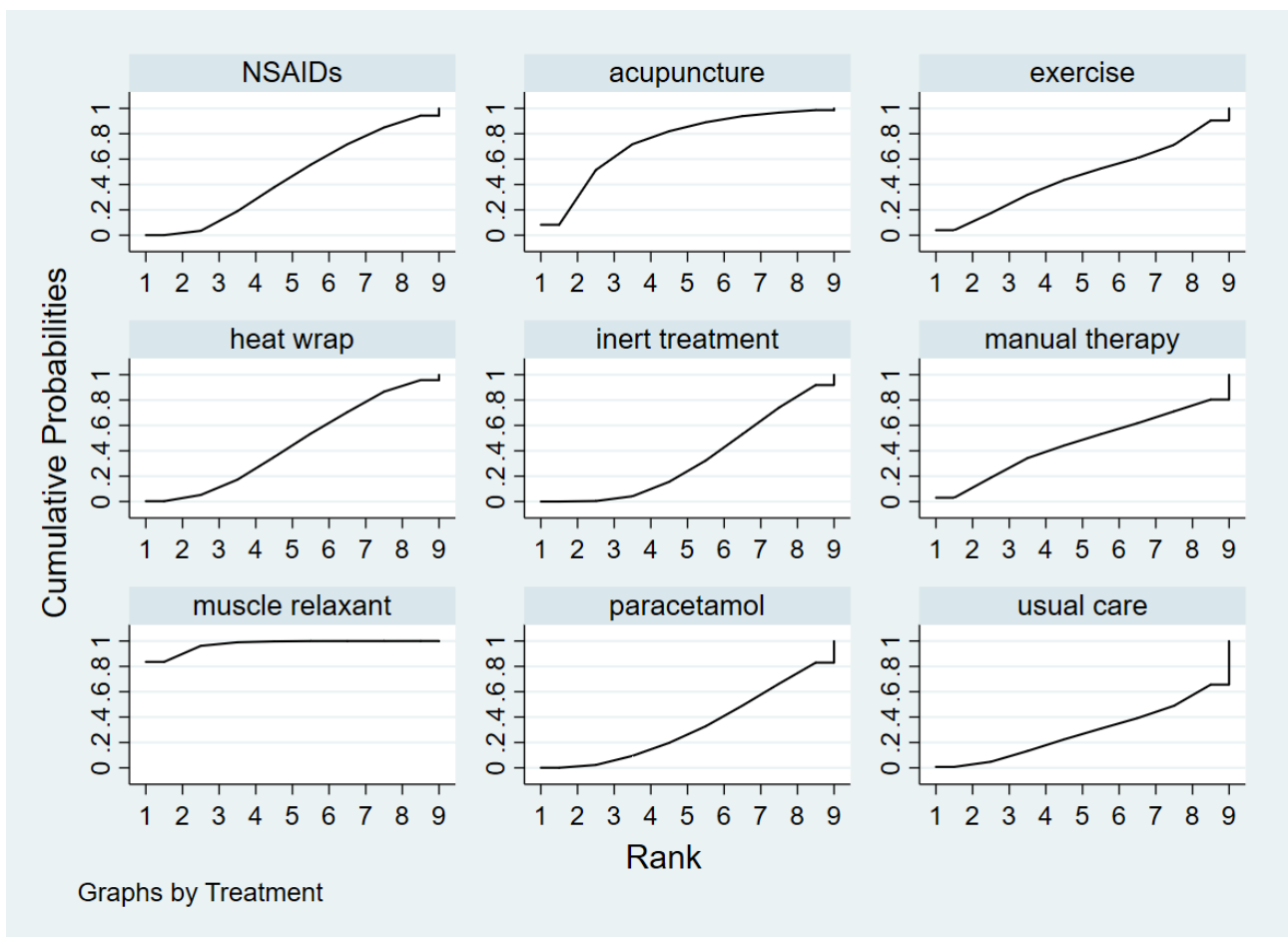


Figure 11. Cumulative SUCRA for disability.

Discussion

Our results showed that muscle relaxant and manual therapy are effective rehabilitative treatments in reducing pain at 1 week of follow up. These results are consistent with those published in literature by a recent systematic review where, for people with acute LBP, muscle relaxants provide clinically significant short-term pain relief (Abdel Shaheed, Maher et al. 2017). As well, international guidelines recommend muscle relaxants drugs as pharmacological treatment for acute low back pain even though with different strength of recommendations (van Tulder, Becker et al. 2006, Qaseem, Wilt et al. 2017, VA/DoD 2017). On the contrary, the NICE guideline recommend the use of NSAIDs for acute low back pain but our results are discordant (2016). This disagreement can be due to our stringent inclusion criteria of clinical trials: we selected only RCTs assessing “pure” treatments in patients with non-specific acute and sub-acute LBP (e.g., no sciatica), excluding all co-interventions (e.g., exercise plus NSAIDs), in order to be more conservative about the best available evidence preventing clinical and methodological intransitivity. We need, anyway, to be cautious in interpreting results on the efficacy of muscle relaxants drugs since, for example, thiocolchicoside has been lately advocated as dangerous with potential for genetic toxicity by the Italian Medicines Agency.

Several systematic reviews investigated the effect of manual therapy but we cannot compare manual therapy results with those published in these reviews since they investigated specific types of manual therapy techniques alone, i.e spinal manipulations or chiropractic treatments. In contrast, our results comprehend a wide range of manual therapy interventions: indeed, in real clinical practice, a health professional does not limit his intervention to a single manual therapy technique. Besides, integrating all treatments into one category, the statistical power of the manual therapy node was amplified.

Manual therapy resulted effective also in reducing disability. In agreement with our findings, the NICE guideline recommend “manual therapy (spinal manipulation, mobilisation or soft tissue techniques such as massage) for managing low back pain with or without sciatica, but only as part of a treatment package including exercise, with or without psychological therapy”(2016).

Overall, the loops in NMA analyses were generally represented by studies with unclear risk of bias. The next step will be to apply the GRADE assessment onto our results: the application of GRADE is more complex in systematic reviews with network meta-analysis than in the traditional systematic reviews with meta-analysis. The quality of a specific trial can affect more estimates in the network in both direct or indirect comparisons (Brignardello-Petersen, Bonner et al. 2018).

Our results derived from several small trials. In general, the NMA, combining indirect and direct evidence, might increase the power and precision of treatment effect estimates. Despite this, some authors suggest to cautiously interpret this analysis because even a network meta analyses may be of

low power and might influence estimates of the effect and the SUCRA values. Clinicians can be mistakenly endorse higher ranked treatments when no clinical relevant differences actually exist: disparities between clinical implications among treatments can be due to “play of chance” rather than real differences (Faltinsen, Storebo et al. 2018). The risk of random errors is a concept currently omnipresent and to which attention has been more and more paid in the evidence synthesis process mostly because of the introduction of the “living systematic reviews” and the “living guidelines” where the evidence is updated as soon as new relevant data emerge (Akl, Meerpohl et al. 2017, Elliott, Synnot et al. 2017). For instance, Faltinsen et al. recently suggested to assess the usefulness of Trial Sequential Analysis for network meta analyses and they called for future research on it (Faltinsen, Storebo et al. 2018).

Conclusion

The NMA is a powerful tool for evidence synthesis generation and appear to be of great promises but it can also be dangerous if not adequately performed and interpreted. Indeed, clinical and methodological assumption and nodes of the intervention network should be clearly defined in a protocol stage in order to not fall into bias besides being observed and controlled throughout all the NMA analysis process. An adequate and appropriate reporting should be also required in the context of evidence quality, statistical power, random errors and treatment rankings. In conclusion, we realized that NMA is resource expensive: it usually involves a lot of studies at each step of the systematic review from screening phase to data extraction, more than a conventional systematic review.

CONCLUSIONS

Despite the use of meta-analysis has been internationally widespread for 40 years by now, signs of a “midlife crisis” have been shown (Gurevitch, Koricheva et al. 2018). This project reflects the endless need in medicine to find and provide the best and the most reliable available evidence for supporting clinical decision making process and health care planning. It has enlightened limitations and threats in scientific process that surrounds evidence synthesis, encompassing both the unit of analysis of meta-analysis, the primary studies, and the next hierarchical level, the meta-analysis itself.

Our studies point out that inadequate reporting of primary studies, underpowered meta-analysis and lack of head-to-head comparisons can have negative influence on clinical-decision making. However, there are innovative methods that can mitigate these problems..

Reporting issues in RCTs and systematic reviews

Evidence syntheses are not immune to shortcoming in primary studies: one of the most common threats is inadequate reporting. The diffusion of meta-analyses to evaluate primary studies was associated with an increasing consciousness of the relevance of reporting. When a clinical trial fails to report essential elements, as we showed in chapter 1, its results should be interpreted with caution. Reporting shortcomings in primary studies focused on a common rehabilitation clinical problem – back pain - are highly frequent. We found that reporting issues are also spread across publication types, and meta-analyses and systematic reviews, including Cochrane reviews, are not immune. Therefore, inadequate reporting is likely to afflict to some extent all kind of publications at each level of hierarchy, possibly extending to tertiary publication types such as guidelines.

Even though many reporting guidance documents and checklists have been developed during the last 20 years, adhere to them seems to be slow and, at best, partial (Jin, Sanger et al. 2018). Our studies show that a deep gap is still present: these checklists seem to have scarce appeal on who should adhere to them. Despite partial failure in changing reporting, reporting checklists and tools have been increasing in terms of number and types of publication coverage, extending to very specific types of publications (e.g. CONSORT for cluster-RCTs) or publication sub-section (e.g. PRISMA-diagnostic test accuracy for abstracts checklist). It is difficult to imagine how all these checklists will make a difference given that the ancestor checklists, such as CONSORT and PRISMA did improve reporting

to a limited extent. Authors of scientific articles are required to follow reporting guidelines: editors and journals emphasise the importance of adhering to checklists, and this often is a pre-requisite to publish research articles. However, a real control of it and how checklists are used seems to not be fully implemented: reporting is a “nice to have” detail that easily gets lost in the publication process. Journals should not just require the submission of a complete guideline checklist with author’s manuscript but they should be stricter in asking to reviewers and editors to check for the real adherence of these tools in order to allow for a significant and real reporting improvement.

For both clinical trials and systematic reviews, registering in advance study design protocol can limit reporting bias. Indeed, the advanced registration is linked to higher methodological quality (Ge, Tian et al. 2018).

A direct factor on which to invest efforts and resources to improve reporting and research conduction is education on high quality research methods. Researchers, medical doctors and professional health practitioners are exposed to light statistical and critical appraisal courses for analysing and interpreting evidence and conduct adequate study design. Until quantitative methods and clinical epidemiology do not become part of academic curricula, at least of those professionals who intend to do research as part of their activities, research reporting and methods will likely to not improve much. As some authors suggested, training in research methods and interpretation of findings should be part of the basic training for higher-degree in medical fields, including research post-graduates, medical doctors and other professional science practitioners (Gurevitch, Koricheva et al. 2018).

Precision, sample size and heterogeneity from RCTs to systematic reviews

The pooled estimate of the effect of an intervention will always have some degree of uncertainty: what is desirable is to find a process to standardise (making replicable) the measurement of precision of estimates. Systematic reviews often include small studies that report greater and more variable effects than those reported by larger studies. Small studies are, indeed, more affected by random noise and publication bias: it has been demonstrated that sample size is associated with methodological quality, with small studies on average having more methodological threats (Kjaergard, Villumsen et al. 2001).

On one hand, our study shows how small studies often do not include a sample size calculation as part of their methods. On the other hand, meta-analyses do not differ, as they seem to have the same problem with low numbers of events or participants, few studies, to achieve a conclusive, reliable and more precise result. It is interesting how the issue of sample size crosses different steps of research

production: its inadequacy can affect precision of results of primary studies and be reflected on the classic meta-analysis results or the network meta-analysis findings altering the assessment of the overall certainty of the evidence at GRADE level.

The concept of sample size and power are of paramount relevance, matching the importance of another methodological concept of systematic review science: the heterogeneity. We have shown how important it is to integrate the statistical manifestation of heterogeneity into an optimal information size calculated for a meta-analysis, as the Trial Sequential Analysis allows. Whereas, the clinical manifestation should be critically explored to avoid imbalance in the distribution of effect modifiers between primary studies. Define the heterogeneity is one of the key assumption in the network meta-analysis conduction.

The GRADE approach is the best attempt that the research community has developed in order to obtain a comprehensive assessment of the certainty of the evidence. In fact, imprecision, sample size and heterogeneity are all considered into the grading of the evidence.

Further efforts are needed to offer a friendlier presentation of certainty of results and increase education: the gap that still exists between research and clinical practice is enlarge by the inability of clinicians and other stakeholders to understand the implications of sample size or heterogeneity limiting factors.

Conclusions

Advances in evidence synthesis have been stimulated by many factors: the need for more precise and reliable results, the need for more standard and transparent methods to assess the body of the evidence, the need for innovative software for cumulative analysis and for multiple comparisons instead of pair-wise comparisons.

Over the years, we have shown how it has been achieved a remarkable level of knowledge in methods to synthetize the evidence with the attempt to offer more reliable interpretation of meta-analysis findings, creating the pre-condition to optimally inform the clinical decision making. Evidence syntheses, systematic reviews and meta-analyses are a unique transparent and accurate tool to offer the best available evidence to clinicians and stakeholders. It is important to remember that a balanced clinical decision should not only take into consideration the best available accurate and reliable evidence but must integrate it with the clinical expertise and the patient's values reflecting the evidence-based medicine model.

Appendix

Appendix 1

Search strategy PubMed

"Adult"[Mesh] OR adult*[Title/Abstract]

AND

back pain[Mesh] OR "acute low back pain" [Title/Abstract] OR "acute back pain"[Title/Abstract] OR
backache [Title/Abstract] OR lumbago [Title/Abstract] OR "back disorder"[Title/Abstract]

AND

Randomized Controlled Trial[ptyp] OR Controlled Clinical Trial[ptyp] OR RCT [Title/Abstract] OR
"Randomized Controlled Trial" [Title/Abstract] OR random* [Title/Abstract] OR trial[Title/Abstract]

AND

"Exercise"[Mesh] OR "Exercise therapy"[Mesh] OR Exercis* [Title/Abstract] OR training[Title/Abstract]
OR "motor control" [Title/Abstract] OR "back school" [Title/Abstract] OR Manipulation, Chiropractic
[Mesh] OR Manipulation, Orthopedic[Mesh] OR Manipulation, Osteopathic [Mesh] OR Manipulation,
Spinal [Mesh] OR Musculoskeletal Manipulations [Mesh] OR Chiropractic [Title/Abstract] OR
manipulation [Title/Abstract] OR manipulate[Title/Abstract] OR "Spinal Manipulation" [Title/Abstract] OR
"Lumbar Manipulation" [Title/Abstract] OR thrust [Title/Abstract] OR manual therap* [Title/Abstract] OR
mobilization [Title/Abstract] OR ACUPUNCTURE [Mesh] OR ACUPUNCTURE THERAPY [Mesh] OR
acupuncture [Title/Abstract] OR electro-acupuncture [Title/Abstract] OR acupressure [Title/Abstract] OR
dry-needling [Title/Abstract] OR Massage[Mesh] OR massage [Title/Abstract] OR "myofascial release"
[Title/Abstract] OR Trigger Points [Mesh] OR "Trigger Points" [Title/Abstract] OR Health Education
[Mesh] OR "Physical Education and Training"[Mesh] OR patient education [Mesh] OR "Patient-Centered
Care"[Mesh] OR "information booklet" [Title/Abstract] OR book* [Title/Abstract] OR pamphlet*
[Title/Abstract] OR leaflet* [Title/Abstract] OR poster* [Title/Abstract] OR education* [Title/Abstract] OR
information* [Title/Abstract] OR Diathermy[Mesh] OR Diatherm* [Title/Abstract] OR
tecar[Title/Abstract] OR Transcutaneous Electrical Nerve Stimulation [Mesh] OR TENS [Title/Abstract] OR
"Electric Nerve Stimulation" [Title/Abstract] OR "Electrical Stimulation therapy" [Title/Abstract] OR
Electrostimulation [Title/Abstract] OR "electric stimulation therapy" [Title/Abstract] OR electroanalgesia
[Title/Abstract] OR electroacupuncture [Title/Abstract] OR electromagnetic [Title/Abstract] OR
electrotherapy*[Title/Abstract] OR taping [Title/Abstract] OR tape*[Title/Abstract] OR
Kinesio[Title/Abstract] OR strap*[Title/Abstract] OR sound [Mesh] OR Ultrasonic Therapy [Mesh]
OR Ultrasonics [Mesh] OR Ultrasonography [Mesh] OR Ultrasonic Waves [Mesh] OR Ultrasonic*
[Title/Abstract] OR ultrasound [Title/Abstract] OR Low-Level Light Therapy [Mesh] OR laser
[Title/Abstract] OR infrared [Title/Abstract] OR ultraviolet [Title/Abstract] OR monochromatic
[Title/Abstract] OR drug therapy [Title/Abstract] OR NSAIDS[Title/Abstract] OR "Cyclooxygenase
Inhibitors" [Pharmacological Action] OR cyclooxygenase [Title/Abstract] OR cyclo-oxygenase
[Title/Abstract] OR Anti-Inflammatory Agents, Non-Steroidal [Pharmacological Action] OR Anti-
Inflammatory Agents, Non-Steroidal [Mesh] OR aspirin [Title/Abstract] OR acetylsalicyl* [Title/Abstract]
OR Salicylic Acid [Title/Abstract] OR carbasalate calcium [Title/Abstract] OR Diflunisal[Title/Abstract] OR
aceclofenac [Title/Abstract] OR alclofenac [Title/Abstract] OR Diclofenac [Title/Abstract] OR Indomethacin

[Title/Abstract] OR Sulindac [Title/Abstract] OR meloxicam [Title/Abstract] OR Piroxicam [Title/Abstract] OR dexibuprofen [Title/Abstract] OR dexketoprofen [Title/Abstract] OR Fenoprofen [Title/Abstract] OR Flurbiprofen [Title/Abstract] OR ibuprofen [Title/Abstract] OR ketoprofen [Title/Abstract] OR Naproxen [Title/Abstract] OR metamizol [Title/Abstract] OR Dipyrone [Title/Abstract] OR phenylbutazone [Title/Abstract] OR phenazone [Title/Abstract] OR Antipyrine [Title/Abstract] OR propyphenazone [Title/Abstract] OR celecoxib [Title/Abstract] OR etoricoxib [Title/Abstract] OR nabumeton [Title/Abstract] OR parecoxib [Title/Abstract] OR rofecoxib [Title/Abstract] OR celecoxib [Title/Abstract] OR valdecoxib [Title/Abstract] OR lumiracoxib [Title/Abstract] OR etoricoxib [Title/Abstract] OR parecoxib [Title/Abstract] OR vioxx [Title/Abstract] OR celebrex [Title/Abstract] OR bextra [Title/Abstract] OR prexige [Title/Abstract] OR arcoxia [Title/Abstract] OR etodolac [Title/Abstract] OR floctafenine [Title/Abstract] OR Meclofenamic Acid [Title/Abstract] OR meclufenamate [Title/Abstract] OR meloxicam [Title/Abstract] OR oxaprozin [Title/Abstract] OR piroxicam [Title/Abstract] OR tenoxicam [Title/Abstract] OR tolmetin [Title/Abstract] OR paracetamol [Title/Abstract] OR Acetaminophen [Mesh] OR “Analgesics, Opioid” [Pharmacological Action] OR M03\$ [Title/Abstract] OR muscle relax* [Title/Abstract] OR anti-spasm* [Title/Abstract] OR calmativ [Title/Abstract] OR carisoprodol [Title/Abstract] OR cyclobenzaprine [Title/Abstract] OR flexeril [Title/Abstract] OR metaxalone [Title/Abstract] OR methocarbamol [Title/Abstract] OR baclofen [Title/Abstract] OR orphenadrine [Title/Abstract] OR tizanidine [Title/Abstract] OR Zanaflex [Title/Abstract] OR dantrolene [Title/Abstract] OR Dantrium [Title/Abstract] OR Quinine [Title/Abstract] OR chlorzoxazone [Title/Abstract] OR norflex [Title/Abstract] OR norgesic [Title/Abstract] OR alprazolam [Title/Abstract] OR xanax [Title/Abstract] OR Triazolam [Title/Abstract] OR Brotizolam [Title/Abstract] OR Oxazepam [Title/Abstract] OR Loprazolam [Title/Abstract] OR Lormetazepam [Title/Abstract] OR Lorazepam [Title/Abstract] OR Ativan [Title/Abstract] OR Temazepam [Title/Abstract] OR Normison [Title/Abstract] OR Temaz* [Title/Abstract] OR Estazolam [Title/Abstract] OR Bromazepam [Title/Abstract] OR Chlordiazepoxide [Title/Abstract] OR Clobazam [Title/Abstract] OR Nimetazepam [Title/Abstract] OR Flunitrazepam [Title/Abstract] OR Nitrazepam [Title/Abstract] OR Clonazepam [Title/Abstract] OR Quazepam [Title/Abstract] OR Diazepam [Title/Abstract] OR Valium [Title/Abstract] OR Phenazepam [Title/Abstract] OR Medazepam [Title/Abstract] OR Prazepam [Title/Abstract] OR Flurazepam [Title/Abstract] OR Clorazepate [Title/Abstract] OR Nordazepam [Title/Abstract] OR NO2A* [Title/Abstract] OR opioid [Title/Abstract] OR analges* [Title/Abstract] OR narcotic* [Title/Abstract] OR morphine [Title/Abstract] OR ordine [Title/Abstract] OR hydromorphone [Title/Abstract] OR dilaudid [Title/Abstract] OR oxycodone [Title/Abstract] OR endone [Title/Abstract] OR targin [Title/Abstract] OR oxymorphone [Title/Abstract] OR OPANA* [Title/Abstract] OR codeine [Title/Abstract] OR dihydrocodeine [Title/Abstract] OR ketobemidone [Title/Abstract] OR pethidine [Title/Abstract] OR Fentanyl [Title/Abstract] OR durogesic [Title/Abstract] OR diphenylpropylamine [Title/Abstract] OR dextromoramide [Title/Abstract] OR piritramide [Title/Abstract] OR dextropropoxyphene [Title/Abstract] OR bezitramide [Title/Abstract] OR methadone [Title/Abstract] OR physeptone [Title/Abstract] OR pentazocine [Title/Abstract] OR phenazocine [Title/Abstract] OR buprenorphine [Title/Abstract] OR norspan [Title/Abstract] OR suboxone [Title/Abstract] OR subutex [Title/Abstract] OR etorphine [Title/Abstract] OR tilidine [Title/Abstract] OR trama* [Title/Abstract] OR tramadol [Title/Abstract] OR dezocine [Title/Abstract] OR tapentadol [Title/Abstract] OR meptazinol [Title/Abstract] OR “benzodiazepines” [Mesh]

Appendix 2.

References of included studies

1. Berry H, Hutchinson DR. A multicentre placebo-controlled study in general practice to evaluate the efficacy and safety of tizanidine in acute low-back pain. *The Journal of international medical research*.1988 Mar-Apr;16(2):75-82.
2. Casale R. Symptomatic treatment with a muscle relaxant drug. *The Clinical journal of pain*.1988 (4):81-88.
3. Dreiser RL, Marty M, Ionescu E, et al. Relief of acute low back pain with diclofenac-K 12.5 mg tablets: a flexible dose, ibuprofen 200 mg and placebo-controlled clinical trial. *Int J Clin Pharm Th*.2003 Sep;41(9):375-385.
4. Eken C, Serinken M, Elicabuk H, et al. Intravenous paracetamol versus dexketoprofen versus morphine in acute mechanical low back pain in the emergency department: a randomised double-blind controlled trial. *Emerg Med J*.2014 Mar;31(3):177-181.
5. Hagen EM, Odélien KH, Lie SA, et al. Adding a physical exercise programme to brief intervention for low back pain patients did not increase return to work. *Scand J Public Health*.2010 Nov;38(7):731-738.
6. Hasegawa TM, Baptista AS, de Souza MC, et al. Acupuncture for acute non-specific low back pain: a randomised, controlled, double-blind, placebo trial. *Acupunct Med*.2014 Apr;32(2):109-115.
7. Hindle TH. Comparison of Carisoprodol, Butabarbital, and Placebo in Treatment of Low Back Syndrome. *Calif Med*.1972;117(2):7-&.
8. Hsieh CYJ, Adams AH, Tobis J, et al. Effectiveness of four conservative treatments for subacute low back pain - A randomized clinical trial. *Spine*.2002 Jun 1;27(11):1142-1148.
9. Jellema P, van der Windt DAWM, van der Horst HE, et al. Should treatment of (sub)acute low back pain be aimed at psychosocial prognostic factors? Cluster randomised clinical trial in general practice. *Brit Med J*.2005 Jul 9;331(7508):84-87.
10. Ketenci A, Ozcan E, Karamursel S. Assessment of efficacy and psychomotor performances of thiocolchicoside and tizanidine in patients with acute low back pain. *Int J Clin Pract*.2005 Jul;59(7):764-770.
11. Li C, Ni J, Wang Z, et al. Analgesic efficacy and tolerability of flupirtine vs. tramadol in patients with subacute low back pain: a double-blind multicentre trial. *Current medical research and opinion*.2008 Dec;24(12):3523-3530.
12. Lindstrom I, Ohlund C, Nachemson A. Physical performance, pain, pain behavior and subjective disability in patients with subacute low back pain. *Scandinavian journal of rehabilitation medicine*.1995 Sep;27(3):153-160.
13. Linton SJ, Andersson T. Can chronic disability be prevented? A randomized trial of a cognitive-behavior intervention and two forms of information for patients with spinal pain. *Spine*.2000 Nov 1;25(21):2825-2831.
14. Little P, Roberts L, Blowers H, et al. Should we give detailed advice and information booklets to patients with back pain? A randomized controlled factorial trial of a self-management booklet and doctor advice to take exercise for back pain. *Spine*.2001 Oct 1;26(19):2065-2072.
15. Machado LAC, Maher CG, Herbert RD, et al. The effectiveness of the McKenzie method in addition to first-line care for acute low back pain: a randomized controlled trial. *BMC medicine*.2010 Jan 26;8.
16. Malmivaara A, Hakkinen U, Aro T, et al. The Treatment of Acute Low-Back-Pain - Bed Rest, Exercises, or Ordinary Activity. *New Engl J Med*.1995 Feb 9;332(6):351-355.
17. Mayer JM, Ralph L, Look M, et al. Treating acute low back pain with continuous low-level heat wrap therapy and/or exercise: a randomized controlled trial. *The spine journal : official journal of the North American Spine Society*.2005 Jul-Aug;5(4):395-403.

18. Moffett JK, Torgerson D, Bell-Syer S, et al. Randomised controlled trial of exercise for low back pain: clinical outcomes, costs, and preferences. *BMJ*.1999 Jul 31;319(7205):279-283.
19. Nadler SF, Steiner DJ, Erasala GN, et al. Continuous low-level heat wrap therapy provides more efficacy than ibuprofen and acetaminophen for acute low back pain. *Spine*.2002 May 15;27(10):1012-1017.
20. Nadler SF, Steiner DJ, Erasala GN, et al. Continuous low-level heatwrap therapy for treating acute nonspecific low back pain. *Arch Phys Med Rehab*.2003 Mar;84(3):329-334.
21. Postacchini F, Facchini M, Palmieri P. Efficacy of various forms of conservative treatments in low back pain. *Neur Orthop*.1988 Dec;6:28-35.
22. del Pozo-Cruz B, Gusi N, del Pozo-Cruz J, et al. Clinical effects of a nine-month web-based intervention in subacute non-specific low back pain patients: a randomized controlled trial. *Clin Rehabil*.2013 Jan;27(1):28-39.
23. Rabin A, Shashua A, Pizem K, et al. A Clinical Prediction Rule to Identify Patients With Low Back Pain Who Are Likely to Experience Short-Term Success Following Lumbar Stabilization Exercises: A Randomized Controlled Validation Study. *J Orthop Sport Phys*.2014 Jan;44(1):6-18.
24. Ralph L, Look M, Wheeler W, et al. Double-blind, placebo-controlled trial of carisoprodol 250-mg tablets in the treatment of acute lower-back spasm. *Current medical research and opinion*.2008 Feb;24(2):551-558.
25. Schneider M, Haas M, Glick R, et al. Comparison of Spinal Manipulation Methods and Usual Medical Care for Acute and Subacute Low Back Pain A Randomized Clinical Trial. *Spine*.2015 Feb 15;40(4):209-217.
26. Seferlis T, Nemeth G, Carlsson AM, et al. Conservative treatment in patients sick-listed for acute low-back pain: a prospective randomised study with 12 months' follow-up. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*.1998;7(6):461-470.
27. Serfer GT, Wheeler WJ, Sacks HK. Randomized, double-blind trial of carisoprodol 250 mg compared with placebo and carisoprodol 350 mg for the treatment of low back spasm. *Current Medical Research and Opinion*.2009; 26:1, 91-99.
28. Shin JS, Ha IH, Lee J, et al. Effects of motion style acupuncture treatment in acute low back pain patients with severe disability: A multicenter, randomized, controlled, comparative effectiveness trial. *Pain*.2013 Jul;154(7):1030-1037.
29. Staal JB, Hlobil H, Twisk JWR, et al. Graded activity for low back pain in occupational health care - A randomized, controlled trial. *Annals of internal medicine*.2004 Jan 20;140(2):77-84.
30. Storheim K, Brox JI, Holm I, et al. Intensive group training versus cognitive intervention in sub-acute low back pain: Short-term results of a single-blind randomized controlled trial. *J Rehabil Med*.2003 May;35(3):132-140.
31. Szpalski M, Hayez JP. Objective Functional Assessment of the Efficacy of Tenoxicam in the Treatment of Acute Low-Back-Pain - a Double-Blind Placebo-Controlled Study. *Brit J Rheumatol*.1994 Jan;33(1):74-78.
32. Takamoto K, Bito I, Urakawa S, et al. Effects of compression at myofascial trigger points in patients with acute low back pain: A randomized controlled trial. *European Journal of Pain*.2015 Sep;19(8):1186-1196.
33. Tuzun F, Unalan H, Oner N, et al. Multicenter, randomized, double-blinded, placebo-controlled trial of thiocolchicoside in acute low back pain. *Joint Bone Spine*.2003 Sep;70(5):356-361.
34. Wand BM, Bird C, McAuley JH, et al. Early intervention for the management of acute low back pain - A single-blind randomized controlled trial of biopsychosocial education, manual therapy, and exercise. *Spine*.2004 Nov 1;29(21):2350-2356.
35. Whitfill T, Haggard R, Bierner SM, et al. Early Intervention Options for Acute Low Back Pain Patients: A Randomized Clinical Trial with One-Year Follow-Up Outcomes. *J Occup Rehabil*.2010 Jun;20(2):256-263.
36. Williams CM, Maher CG, Latimer J, et al. Efficacy of paracetamol for acute low-back pain: a double-blind, randomised controlled trial. *Lancet*.2014 Nov 1;384(9954):1586-1596.

Appendix 3.

Study and Patient characteristics (n=36) for transitivity assessment.

ID	Author	Year	Setting	Stage of LBP	Length of treatment	Outcomes	Week of FU	Sample size	Treatments	Age mean	Age variance (SD)	% of male
1	Berry	1988	Single center	Acute LBP (less than 6 weeks)	1 week	Pain	1 week	113	1.Tizanidine 2.placebo	44 38	13 13	51 51
2	Casale	1988	Single center	Acute LBP (less than 6 weeks)	1 week	Pain	1 week	20	1.Dantrolene sodium 2.placebo	46,70 47,10	2,3 2,2	70 80
3	Dreiser	2003	Multi-center	Acute LBP (less than 6 weeks)	1 week	Pain; disability	Day 3; day 8	369	1.Diclofenac NSAIDs 2. Ibuprofen 3. placebo	40,9 40,6 41	10,9 11,6 11,3	48,4 52,2 47,2
4	Eken	2014	Single center	Acute LBP (less than 6 weeks)	1 day	Pain	1 day	137	1. Morphine 2.paracetamol 3. Dexketoprofen	31,5	9,5	60,6
5	Hagen	2010	Single center	Subcute LBP (6-12 weeks)	8 weeks	Pain; disability	6 months; 12 months	246	1. exercise 2.minimal intervention	40,7 41,6	10,5 11	48 50
6	Hasegawa	2014	Single center	Acute LBP (less than 6 weeks)	1 week	Pain; disability	1 week; 1 month	80	1. acupuncture 2. sham acupuncture	47 43,9	9,8 10,9	37,5 35
7	Hindle	1972	Single center	Acute LBP (less than 6 weeks)	not reported	Pain; disability	1 week	48	1. cariprosodol 2. butabarbital 3. placebo	37 34,6 43,5	NA NA NA	56 50 62
8	Hsieh	2002	Single center	Mixed (acute and subacute)	3 weeks	Pain; disability	3 weeks	148	1.manipulation 2.myofascial therapy 3.exercise	47,4 49 47,9	14 14,8 13,7	67,3 66,7 60,4

9	Jallama	2005	Multi-center	Mixed LBP (less than 12 weeks)	1 week	Pain; disability	6, 13, 26, 52 weeks	314	1.minimal intervention strategy as behavioural therapy 2.usual care	43,2 42	11,1 12	52 53
10	Ketenci	2005	Single center	Acute LBP (less than 6 weeks)	1 week	Pain	1 week	97	1.thiocolchicoside 2. tizanidine 3.placebo	37 37 40	NA NA NA	58 37 48
11	Li	2008	Multi-center	Subacute LBP (6-12 weeks)	1 week	Pain	1 week	220	1. flupirtine 2.tramadol	46 46,5	15,1 15,7	35 46
12	Linstrom	1992	Single center	Mixed LBP (less than 12 weeks)	not reported	Pain; disability	12 months	103	1.cognitive behavioural therapy 2.usual care	39,4 42,4	10,7 10,9	76,5 64
13	Linton	2000	Single center	Subacute LBP (6-12 weeks)	not reported	Pain; disability	12 months	243	1.cognitive-behavioural therapy 2.phamphlet-education 3.infopack-education	44 45 44	NA NA NA	30 29 26
14	Little	2001	Multi-center	Subacute LBP (6-12 weeks)	3 weeks	Pain; disability	1 week; 1month	234	1.no treatment 2 booklet-education 3.exercise	47 42 47	17 14 14	43 43 43
15	Machado	2010	Multi-center	Acute LBP (less than 6 weeks)	3 weeks	Pain; disability	1 week; 3 weeks	146	1.first line-usual care 2.McKenzie exercise	45,9 47,5	14,9 14,4	52 48
16	Malmivaara	1995	Multi-center	Acute LBP (less than 6 weeks)	Not reported	Pain; disability	3 weeks; 12 weeks	119	1.exercise 2.usual care 3. bad rest	40,8 41,1 NA	NA NA NA	40 29 NA
17	Mayer	2005	Multi-center	Mixed (acute and subacute)	5 days	Pain; disability	2-4 day; 1 week	76	1.heat wrap 2.exercise 3.education	29,3 32,6 31,3	9,9 10,3 10,9	NA NA NA
18	Moffett	1999	Multi-center	Subacute LBP (6-12 weeks)	4 weeks	Pain; disability	6 weeks; 6 months; 12 months	187	1.behavioural exercise 2.usual care	41,1 42,6	9,21 8,62	43 44

19	Nadler	2002	Multi-center	Subacute LBP (6-12 weeks)	2 days	Pain; disability	4 days	371	1. paracetamol ** 2. ibuprofen ** 3. heat wrap 4. unheated wrap 5. oral placebo	35,82 34,90 36,61 36,79 38,00	10,54 11,29 10,4 9,32 9,07	41,6 43,4 40,6 42,1 40
20	Nadler	2003	Multi-center	Subacute LBP (6-12 weeks)	3 days	Pain; disability	5 days	191	1.heat wrap** 2.placebo ** 3.unheated wrap 4.(ibuprofen) NSAIDs **	35,55 36,73 36,25 34,88	11,57 10,82 11,56 11,32	45,7 45,7 45,7 45,7
21	Postacchini	1988	Multi-center	Acute LBP (less than 6 weeks)	2 weeks	Pain	3 weeks; 3 months	61	1.bad rest 2.manipulation manual therapy 3. (diclofenac) NSAIDs 4.placebo	36,3	NA	55
22	Pozo-Cruz	2012	Multi-center	Subacute LBP (6-12 weeks)	36 weeks	Pain; disability	12 months	90	1.occupational exercise 2. education	46,83 45,50	9,13 7,02	15,2 11,4
23	Rabin	2014	Multi-center	Mixed (acute and subacute)	7weeks	Pain; disability	8 weeks	105	1.exercise 2.manual therapy	38,3 35,5	10,5 9,1	47,9 45,6
24	Ralph	2008	Multi-center	Acute LBP (less than 6 weeks)	1 days	Pain; disability	3 days	562	1.cariprosodol 2.placebo	39,3 41,5	11,82 11,7	51,3 45
25	Schneider	2015	Single center	Mixed (acute and subacute)	4 weeks	Pain; disability	4 weeks; 3 months; 6 months	112	1.manual manipulation 2.mechanical assisted manipulation 3.usual care	40,4 41,4 40,3	15,9 15,3 11,6	40 32,4 40
26	Seferlis	1998	Multi-center	Acute LBP (less than 6 weeks)	8 weeks	Pain; disability	1 months; 3 months; 12 months	180	1.exercise 2.general practitioner program-usual care 3. intensive training program mixed	39	19-64 range	52,8
27	Serfer	2010	Single center	Acute LBP (less than 6 weeks)	1 week	Pain; disability	1 week	806	1.carisoprodol mg (myorelaxant)	250 40,9 40,5 40,7	11,7 12,4 13,1	47,7 44,3 39,4

									2. carisoprodol 350 mg (myorelaxant)			
									3.placebo			
28	Shin	2013	Multi-center	Acute LBP (less than 6 weeks)	1 week	Pain; disability	2 weeks; 4 weeks; 24 weeks	58	1.acupuncture	37,93	7,37	66
									2.diclofenac	38,69	8,64	52
29	Staal	2004	Single center	Mixed LBP (less than 12 weeks)	12 weeks	Pain; disability	3 months; 6 months	134	1.behavioural graded activity	39	9	95,5
									2. usual care	37	8	95,5
30	Storheim	2003	Single center	Subacute LBP (6-12 weeks)	15 weeks	Pain; disability	18 weeks	93	1.behavioural exercise	42,3	9,2	46,7
									2. exercise	41,3	9,4	52,9
									3. control-usual care	38,9	11,9	44,8
31	Szpalski	1994	Single center	Acute LBP (less than 6 weeks)	2 weeks	Pain	1 week; 2 weeks	73	1.tenoxicam	37,5	9,2	62,2
									2.placebo	38,9	10,4	66,7
32	Takamoto	2015	Multi-center	Acute LBP (less than 6 weeks)	2 weeks	Pain; disability	1 week; 1 month	63	1.manual therapy	38	3	45,4
									2.placebo	38,1	3,8	47,1
									3.effleurage massage	Na	3	37,5
33	Tuzun	2005	Multi-center	Acute LBP (less than 6 weeks)	5 days	Pain	5 days	143	1.thiocolchicoside	40,7	10,3	50
									2.placebo	41	11	42
34	Wand	2004	Single center	Acute LBP (less than 6 weeks)	6 weeks	Pain; disability	6 weeks; 3 months; 6 months	102	1.exercise	34	9	55,8
									2. usual care	35	7,9	45,09
35	Whitfill	2010	Single center	Mixed LBP (less than 12 weeks)	10 weeks	Pain	12 months	102	1.cognitive behavioural therapy	42,3	Na	50
									2.usual care	37,6	Na	43,2
36	Williams	2014	Multi-center	Acute LBP (less than 6 weeks)	4 weeks	Pain; disability	1week; 1 month; 3 month;	1641	1.paracetamol	44,1	14,8	52
									2.paracetamol as needed	45,5	16,7	53
									3.placebo	45,4	15,9	55

**arms dropped

Bibliography

(2010). "Feasibility and challenges of independent research on drugs: the Italian medicines agency (AIFA) experience." Eur J Clin Invest **40**(1): 69-86.

(2011). TSA software. Copenhagen Trial Unit, Centre for Clinical Intervention Research, Copenhagen, Denmark.

(2016). Low Back Pain and Sciatica in Over 16s: Assessment and Management. London.

(MTM)., M.-T. M.-a. "A framework for evaluating and ranking multiple healthcare technologies." Retrieved <http://www.mtm.uoi.gr/> (accessed 8 February 2017).

Abdel Shaheed, C., C. G. Maher, K. A. Williams and A. J. McLachlan (2017). "Efficacy and tolerability of muscle relaxants for low back pain: Systematic review and meta-analysis." Eur J Pain **21**(2): 228-237.

Abdul Latif, L., J. E. Daud Amadera, D. Pimentel, T. Pimentel and F. Fregni (2011). "Sample size calculation in physical medicine and rehabilitation: a systematic review of reporting, characteristics, and results in randomized controlled trials." Arch Phys Med Rehabil **92**(2): 306-315.

Akl, E. A., J. J. Meerpohl, J. Elliott, L. A. Kahale, H. J. Schunemann and N. Living Systematic Review (2017). "Living systematic reviews: 4. Living guideline recommendations." J Clin Epidemiol **91**: 47-53.

Al-Marzouki, S., I. Roberts, S. Evans and T. Marshall (2008). "Selective reporting in clinical trials: analysis of trial protocols accepted by The Lancet." Lancet **372**(9634): 201.

AlBalawi, Z., F. A. McAlister, K. Thorlund, M. Wong and J. Wetterslev (2013). "Random error in cardiovascular meta-analyses: how common are false positive and false negative results?" Int J Cardiol **168**(2): 1102-1107.

Antes, G. (2010). "The new CONSORT statement." BMJ **340**: c1432.

Anttila, S., J. Persson, N. Vareman and N. E. Sahlin (2016). "Conclusiveness resolves the conflict between quality of evidence and imprecision in GRADE." J Clin Epidemiol **75**: 1-5.

Armijo-Olivo, S., S. Warren, J. Fuentes and D. J. Magee (2011). "Clinical relevance vs. statistical significance: Using neck outcomes in patients with temporomandibular disorders as an example." Man Ther **16**(6): 563-572.

Ayeni, O., L. Dickson, T. A. Ignacy and A. Thoma (2012). "A systematic review of power and sample size reporting in randomized controlled trials within plastic surgery." Plast Reconstr Surg **130**(1): 78e-86e.

Balshem, H., M. Helfand, H. J. Schunemann, A. D. Oxman, R. Kunz, J. Brozek, G. E. Vist, Y. Falck-Ytter, J. Meerpohl, S. Norris and G. H. Guyatt (2011). "GRADE guidelines: 3. Rating the quality of evidence." J Clin Epidemiol **64**(4): 401-406.

Bassler, D., V. M. Montori, M. Briel, P. Glasziou and G. Guyatt (2008). "Early stopping of randomized clinical trials for overt efficacy is problematic." J Clin Epidemiol **61**(3): 241-246.

Beaton, D. E., C. Bombardier, F. Guillemin and M. B. Ferraz (2000). "Guidelines for the process of cross-cultural adaptation of self-report measures." Spine (Phila Pa 1976) **25**(24): 3186-3191.

- Bender, R., C. Bunce, M. Clarke, S. Gates, S. Lange, N. L. Pace and K. Thorlund (2008). "Attention should be given to multiplicity issues in systematic reviews." J Clin Epidemiol **61**(9): 857-865.
- Bjelakovic, G., L. L. Gluud, D. Nikolova, K. Whitfield, J. Wetterslev, R. G. Simonetti, M. Bjelakovic and C. Gluud (2014). "Vitamin D supplementation for prevention of mortality in adults." Cochrane Database Syst Rev(1): CD007470.
- Boutron, I., D. Moher, D. G. Altman, K. F. Schulz and P. Ravaud (2008). "Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration." Ann Intern Med **148**(4): 295-309.
- Brignardello-Petersen, R., A. Bonner, P. E. Alexander, R. A. Siemieniuk, T. A. Furukawa, B. Rochweg, G. S. Hazlewood, W. Alhazzani, R. A. Mustafa, M. H. Murad, M. A. Puhan, H. J. Schunemann, G. H. Guyatt and G. W. Group (2018). "Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis." J Clin Epidemiol **93**: 36-44.
- Brok, J., L. D. Huusom and K. Thorlund (2012). "Conclusive meta-analyses on antenatal magnesium may be inconclusive! Are we underestimating the risk of random error?" Acta Obstet Gynecol Scand **91**(11): 1247-1251.
- Brok, J., K. Thorlund, J. Wetterslev and C. Gluud (2009). "Apparently conclusive meta-analyses may be inconclusive--Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses." Int J Epidemiol **38**(1): 287-298.
- Brok, J., K. Thorlund, J. Wetterslev and C. Gluud (2009). "Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses." International Journal of Epidemiology **38**(1): 287-298.
- Buchbinder, R., C. Maher and I. A. Harris (2015). "Setting the research agenda for improving health care in musculoskeletal disorders." Nat Rev Rheumatol.
- Calvert, M., J. Blazeby, D. G. Altman, D. A. Revicki, D. Moher and M. D. Brundage (2013). "Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension." JAMA **309**(8): 814-822.
- Castellini, G., S. Gianola, G. Banfi, S. Bonovas and L. Moja (2016). "Mechanical Low Back Pain: Secular Trend and Intervention Topics of Randomized Controlled Trials." Physiotherapy Canada **68**(1): 61-63.
- Castellini, G., S. Gianola, S. Bonovas and L. Moja (2016). "Improving Power and Sample Size Calculation in Rehabilitation Trial Reports: A Methodological Assessment." Arch Phys Med Rehabil **97**(7): 1195-1201.
- Castellini, G., S. Gianola, S. Bonovas and L. Moja (2016). "Improving Power and Sample Size Calculation in Rehabilitation Trial Reports: A Methodological Assessment." Arch Phys Med Rehabil.
- Castellini, G., S. Gianola, P. Frigerio, M. Agostini, R. Bolotta, D. Corbetta, M. Gasparini, P. Gozzer, E. Guariento, L. Li, V. Pecoraro, V. Sirtori, A. Turolla and L. Moja (2015). "Completeness of outcome description in studies for low back pain rehabilitation interventions: a survey of trials included in Cochrane reviews." Trials **16**(Suppl 1): P24-P24.
- Chaimani, A., D. M. Caldwell, T. Li, J. P. Higgins and G. Salanti (2017). "Additional considerations are required when preparing a protocol for a systematic review with multiple interventions." J Clin Epidemiol.
- Chaimani, A., J. P. Higgins, D. Mavridis, P. Spyridonos and G. Salanti (2013). "Graphical tools for network meta-analysis in STATA." PLoS One **8**(10): e76654.

Chalmers, I. and R. Matthews (2006). "What are the implications of optimism bias in clinical research?" Lancet **367**(9509): 449-450.

Chan, A. W. and D. G. Altman (2005). "Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors." BMJ **330**(7494): 753.

Chan, A. W., A. Hrobjartsson, M. T. Haahr, P. C. Gotzsche and D. G. Altman (2004). "Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles." JAMA **291**(20): 2457-2465.

Chan, A. W., J. M. Tetzlaff, D. G. Altman, A. Laupacis, P. C. Gotzsche, A. J. K. Krle, A. Hrobjartsson, H. Mann, K. Dickersin, J. A. Berlin, C. J. Dore, W. R. Parulekar, W. S. Summerskill, T. Groves, K. F. Schulz, H. C. Sox, F. W. Rockhold, D. Rennie and D. Moher (2015). "SPIRIT 2013 Statement: defining standard protocol items for clinical trials." Rev Panam Salud Publica **38**(6): 506-514.

Chan, K. B., M. Man-Son-Hing, F. J. Molnar and A. Laupacis (2001). "How well is the clinical importance of study results reported? An assessment of randomized controlled trials." CMAJ **165**(9): 1197-1202.

Chapter 5.2.4.2 Imprecision in in systematic reviews in Schünemann H, B. J., Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from guidelinedevelopment.org/handbook. .

Charles, P., B. Giraudeau, A. Dechartres, G. Baron and P. Ravaud (2009). "Reporting of sample size calculation in randomised controlled trials: review." BMJ **338**: b1732.

Clarke, J. A., M. W. van Tulder, S. E. Blomberg, H. C. de Vet, G. J. van der Heijden, G. Bronfort and L. M. Bouter (2007). "Traction for low-back pain with or without sciatica." Cochrane Database Syst Rev(2): CD003010.

Cochrane Cochrane Library, Wiley Online Library.

COEVIDENCE "<https://www.covidence.org/reviews>." accessed on January 2018.

Cook, J. A., J. Hislop, D. G. Altman, P. Fayers, A. H. Briggs, C. R. Ramsay, J. D. Norrie, I. M. Harvey, B. Buckley, D. Fergusson, I. Ford and L. D. Vale (2015). "Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers." Trials **16**(1): 12.

Copay, A. G., B. R. Subach, S. D. Glassman, D. W. Polly, Jr. and T. C. Schuler (2007). "Understanding the minimum clinically important difference: a review of concepts and methods." Spine J **7**(5): 541-546.

Crowe, S., M. Fenton, M. Hall, K. Cowan and I. Chalmers (2015). "Patients', clinicians' and the research communities' priorities for treatment research: there is an important mismatch." Res Involv Engagem **1**: 2.

DerSimonian, R. and N. Laird (1986). "Meta-analysis in clinical trials." Control Clin Trials **7**(3): 177-188.

Dias, S., N. J. Welton, D. M. Caldwell and A. E. Ades (2010). "Checking consistency in mixed treatment comparison meta-analysis." Stat Med **29**(7-8): 932-944.

Dwan, K., D. G. Altman, J. A. Arnaiz, J. Bloom, A. W. Chan, E. Cronin, E. Decullier, P. J. Easterbrook, E. Von Elm, C. Gamble, D. Ghersi, J. P. Ioannidis, J. Simes and P. R. Williamson (2008). "Systematic review of the empirical evidence of study publication bias and outcome reporting bias." PLoS One **3**(8): e3081.

- Ebadi, S., N. Henschke, N. Nakhostin Ansari, E. Fallah and M. W. van Tulder (2014). "Therapeutic ultrasound for chronic low-back pain." Cochrane Database Syst Rev **3**: CD009169.
- Elliott, J. H., A. Synnot, T. Turner, M. Simmonds, E. A. Akl, S. McDonald, G. Salanti, J. Meerpohl, H. MacLehose, J. Hilton, D. Tovey, I. Shemilt, J. Thomas and N. Living Systematic Review (2017). "Living systematic review: 1. Introduction-the why, what, when, and how." J Clin Epidemiol **91**: 23-30.
- Faltinsen, E. G., O. J. Storebo, J. C. Jakobsen, K. Boesen, T. Lange and C. Gluud (2018). "Network meta-analysis: the highest level of medical evidence?" BMJ Evid Based Med **23**(2): 56-59.
- Fisher, S. A., S. J. Brunskill, C. Doree, A. Mathur, D. P. Taggart and E. Martin-Rendon (2014). "Stem cell therapy for chronic ischaemic heart disease and congestive heart failure." Cochrane Database Syst Rev(4): CD007888.
- Fisher, S. A., C. Doree, A. Mathur and E. Martin-Rendon (2015). "Meta-analysis of cell therapy trials for patients with heart failure." Circ Res **116**(8): 1361-1377.
- Fisher, S. A., C. Doree, D. P. Taggart, A. Mathur and E. Martin-Rendon (2016). "Cell therapy for heart disease: Trial sequential analyses of two Cochrane reviews." Clin Pharmacol Ther **100**(1): 88-101.
- Fitzner, K. and E. Heckinger (2010). "Sample size calculation and power analysis: a quick review." Diabetes Educ **36**(5): 701-707.
- Freiman, J. A., T. C. Chalmers, H. Smith, Jr. and R. R. Kuebler (1978). "The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials." N Engl J Med **299**(13): 690-694.
- Froud, R., S. Patterson, S. Eldridge, C. Seale, T. Pincus, D. Rajendran, C. Fossum and M. Underwood (2014). "A systematic review and meta-synthesis of the impact of low back pain on people's lives." BMC Musculoskelet Disord **15**: 50.
- Furlan, A. D., M. Imamura, T. Dryden and E. Irvin (2008). "Massage for low-back pain." Cochrane Database Syst Rev(4): CD001929.
- Gagnier, J. J. and P. J. Kellam (2013). "Reporting and methodological quality of systematic reviews in the orthopaedic literature." J Bone Joint Surg Am **95**(11): e771-777.
- Garattini, S., J. C. Jakobsen, J. Wetterslev, V. Bertele, R. Banzi, A. Rath, E. A. Neugebauer, M. Laville, Y. Masson, V. Hivert, M. Eikermann, B. Aydin, S. Ngwabyt, C. Martinho, C. Gerardi, C. A. Szmigielski, J. Demotes-Mainard and C. Gluud (2016). "Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them." Eur J Intern Med.
- Garner, P., S. Hopewell, J. Chandler, H. MacLehose, H. J. Schunemann, E. A. Akl, J. Beyene, S. Chang, R. Churchill, K. Dearness, G. Guyatt, C. Lefebvre, B. Liles, R. Marshall, L. Martinez Garcia, C. Mavergames, M. Nasser, A. Qaseem, M. Sampson, K. Soares-Weiser, Y. Takwoingi, L. Thabane, M. Trivella, P. Tugwell, E. Welsh and E. C. Wilson (2016). "When and how to update systematic reviews: consensus and checklist." BMJ **354**: i3507.
- Ge, L., J. H. Tian, Y. N. Li, J. X. Pan, G. Li, D. Wei, X. Xing, B. Pan, Y. L. Chen, F. J. Song and K. H. Yang (2018). "Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study." J Clin Epidemiol **93**: 45-55.
- Geha, N. N., A. M. Moseley, M. R. Elkins, L. D. Chiavegato, S. R. Shiwa and L. O. Costa (2013). "The quality and reporting of randomized trials in cardiothoracic physical therapy could be substantially improved." Respir Care **58**(11): 1899-1906.

- Gianola, S., A. Andreano, G. Castellini, L. Moja and M. G. Valsecchi (2018). "Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: the need to present minimal important differences units in meta-analyses." Health Qual Life Outcomes **16**(1): 91.
- Gianola, S., G. Castellini, M. Agostini, R. Bolotta, D. Corbetta, P. Frigerio, M. Gasparini, P. Gozzer, E. Guariento, L. C. Li, V. Pecoraro, V. Sirtori, A. Turolla, A. Andreano and L. Moja (2016). "Reporting of Rehabilitation Intervention for Low Back Pain in Randomized Controlled Trials: Is the Treatment Fully Replicable?" Spine (Phila Pa 1976) **41**(5): 412-418.
- Gianola, S., G. Castellini, M. Agostini, R. Bolotta, D. Corbetta, P. Frigerio, M. Gasparini, P. Gozzer, E. Guariento, L. C. Li, V. Pecoraro, V. Sirtori, A. Turolla, A. Andreano and L. Moja (2016). "Reporting of Rehabilitation Intervention for Low Back Pain in Randomized Controlled Trials: Is the Treatment Fully Replicable?" Spine **41**(5): 412-418.
- Gianola, S., P. Frigerio, M. Agostini, R. Bolotta, G. Castellini, D. Corbetta, M. Gasparini, P. Gozzer, E. Guariento, L. C. Li, V. Pecoraro, V. Sirtori, A. Turolla, A. Andreano and L. Moja (2016). "Completeness of Outcomes Description Reported in Low Back Pain Rehabilitation Interventions: A Survey of 185 Randomized Trials." Physiotherapy Canada: 1-8.
- Grimshaw, J. M., M. P. Eccles, J. N. Lavis, S. J. Hill and J. E. Squires (2012). "Knowledge translation of research findings." Implement Sci **7**: 50.
- Gurevitch, J., J. Koricheva, S. Nakagawa and G. Stewart (2018). "Meta-analysis and the science of research synthesis." Nature **555**(7695): 175-182.
- Guyatt, G., A. D. Oxman, E. A. Akl, R. Kunz, G. Vist, J. Brozek, S. Norris, Y. Falck-Ytter, P. Glasziou, H. DeBeer, R. Jaeschke, D. Rind, J. Meerpohl, P. Dahm and H. J. Schunemann (2011). "GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables." J Clin Epidemiol **64**(4): 383-394.
- Guyatt, G. H., A. D. Oxman, R. Kunz, J. Brozek, P. Alonso-Coello, D. Rind, P. J. Devereaux, V. M. Montori, B. Freyschuss, G. Vist, R. Jaeschke, J. W. Williams, Jr., M. H. Murad, D. Sinclair, Y. Falck-Ytter, J. Meerpohl, C. Whittington, K. Thorlund, J. Andrews and H. J. Schunemann (2011). "GRADE guidelines 6. Rating the quality of evidence--imprecision." J Clin Epidemiol **64**(12): 1283-1293.
- Guyatt, G. H., A. D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, Y. Falck-Ytter, R. Jaeschke, G. Vist, E. A. Akl, P. N. Post, S. Norris, J. Meerpohl, V. K. Shukla, M. Nasser, H. J. Schunemann and G. W. Group (2011). "GRADE guidelines: 8. Rating the quality of evidence--indirectness." J Clin Epidemiol **64**(12): 1303-1310.
- Guyatt, G. H., A. D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, P. Glasziou, R. Jaeschke, E. A. Akl, S. Norris, G. Vist, P. Dahm, V. K. Shukla, J. Higgins, Y. Falck-Ytter, H. J. Schunemann and G. W. Group (2011). "GRADE guidelines: 7. Rating the quality of evidence--inconsistency." J Clin Epidemiol **64**(12): 1294-1302.
- Guyatt, G. H., A. D. Oxman, V. Montori, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, B. Djulbegovic, D. Atkins, Y. Falck-Ytter, J. W. Williams, Jr., J. Meerpohl, S. L. Norris, E. A. Akl and H. J. Schunemann (2011). "GRADE guidelines: 5. Rating the quality of evidence--publication bias." J Clin Epidemiol **64**(12): 1277-1282.
- Guyatt, G. H., A. D. Oxman, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, V. Montori, E. A. Akl, B. Djulbegovic, Y. Falck-Ytter, S. L. Norris, J. W. Williams, Jr., D. Atkins, J. Meerpohl and H. J. Schunemann (2011). "GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias)." J Clin Epidemiol **64**(4): 407-415.

Guyatt, G. H., A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello and H. J. Schunemann (2008). "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations." *BMJ* **336**(7650): 924-926.

Guyatt, G. H., K. Thorlund, A. D. Oxman, S. D. Walter, D. Patrick, T. A. Furukawa, B. C. Johnston, P. Karanicolas, E. A. Akl, G. Vist, R. Kunz, J. Brozek, L. L. Kupper, S. L. Martin, J. J. Meerpohl, P. Alonso-Coello, R. Christensen and H. J. Schunemann (2013). "GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes." *J Clin Epidemiol* **66**(2): 173-183.

Hartwell, D., J. Colquitt, E. Loveman, A. J. Clegg, H. Brodin, N. Waugh, P. Royle, P. Davidson, L. Vale and L. MacKenzie (2005). "Clinical effectiveness and cost-effectiveness of immediate angioplasty for acute myocardial infarction: systematic review and economic evaluation." *Health Technol Assess* **9**(17): 1-99, iii-iv.

Hayden, J. A., M. W. van Tulder, A. Malmivaara and B. W. Koes (2005). "Exercise therapy for treatment of non-specific low back pain." *Cochrane Database Syst Rev*(3): CD000335.

Henschke, N., R. W. Ostelo, M. W. van Tulder, J. W. Vlaeyen, S. Morley, W. J. Assendelft and C. J. Main (2010). "Behavioural treatment for chronic low-back pain." *Cochrane Database Syst Rev*(7): CD002014.

Heymans, M. W., M. W. van Tulder, R. Esmail, C. Bombardier and B. W. Koes (2004). "Back schools for non-specific low-back pain." *Cochrane Database Syst Rev*(4): CD000261.

Higgins, J., J. Deeks and D. Altman (2011). "Chapter 16: Special topics in statistics. In: Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions*.

The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Higgins, J. P., D. G. Altman, P. C. Gotzsche, P. Juni, D. Moher, A. D. Oxman, J. Savovic, K. F. Schulz, L. Weeks and J. A. Sterne (2011). "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials." *BMJ* **343**: d5928.

Higgins, J. P., D. Jackson, J. K. Barrett, G. Lu, A. E. Ades and I. R. White (2012). "Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies." *Res Synth Methods* **3**(2): 98-110.

Higgins, J. P., S. G. Thompson, J. J. Deeks and D. G. Altman (2003). "Measuring inconsistency in meta-analyses." *BMJ* **327**(7414): 557-560.

Higgins, J. P., A. Whitehead and M. Simmonds (2011). "Sequential methods for random-effects meta-analysis." *Stat Med* **30**(9): 903-921.

Higgins JPT and Green S (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011], The Cochrane Collaboration.

Hoffmann, T. C., C. Eructi and P. P. Glasziou (2013). "Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials." *BMJ* **347**: f3755.

Hoffmann, T. C., S. T. Thomas, P. N. Shin and P. P. Glasziou (2014). "Cross-sectional analysis of the reporting of continuous outcome measures and clinical significance of results in randomized trials of non-pharmacological interventions." *Trials* **15**: 362.

Hopewell, S., S. Dutton, L. M. Yu, A. W. Chan and D. G. Altman (2010). "The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed." *BMJ* **340**: c723.

- Hoy, D., L. March, P. Brooks, F. Blyth, A. Woolf, C. Bain, G. Williams, E. Smith, T. Vos, J. Barendregt, C. Murray, R. Burstein and R. Buchbinder (2014). "The global burden of low back pain: estimates from the Global Burden of Disease 2010 study." Ann Rheum Dis **73**(6): 968-974.
- Hrobjartsson, A. and P. C. Gotzsche (2010). "Placebo interventions for all clinical conditions." Cochrane Database Syst Rev(1): CD003974.
- Hu, M., J. C. Cappelleri and K. K. Lan (2007). "Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes." Clin Trials **4**(4): 329-340.
- Hutton, B., G. Salanti, D. M. Caldwell, A. Chaimani, C. H. Schmid, C. Cameron, J. P. Ioannidis, S. Straus, K. Thorlund, J. P. Jansen, C. Mulrow, F. Catala-Lopez, P. C. Gotzsche, K. Dickersin, I. Boutron, D. G. Altman and D. Moher (2015). "The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations." Ann Intern Med **162**(11): 777-784.
- Imberger, G., C. Gluud, J. Boylan and J. Wetterslev (2015). "Systematic Reviews of Anesthesiologic Interventions Reported as Statistically Significant: Problems with Power, Precision, and Type 1 Error Protection." Anesth Analg **121**(6): 1611-1622.
- Imberger, G., K. Thorlund, C. Gluud and J. Wetterslev (2016). "False-positive findings in Cochrane meta-analyses with and without application of trial sequential analysis: an empirical review." BMJ Open **6**(8).
- Ioannidis, J. P. (2005). "Why most published research findings are false." PLoS Med **2**(8): e124.
- Ioannidis, J. P. (2016). "The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses." Milbank Q **94**(3): 485-514.
- Ioannidis, J. P., S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz and R. Tibshirani (2014). "Increasing value and reducing waste in research design, conduct, and analysis." Lancet **383**(9912): 166-175.
- Jakobsen, J. C., C. Gluud, P. Winkel, T. Lange and J. Wetterslev (2014). "The thresholds for statistical and clinical significance - a five-step procedure for evaluation of intervention effects in randomised clinical trials." BMC Med Res Methodol **14**: 34.
- Jakobsen, J. C., J. Wetterslev, P. Winkel, T. Lange and C. Gluud (2014). "Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods." BMC Med Res Methodol **14**: 120.
- James, A., A. Yavchitz and I. Boutron (2018). "Importance of the methods used to support the node-making process in network meta-analysis." J Clin Epidemiol.
- James, A., A. Yavchitz, P. Ravau and I. Boutron (2018). "Node-making process in network meta-analysis of nonpharmacological treatment are poorly reported." J Clin Epidemiol **97**: 95-102.
- Jin, Y., N. Sanger, I. Shams, C. Luo, H. Shahid, G. Li, M. Bhatt, L. Zielinski, B. Bantoto, M. Wang, L. P. Abbade, I. Nwosu, A. Leenus, L. Mbuagbaw, M. Maaz, Y. Chang, G. Sun, M. A. Levine, J. D. Adachi, L. Thabane and Z. Samaan (2018). "Does the medical literature remain inadequately described despite having reporting guidelines for 21 years? - A systematic review of reviews: an update." J Multidiscip Healthc **11**: 495-510.
- Junger, D. (1995). "The rhetoric of research. Embrace scientific rhetoric for its power." BMJ **311**(6996): 61.

- Kamper, S. J., A. T. Apeldoorn, A. Chiarotto, R. J. Smeets, R. W. Ostelo, J. Guzman and M. W. van Tulder (2014). "Multidisciplinary biopsychosocial rehabilitation for chronic low back pain." Cochrane Database Syst Rev(9): CD000963.
- Kessler, K. M. (2002). "The CONSORT statement: explanation and elaboration. Consolidated Standards of Reporting Trials." Ann Intern Med **136**(12): 926-927; author reply 926-927.
- Khadilkar, A., D. O. Odebiyi, L. Brosseau and G. A. Wells (2008). "Transcutaneous electrical nerve stimulation (TENS) versus placebo for chronic low-back pain." Cochrane Database Syst Rev(4): CD003008.
- Kjaergard, L. L., J. Villumsen and C. Gluud (2001). "Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses." Ann Intern Med **135**(11): 982-989.
- Koletsis, D., N. Pandis and P. S. Fleming (2014). "Sample size in orthodontic randomized controlled trials: are numbers justified?" Eur J Orthod **36**(1): 67-73.
- Kulinskaya, E. and J. Wood (2014). "Trial sequential methods for meta-analysis." Res Synth Methods **5**(3): 212-220.
- Landis, J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data." Biometrics **33**(1): 159-174.
- Lee, A., S. K. Chan and L. T. Fan (2015). "Stimulation of the wrist acupuncture point PC6 for preventing postoperative nausea and vomiting." Cochrane Database Syst Rev(11): CD003281.
- Liberati, A. (2004). "An unfinished trip through uncertainties." BMJ **328**: 531.
- Maggard, M. A., J. B. O'Connell, J. H. Liu, D. A. Etzioni and C. Y. Ko (2003). "Sample size calculations in surgery: are they done correctly?" Surgery **134**(2): 275-279.
- Maniadakis, N. and A. Gray (2000). "The economic burden of back pain in the UK." Pain **84**(1): 95-103.
- March, L., E. U. Smith, D. G. Hoy, M. J. Cross, L. Sanchez-Riera, F. Blyth, R. Buchbinder, T. Vos and A. D. Woolf (2014). "Burden of disability due to musculoskeletal (MSK) disorders." Best Pract Res Clin Rheumatol **28**(3): 353-366.
- McKeown A, G. J., McDermott MP, Pawlowski JR, Poli JJ, Rothstein D, Farrar JT, Gilron I, Katz NP, Lin AH, Rappaport BA, Rowbotham MC, Turk DC, Dworkin RH, Smith SM (2015). "Reporting of Sample Size Calculations in Analgesic Clinical Trials: ACTION Systematic Review." The Journal of Pain **16**(3 (March)): 199-206.
- Miladinovic, B., I. Hozo, A. Chaimani and B. Djulbegovic (2014). "Indirect treatment comparison." Stata Journal **14**(1): 76-86.
- Moher, D., C. S. Dulberg and G. A. Wells (1994). "Statistical power, sample size, and their reporting in randomized controlled trials." JAMA **272**(2): 122-124.
- Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger and D. G. Altman (2010). "CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials." BMJ **340**: c869.
- Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman and P. Group (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." PLoS Med **6**(7): e1000097.
- Moher, D., L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle and L. A. Stewart (2015). "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement." Syst Rev **4**: 1.

- Moher, D., L. Shamseer, M. Clarke, D. Gherzi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart and P.-P. Group (2015). "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement." Syst Rev **4**: 1.
- Moher, D., L. Stewart and P. Shekelle (2016). "Implementing PRISMA-P: recommendations for prospective authors." Syst Rev **5**: 15.
- Moher, D., J. Tetzlaff, A. C. Tricco, M. Sampson and D. G. Altman (2007). "Epidemiology and reporting characteristics of systematic reviews." PLoS Med **4**(3): e78.
- Molnar, F. J., M. Man-Son-Hing and D. Fergusson (2009). "Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia." J Am Geriatr Soc **57**(3): 536-546.
- Ostelo, R. W., R. A. Deyo, P. Stratford, G. Waddell, P. Croft, M. Von Korff, L. M. Bouter and H. C. de Vet (2008). "Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change." Spine (Phila Pa 1976) **33**(1): 90-94.
- Oxman, A. D. and G. H. Guyatt (1993). "The science of reviewing research." Ann N Y Acad Sci **703**: 125-133; discussion 133-124.
- Page, M. J., L. Shamseer, D. G. Altman, J. Tetzlaff, M. Sampson, A. C. Tricco, F. Catala-Lopez, L. Li, E. K. Reid, R. Sarkis-Onofre and D. Moher (2016). "Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study." PLoS Med **13**(5): e1002028.
- Pandis, N., P. S. Fleming, H. Worthington and G. Salanti (2015). "The Quality of the Evidence According to GRADE Is Predominantly Low or Very Low in Oral Health Systematic Reviews." PLoS One **10**(7): e0131644.
- Pocock, S. J., N. L. Geller and A. A. Tsiatis (1987). "The analysis of multiple endpoints in clinical trials." Biometrics **43**(3): 487-498.
- Pocock, S. J. and G. W. Stone (2016). "The Primary Outcome Is Positive - Is That Good Enough?" N Engl J Med **375**(10): 971-979.
- Qaseem, A., T. J. Wilt, R. M. McLean, M. A. Forcica and P. Clinical Guidelines Committee of the American College of (2017). "Noninvasive Treatments for Acute, Subacute, and Chronic Low Back Pain: A Clinical Practice Guideline From the American College of Physicians." Ann Intern Med **166**(7): 514-530.
- Rising, K., P. Bacchetti and L. Bero (2008). "Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation." PLoS Med **5**(11): e217; discussion e217.
- Riva, N., L. Puljak, L. Moja, W. Ageno, H. Schunemann, N. Magrini and A. Squizzato (2017). "Multiple overlapping systematic reviews facilitate the origin of disputes: the case of thrombolytic therapy for pulmonary embolism." J Clin Epidemiol.
- Rubinstein, S. M., C. B. Terwee, W. J. Assendelft, M. R. de Boer and M. W. van Tulder (2013). "Spinal manipulative therapy for acute low back pain: an update of the cochrane review." Spine (Phila Pa 1976) **38**(3): E158-177.
- Rubinstein, S. M., M. van Middelkoop, W. J. Assendelft, M. R. de Boer and M. W. van Tulder (2011). "Spinal manipulative therapy for chronic low-back pain." Cochrane Database Syst Rev(2): CD008112.

- Rutterford, C., M. Taljaard, S. Dixon, A. Copas and S. Eldridge (2014). "Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review." J Clin Epidemiol.
- Sackett, D. L., W. M. Rosenberg, J. A. Gray, R. B. Haynes and W. S. Richardson (1996). "Evidence based medicine: what it is and what it isn't." BMJ **312**(7023): 71-72.
- Salanti, G. (2012). "Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool." Res Synth Methods **3**(2): 80-97.
- Salanti, G., A. E. Ades and J. P. Ioannidis (2011). "Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial." J Clin Epidemiol **64**(2): 163-171.
- Santaguida, P., M. Oremus, K. Walker, L. R. Wishart, K. L. Siegel and P. Raina (2012). "Systematic reviews identify important methodological flaws in stroke rehabilitation therapy primary studies: review of reviews." J Clin Epidemiol **65**(4): 358-367.
- Saragiotto, B. T., C. G. Maher, T. P. Yamato, L. O. Costa, L. C. Menezes Costa, R. W. Ostelo and L. G. Macedo (2016). "Motor control exercise for chronic non-specific low-back pain." Cochrane Database Syst Rev **1**: CD012004.
- Schork, M. A. (2003). "Publication bias and meta analysis." J Hypertens **21**(2): 243-245.
- Schulz, K. F., I. Chalmers, R. J. Hayes and D. G. Altman (1995). "Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials." JAMA **273**(5): 408-412.
- Schulz, K. F. and D. A. Grimes (2005). "Sample size calculations in randomised trials: mandatory and mystical." Lancet **365**(9467): 1348-1353.
- Schünemann H, B. J., Guyatt G, Oxman A, editors. (2013). GRADE handbook for grading quality of evidence and strength of recommendations., The GRADE Working Group
- Schunemann, H. J. (2016). "Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision?" J Clin Epidemiol **75**: 6-15.
- Shea, B. J., J. M. Grimshaw, G. A. Wells, M. Boers, N. Andersson, C. Hamel, A. C. Porter, P. Tugwell, D. Moher and L. M. Bouter (2007). "Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews." BMC Med Res Methodol **7**: 10.
- Sim, J. and M. Lewis (2012). "The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency." J Clin Epidemiol **65**(3): 301-308.
- Simera, I., D. G. Altman, D. Moher, K. F. Schulz and J. Hoey (2008). "Guidelines for reporting health research: the EQUATOR network's survey of guideline authors." PLoS Med **5**(6): e139.
- Simmonds, M., G. Salanti, J. McKenzie, J. Elliott and N. Living Systematic Review (2017). "Living systematic reviews: 3. Statistical methods for updating meta-analyses." J Clin Epidemiol **91**: 38-46.
- Song, F., S. Parekh, L. Hooper, Y. K. Loke, J. Ryder, A. J. Sutton, C. Hing, C. S. Kwok, C. Pang and I. Harvey (2010). "Dissemination and publication of research findings: an updated review of related biases." Health Technol Assess **14**(8): iii, ix-xi, 1-193.
- Stata-IC (2017). Stata Statistical Software, v 15. College Station, TX: StataCorp LLC.
- StataCorp.2003. Stata Statistical Software: Release 8. College Station, TX: StataCorp LP.

- Thorlund K, Wetterslev J, Brok J, Imberger G and Gluud G (2011). Trial Sequential Analysis (TSA) manual. Copenhagen, Denmark.
- Thorlund, K., G. Imberger, B. C. Johnston, M. Walsh, T. Awad, L. Thabane, C. Gluud, P. J. Devereaux and J. Wetterslev (2012). "Evolution of heterogeneity (I²) estimates and their 95% confidence intervals in large meta-analyses." PLoS One **7**(7): e39471.
- Thorlund, K., G. Imberger, M. Walsh, R. Chu, C. Gluud, J. Wetterslev, G. Guyatt, P. J. Devereaux and L. Thabane (2011). "The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis--a simulation study." PLoS One **6**(10): e25491.
- Tunis, A. S., M. D. McInnes, R. Hanna and K. Esmail (2013). "Association of study quality with completeness of reporting: have completeness of reporting and quality of systematic reviews and meta-analyses in major radiology journals changed since publication of the PRISMA statement?" Radiology **269**(2): 413-426.
- Turner, E. H., A. M. Matthews, E. Linardatos, R. A. Tell and R. Rosenthal (2008). "Selective publication of antidepressant trials and its influence on apparent efficacy." N Engl J Med **358**(3): 252-260.
- Turner, R. M., S. M. Bird and J. P. Higgins (2013). "The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews." PLoS One **8**(3): e59202.
- Urrutia, G., A. K. Burton, A. Morral, X. Bonfill and G. Zanolli (2004). "Neuroreflexotherapy for non-specific low-back pain." Cochrane Database Syst Rev(2): CD003009.
- Uthman, O. A., D. A. van der Windt, J. L. Jordan, K. S. Dziedzic, E. L. Healey, G. M. Peat and N. E. Foster (2014). "Exercise for lower limb osteoarthritis: systematic review incorporating trial sequential analysis and network meta-analysis." Br J Sports Med **48**(21): 1579.
- VA/DoD (2017). Clinical practice guideline for diagnosis and treatment of low back pain. D. o. V. A. D. o. Defense.
- van der Roer, N., R. W. Ostelo, G. E. Bekkering, M. W. van Tulder and H. C. de Vet (2006). "Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain." Spine (Phila Pa 1976) **31**(5): 578-582.
- van Tulder, M., A. Becker, T. Bekkering, A. Breen, M. T. del Real, A. Hutchinson, B. Koes, E. Laerum and A. Malmivaara (2006). "Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care." Eur Spine J **15 Suppl 2**: S169-191.
- van Tulder, M., A. Malmivaara, J. Hayden and B. Koes (2007). "Statistical significance versus clinical importance: trials on exercise therapy for chronic low back pain as example." Spine (Phila Pa 1976) **32**(16): 1785-1790.
- Watson, P. F. and A. Petrie (2010). "Method agreement analysis: a review of correct methodology." Theriogenology **73**(9): 1167-1179.
- Wegner, I., I. S. Widyahening, M. W. van Tulder, S. E. Blomberg, H. C. de Vet, G. Bronfort, L. M. Bouter and G. J. van der Heijden (2013). "Traction for low-back pain with or without sciatica." Cochrane Database Syst Rev **8**: CD003010.
- Wetterslev, J., J. C. Jakobsen and C. Gluud (2017). "Trial Sequential Analysis in systematic reviews with meta-analysis." BMC Med Res Methodol **17**(1): 39.
- Wetterslev, J., C. S. Meyhoff, L. N. Jorgensen, C. Gluud, J. Lindschou and L. S. Rasmussen (2015). "The effects of high perioperative inspiratory oxygen fraction for adult surgical patients." Cochrane Database Syst Rev(6): CD008884.

- Wetterslev, J., K. Thorlund, J. Brok and C. Gluud (2008). "Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis." J Clin Epidemiol **61**(1): 64-75.
- Wetterslev, J., K. Thorlund, J. Brok and C. Gluud (2009). "Estimating required information size by quantifying diversity in random-effects model meta-analyses." BMC Med Res Methodol **9**: 86.
- White, I. (2011). "Multivariate random-effects meta-regression:updates to mvmeta." The STATA Journal **11**: 255–270.
- White, I., J. Barrett, D. Jackson and J. Higgins (2012). "Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression." Res Synth Methods **Jun;3**(2): 111-125.
- Whiting, P., J. Savovic, J. P. Higgins, D. M. Caldwell, B. C. Reeves, B. Shea, P. Davies, J. Kleijnen and R. Churchill (2016). "ROBIS: A new tool to assess risk of bias in systematic reviews was developed." J Clin Epidemiol **69**: 225-234.
- Wright, J. G. (1996). "The minimal important difference: who's to say what is important?" J Clin Epidemiol **49**(11): 1221-1222.
- Yousefi-Nooraie, R., E. Schonstein, K. Heidari, A. Rashidian, V. Pennick, M. Akbari-Kamrani, S. Irani, B. Shakiba, S. A. Mortaz Hejri, S. O. Mortaz Hejri and A. Jonaidi (2008). "Low level laser therapy for nonspecific low-back pain." Cochrane Database Syst Rev(2): CD005107.
- Zanoli, G. (2005). "Outcome assessment in lumbar spine surgery." Acta Orthop Suppl **76**(318): 5-47.