Dipartimento di
Medicina e Chirurgia

Dottorato di Ricerca in Sanità Pubblica Ciclo XXIX

Curriculum in Health Technology Assessment

Evidence Synthesis in Mechanical Low Back Pain Rehabilitation Interventions: translate research into practice

Surname: Gianola   Name: Silvia Eleonora

Matricola 787763

Tutor: Dr. Anita Andreano

Co-tutor: Dr. Lorenzo Moja

Supervisor: Prof. Maria Grazia Valsecchi

Coordinator: Prof. Guido Grassi

ANNO ACCADEMICO 2017-2018

Table of Contents

Acknowledgements

List of Figures

List of Tables

List of Appendix

List of abbreviations

COMET, Core Outcome Measures in Effectiveness Trials
CI, Confidence Intervals
CONSORT, Consolidated Standards of Reporting Trials
COS, Core Outcome Set
EQUATOR, Enhancing the QUAlity and Transparency Of health Research
ES, Effect Size
FU, Follow-Up
HRQL, Health-Related Quality of Life
IQR, Inter Quartile Range (I and III quartiles)
LBP, Low Back Pain;
KT, Knowledge Translation
MA, Meta-analysis
MID, Minimal Important Difference
MD, Mean Difference
MBR, Multidisciplinary Biopsychosocial Rehabilitation
NMA, Network Meta-Analysis
NSAID, Non-Steroid Anti-Inflammatory Drug
NPT, Non-Pharmacological Therapies
NRS, Numerical Rating Scale
ODI, Oswestry Disability Index
OR, Odds Ratio
PROs, Patients Reported Outcomes
PRISMA, Preferred Reporting Items for Systematic Review and Meta-Analyses
RCT, Randomized Controlled Trials
RMQ, Roland and Morris Questionnaire
ROB, Risk of Bias Assessment
SPIRIT, Standard Protocol Items: Recommendations for Interventional Trials
SD, Standard Deviation
SDR, Standard Deviation Ratio
SEM, Standard Error Measurement
SF-36, Medical Outcomes Study 36-Item Short Form Health Survey
SMD, Standardized Mean Difference
SR, Systematic Review
SUCRA, Surface Under the Cumulative Ranking Curve
TIDieR, Template for Intervention Description and Replication
VAS, Visual Analogue Scale

EXECUTIVE SUMMARY

INTRODUCTION

Mechanical low-back pain (LBP) is the musculoskeletal disorder with the highest prevalence in adults (9.4%, 95% Confidence Interval [CI] 9.0 to 9.8). Consequently, it has a high economic and social burden. There are many different therapeutic interventions for mechanical LBP, but none of them is universally accepted. In order to determine which treatment is more effective according to available scientific evidence, we primarily had to critical appraise the quality of the study design of available evidence in this field. The best study design to assess the efficacy of an intervention is a randomized clinical trial (RCT). Anyway, critical evaluation of the reported efficacy depends on several dimensions of conduct and interpretation of trials. First of all, the accurate reporting of population and sample size, intervention and comparison, and outcomes. The description of these dimensions guarantees the validity and generalizability of the effectiveness of an intervention. Secondarily, we need to consider the parameter adopted to declare a treatment as having a beneficial effect. Usually findings in trials are interpreted in terms of statistical significance, anyway results can be translated in terms of clinical relevance. After a careful appraisal of the limits of study design in rehabilitation field and having expressed efficacy in clinically meaningful units, a meta-analysis (MA) of multiple interventions will allow to determine the best treatment among different options on the basis of evidence, and not only according to expert's opinion.

AIMS

The aim of this dissertation was to evaluate mechanical LBP rehabilitation interventions based on RCTs. In particular, the quality of study design was evaluated trough systematic assessment of completeness in reporting of interventions, outcome and sample size dimensions (aim of Part I). Then, the reporting in terms of clinically meaningful effects was assessed in RCTs and in MA of RCTs, which are considered the gold standard to disseminate the synthesis of evidence (aim of Part II). Finally, in order to determine the most efficacious treatment for acute LBP, a network meta-analysis (NMA) comparing multiple interventions was performed.

METHODS

To assess reporting of different components of study design and assess the clinical relevance in RCTs, a systematic search for all RCTs included in Cochrane systematic reviews (SRs) on LBP published in *The Cochrane Database of Systematic Reviews* was performed. The description of sample size, interventions, outcomes, and clinical relevance of each RCT was evaluated, independently by two of the investigators, using dedicate extraction forms.

To assess the clinical relevance in MA, a Cochrane review focusing on multidisciplinary biopsychosocial rehabilitation (MBR) in short, medium and long terms was selected as a case-study: we re-analysed the data using a MID (minimal important difference) units approach and discussed the implications of this approach compared to the traditional one. Results were expressed in MID units and gave a clinical meaningful interpretation.

To estimate the best efficacious intervention, a NMA was implemented using random-effects models within a frequentist setting assuming equal heterogeneity across all comparisons and all ranking probabilities for each treatment outcome in acute LBP interventions as case study were estimated.

RESULTS

To appraise reporting quality, 185 eligible RCTs from all Cochrane SRs focused on LBP rehabilitation interventions were found. All items necessary for a full replication of the intervention were present only in 33 RCTs (17.8%). Thirty-six different outcomes were investigated across all RCTs. The 2 most commonly reported outcomes were pain (n= 165 RCTs; 89.2%) and disability (n= 118 RCTs; 63.8%). Pain and disability outcomes were found replicable in only 10.3% (n= 17) and 10.2% (n= 12) of the RCTs, respectively. Of the 80 RCTs reporting sample size calculation, only thirteen (16.3%) gave an adequate description of the a priori sample size calculation, with all elements provided in compliance with the Consolidated Standards of Reporting Trials (CONSORT) guidelines. We then identified 42 RCTs (41,6%) having both a sample size calculation and a planned MID. Overall, we found that more than one-third of RCTs (37,5%, n=15) were both statistically and clinically significant whereas few (23,8 %, n=10) were statistically significant but not clinically relevant.

For the clinical interpretation of MAs we used the MID units approach. Pooling standardized mean differences (SMDs), the 95%CI of the summary estimate did not include the zero in all three MAs, indicating a statistically significant effect in favour of MBR over usual care in terms of pain relief for LBP patients in short, medium and long terms. However, in terms of clinical relevance, MBR improved back pain in an appreciable number of patients only at short term (MID lower than but close to 1) whereas in longer time MBR showed little or no effect for the majority of patients (MID close to 0).

For multiple comparisons, a NMA was performed on acute and subacute LBP discovering that the best efficacious treatments for pain at short term of follow-up (FU) are muscle relaxant drugs (34.5% probability to be the first treatment) and manual therapy (18.8% probability to be the first treatment). Inconsistency was not found in the network.

CONCLUSIONS

Despite the remarkable amount of resources spent performing RCTs in the LBP rehabilitation field, the majority of RCTs failed to report sufficient information for sample size, interventions, outcomes and clinical relevance. Moreover, almost a quarter of trials are statistically significant but not clinically significant. Improving the quality of reporting and introducing interpretation of effects in terms of clinical relevance can increase validity and efficacy of research findings, promoting the knowledge translation (KT) of valid results into rehabilitation practices. Only in this light NMA can be used for clinical decision making, based on a strong evidence and a useful tool, for all stakeholders of LBP condition.

BACKGROUND AND RATIONALE

Low back pain

Non-specific LBP is a symptom, not a disease [1]. Various spinal structures, including ligaments, facet joints, paravertebral musculature and fascia, intervertebral discs, and spinal nerve roots, have been implicated as pain generators [2]. Nevertheless, 85% of patients with isolated back pain still do not have a definitive identified cause for their symptoms [3]. The aetiologies can be subdivided into mechanical, systemic, and referred. By far, the most frequent cause is mechanical (97%) [2] with the most common form of "non-specific LBP" [4]. This definition is used when the anatomo-physio-pathological cause of the pain cannot be precisely determined [1], and it is based on the exclusion of patients with a specific cause (e.g., fracture, infection, cancer) [4]. LBP is commonly defined as pain or discomfort localized in the area of the posterior aspect of the body, from the lower margin of the twelfth ribs to the lower gluteal folds, with or without pain referred into one or both lower limbs, that lasts for at least one day [5].

Non-specific LBP may be classified by duration as acute (pain lasting less than 6 weeks), sub-chronic (6 to 12 weeks), or chronic (more than 12 weeks) [6]. Acute LBP is one of the most common reasons for adults to see a general practitioner [7], experiencing moderate to severe pain, and being debilitated in motor and psychological functions [7].

It is associated with high disability and costs for the society [8]. Out of the 291 pathological conditions studied, LBP ranked highest in terms of disability, and sixth in terms of overall burden expressed as disability adjusted life-year (DALYs). The worldwide point prevalence of LBP in 2010was 9.4% (95%CI, 9.0-9.8). DALYs increased from 58.2 million (95% CI, 39.9-78.1 million) in 1990 to 83.0 (95%CI, 56.6-111.9 million) in 2010. Prevalence and burden increase with age [5]. Despite its widespread prevalence, acute LBP is considered to be typically self-limiting, with a recovery rate of 90% within 6 weeks of the initial episode [9], whereas 2% to 7% of patients develop chronic LBP [4]. However, its chronicity is associated with considerable disability and costs for the

society [8]. In fact, chronic evolution of LBP is often considered as a biopsychosocial problem, as it is characterised by a combination of physical, psychological and social dysfunctions [10]. Those manifestations are typically patient reported and have a subjective nature.

Gaps in LBP interventions

In order to determine which is the best intervention for mechanical and aspecific LBP, a corrected methodology has to be established in order to assess the effectiveness of interventions for this condition. In fact there are several limitations in many studies evaluating LBP interventions (pharmacological as well as rehabilitation treatments). In rehabilitation field, most of studies are empiric or are based on clinical observation with a trivial sample size. On the contrary, evaluation of therapeutic interventions should be based on well-designed, adequately powered, and properly conducted RCTs, which are the most reliable design study used to develop useable forms of evidence in KT. They can also be the basis to inform end users about the evidence and promote change in practice [11]. Clinical trials seek to evaluate whether an intervention is more effective than a comparator [12]. However the measured effectiveness of an intervention depends on several dimensions, which need to be carefully considered by researchers when planning or interpreting the design of a trial: patients and sample size, definition of intervention and comparison, outcome definition and measurement. Proper specification of these dimensions guarantee the validity and generalizability of an intervention. If authors do not provide sufficient details concerning the conduct of their study, readers cannot judge the validity and generalizability of the results [13]. As a consequence, invalid study results may be used by health care professionals, causing harm to patients and shifting resources on ineffective treatments [14]. Moreover, incomplete reporting of the applied methodology limits the reproducibility of the study results in clinical practice and prevents effective dissemination of a new efficacious procedure [14].

Secondarily, effects of rehabilitation interventions are substantially smaller than those of other interventions, such as surgical or pharmacological ones, that reduce substantially pain and/or

disability. Small effects are more difficult to detect and require larger sample sizes for clinical studies, making them more difficult to carry out [15]. Once a study is well conducted and adequately reported, the interpretation of the effectiveness of an intervention depends on the cut-off adopted to consider a treatment successful or not. It is important to consider not only the statistical significance but also the clinical relevance, defined as a clinically important and meaningful change. In fact, an observed statistically significant difference between two interventions does not necessarily imply that this difference is clinically important [16] or that changes were clinically relevant for patients [17]. This assumes a huge importance in rehabilitation, because outcomes are measured by self-reported scale or questionnaires rather than being assessed by the researcher. Patient reported outcomes (PROs) include pain, disability, health-related quality of life (HRQL), presence and intensity of symptoms, and satisfaction ratings.

Finally, the increase in alternative medical treatment options has led to the need for comparative effectiveness research [18]. Trials comparing many treatment options are usually not feasible, so other methodological approaches are needed. A MA of studies included in a SR is a useful statistical tool that provides a summary estimate of treatment effect combining data from many studies. However, a key limitation of pairwise (standard) MAs is that they can compare only 2 interventions at a time. When several treatment options are available, a series of individual MAs provides only partial information, because it can tell us only which of two treatment is more effective. This does not support optimal clinical decision making [19]. The NMA, also called multiple treatments meta-analysis or mixed-treatment comparison, has been developed to assess the relative effectiveness of several interventions [20]. The method is based on the simultaneous analysis of direct evidence (which comes from studies directly randomizing treatments of interest) and indirect evidence (which comes from studies comparing treatments of interest with a common comparator) [21].

Organization of the project

The PhD research project and the thesis were organized around the following three main objectives:

In Part 1, the reporting of all essential elements of a RCT evaluating rehabilitation intervention of mechanical LBP was investigated, in terms of sample size, information necessary for the replication of an intervention and assessment of the outcome.

In Part 2, the clinical relevance of the effects found in RCTs and MAs evaluating rehabilitation intervention for LBP was assessed.

In Part 3 a NMA of intervention for acute and subacute mechanical LBP was performed.

A general discussion and conclusion follows in the last section.

PART 1. Evaluating study design through reporting

1. Introduction

Improving quality of research through reporting guidelines

Scientific publications represent the most important output of research and they are the best channel of communication among researchers. The quality and relevance of biomedical studies can be judged almost only on the basis of what is reported in the publication [14]. In fact, we can verify how a study was planned only when the protocol is registered before the study begins, and it is almost impossible to verify how a study was conducted in practice. However, there is evidence that inadequate reporting is associated with poorly conducted research [14]. Consequently, evaluation of reporting can be considered a proxy for evaluation of research quality.

Lack of complete and adequate reporting has many drawbacks. First, if authors do not provide sufficient details concerning the conduct of their study, readers cannot judge the validity and generalizability of the results [13]. As a consequence, invalid study results may be used by patients and health care providers: patients may be harmed and resources shifted to ineffective treatments [14]. Secondly, incomplete reporting of methodology limits the reproducibility of the study in clinical practice and prevents effective dissemination of a new efficacious procedure [14].

There is evidence that several published medical research are poorly reported [22]. This is pervasive to almost every area of health research [13] and rehabilitation is not an exception [23].

Initiatives to promote reporting guidelines, such as the Enhancing the QUAlity and Transparency Of health Research (EQUATOR), aimed at improving the clarity and transparency of reporting in health research. To reach this goal, they developed resources and training for robust reporting guidelines [13].

The first attempt was the CONSORT Statement, aimed at improving the reporting of RCTs [24]. Since its development in 1996, several other guidelines have been proposed for other types of research studies. Examples include QUOROM (Quality of Reporting of Meta-analyses) for MAs

of RCTs [25], recently replaced by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [26], STARD (Standards for Reporting of Diagnostic Accuracy Studies) [27], and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [28].

The most reliable study design to test the efficacy of an healthcare intervention is a well-designed and properly executed RCT. This is true to the point that Rennie affirmed that "The whole of medicine depends on the transparent reporting of clinical trials" [29]. The RCT is a form of KT used to increase awareness of a problem, develop useable/ actionable forms of evidence, inform end users about the evidence, and promote change in practice [11]. However, trials presenting an inadequate reporting of their methodology are associated with biases and overestimation of the treatment effect [30]. The critical appraisal of the quality of an RCT is possible only if its design, conduct, and analysis are thoroughly and accurately described in the report. DerSimonian and colleagues suggested that "editors could greatly improve the reporting of clinical trials by providing authors with a list of items that they expected to be strictly reported" [31].

In the 1990s, two groups of journal editors, trialists, and methodologists independently published recommendations on the reporting of RCTs [32,33]. Subsequently, the two groups developed a common set of recommendations [34] resulting in the CONSORT statement, firstly published in 1996 [35]. The CONSORT included 21 checklist items pertaining to the rationale, design, analysis, and interpretation of a trial, and a flow diagram outlining the progress of participants involved. The aim was to standardize the reporting of RCTs, improving completeness and transparency, and aiding the critical appraisal and interpretation of trials results.

In 2001, the CONSORT checklist was updated to 22 items to increase the ability to judge the validity or relevance of trial findings [24]. The second revision of the CONSORT Statement (CONSORT 2010) was published in March 2010 [36] with an updated checklist of 25-items [37].

To date, over 600 journals have endorsed the CONSORT Statement and many of them suggest or require the adoption of the CONSORT checklist and flow-chart for the publication of a clinical trial [38].

Finally, extensions of the CONSORT Statement, have been developed in order to have additional guidance when dealing with specific designs (e.g., Non-Inferiority and Equivalence Trials), data (i.e., Harms) and interventions (i.e., Non-Pharmacological Therapies [NPT]). In fact, clinical trials evaluate the effectiveness of an intervention [12], but this effectiveness depends on the included population, the characteristics of the intervention, the performed comparison, and the chosen outcome measure. All these dimensions need to be carefully defined in planning and interpreting a clinical trial.

These are the reasons why the focus of part I of the dissertation was the evaluation of how sample size calculation (Chapter 2), interventions (Chapter 3) and outcomes assessment (Chapter 4) are reported in RCTs on interventions for LBP.

The reporting of the sample size calculation

Sample size is related to statistical power, which is the complement to one of type II error [39,40]: it represents the likelihood of failure to reject the null hypothesis when, in fact, it should be rejected. The investigator's aim is to control for this type of error by defining an adequate sample size. Sample size calculation is essential in study design because a low-powered study may fail to yield significant results when the treatment as a relevant clinical effect. It needs to be described in any published report, so that readers can judge if the assumptions made in the sample size calculation were realistic. Although the number of published RCTs in rehabilitation has been increasing [41], the majority of those studies have a sample sizes that is too small to reject the null hypothesis when it is false with a high enough probability [42].

In order 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. At a minimum they have to report: the expected difference in outcomes between groups (minimum important treatment effect or effect size [ES]), the level of significance (type I error), the statistical power (1- type II error), and the estimated outcome

variability [43-45]. In addition, the CONSORT guidelines also recommend reporting the primary outcome on which important differences between two groups are determined. Authors should therefore decide and state a priori the fixed values for parameter assumptions.

The reporting of description intervention

In primary care and general medicine, a study including 80 RCTs and SRs selected by the journal "Evidence-Based Medicine" for their relevance and newsworthiness, showed that 51% of the articles had an incomplete description of the treatment and information was better reported for pharmacological treatments than for NPT [46]. NPTs include surgery, technical procedures, devices, psychotherapy, behavioural interventions, complementary, alternative medicine and rehabilitation. For NPTs a detailed description of the intervention is essential, more than for pharmacologic therapies, to allow its replication in the practice. Another study based on 51 RCTs found that 57% of the interventions published in the British Medical Journal, a general medical journal, could not be replicated based on the description of the treatment as published [47]. Pharmaceutical studies were better described than those on non-drug treatments, with 33% (7/21) of drug trials and 73% (22/30) of non-drug trials deemed not replicable. Like others NPTs, rehabilitation interventions are often not adequately reported [48].Trials for NPTs usually test complex interventions involving several components [49] that are difficult to standardize, reproduce, and administer consistently to all patients. All of these variations could impact on the treatment effect. In addition, care providers' expertise, patient confidence with care provider, and volume of the centres can also influence the estimate of the treatment effect for NPTs to a greater extent than for drugs [50].

In order to answer to specific issues in the assessment of NPTs, the extension for the CONSORT Statement for "Non-Pharmacological Therapies" was introduced in 2008 [48]: the CONSORT flow diagram was modified to include data on the number of care providers, centers in each group and the number of patients treated by each care provider. However, a comprehensive checklist for the description of all aspect of the intervention was presented only in 2012 [47] and recently more

detailed reporting guidelines have been introduced by the TIDieR (Template for Intervention Description and Replication) Checklist [51].

The reporting of the outcome assessment

The chosen type of primary outcome measure influences both the size of the clinically important difference attributable to the intervention and the definition of "successful intervention" in a RCT [52]. Success is not only the demonstration of the statistically significant difference in terms of the main outcome between treated and not treated but also the evaluation of its clinical importance. Furthermore, the chosen outcome will influence the required sample size of the trial [53] and the extension of the FU needed to cumulate a sufficient number of events. Finally the choice of the outcome has been related to a peculiar type of bias: the selective outcome reporting bias, i.e. the outcome measure that results to be statistically significant is the one reported. In other words, in the presence of selective outcome reporting bias, published results are prone to the 'statistically significant' cliché: new statistically significant outcomes not included in the design are introduced at the time of publication; statistically significant secondary outcomes are upgraded to primary; and non-significant primary outcomes are possibly omitted from reports [54]. Consequently, the selection of an inappropriate outcome measure can distort the results [52]. The best designed and most rigorously executed trials cannot make up for a poorly chosen outcome measure [52].

A variety of outcome measures have been used in studies of rehabilitation interventions for LBP management [55]. Even when studies are concordant in demonstrating benefit for a given treatment, the use of different outcome measures has made challenging to capture the magnitude of the treatment effects across various studies [56], preventing patients and clinicians to choose the most effective therapeutic option.

In rehabilitation, as well as for others NPT, outcomes are often measured by self-reported scale or questionnaires rather than being observer reported. PROs are mainly continuous variables and include pain, disability, HRQL, presence and intensity of symptoms, and satisfaction ratings.

PRO data from RCTs are increasingly used to inform patient-centred care, clinical decision making, and health policy [57]. However, there is lacks of guidance on how to report PROs, which are often inadequately described in trials, thus limiting the value of these studies. In order to avoid outcome reporting bias, a better reporting of PRO is needed. An evidence-based extension of the CONSORT statement for PRO in RCTs was developed, the CONSORT PRO [57]. It recommends the identification of primary and secondary outcomes in the abstract, and the description of the scientific rationale of a specific patient reported outcome and its relevant domains (ie, if a multidimensional PRO tool has been used). It also prescribe to validate or to cite a previous work demonstrating the validity and reliability of the PRO instrument, to explicit the statistical approaches for missing data, and to discuss the specific limitations, generalizability and use in clinical practice of the selected PRO.

Open questions and aims in LBP rehabilitation interventions

Patients and sample size, intervention and comparison, outcomes are important dimensions that guarantee the validity and generalizability of an intervention. In fact, incomplete reporting of methodology limits the reproducibility of the study in clinical practice and prevents effective dissemination of a new efficacious procedure [14]. For our knowledge nobody has investigate yet the reporting of sample size, interventions, outcome in the rehabilitation field.

2. Improving Power and Sample Size calculation in Rehabilitation Trial Reports

The purpose was to systematically assess the quality of reporting of power and sample size calculation in RCTs comparing mechanical LBP rehabilitation interventions and included in Cochrane SRs.

Methods

Search strategy, eligibility criteria and study selection

We searched all Cochrane SRs published up to December 2013 in the Cochrane database, using the terms 'back pain' and 'rehabilitation' in adult treatments. We focused on Cochrane SRs because they represent a gold standard for identifying all relevant RCTs in a field [57,58] through highly sensitive search strategies.

A Cochrane SR was included if mechanical LBP was the target disease and rehabilitation was the intervention. Rehabilitation included all forms of therapeutic interventions defined by the National Library of Medicine as the "restoration of human functions to the maximum degree possible in a person or persons suffering from disease or injury" [59] delivered by health professionals of rehabilitation. SRs focusing on interventions other than therapeutic rehabilitation (e.g., prevention) or based on population subgroups (e.g., pregnancy) were excluded.

From eligible SRs, we extracted all RCTs published in English, Italian, Spanish, or French. Three authors independently screened the SRs (title and abstract) for eligibility and subsequently reviewed all identified RCTs. Disagreements were resolved by negotiation among the authors.

Data Extraction

The following general characteristics were collected from the included RCTs: year of publication, number of authors, first author's geographic region (Europe, North and South America, Asia and Australia) and journal that published the study, and funding source. A detailed extraction form derived from the CONSORT checklist was developed to extract data on sample size calculation. The checklist was upload on Distiller SR [60], a web-based database for data management.

It was assessed whether the RCT report included a power analysis in the Methods section and, if so, whether the description of the sample size calculation was CONSORT-compliant. Following the CONSORT checklist [44], the description for reporting of six sample size calculation components was assessed: (1) type I error (alpha), (2) type II error (beta) or power (1- beta), (3) assumption of expected treatment effect of the intervention (i.e., the difference between group means as ES or MID and relative risk), (4) the assumed variability expressed as a standard deviation (SD) or a variance or an interclass correlation coefficient., (5) the outcome on which sample size calculation was based, and (6) whether there was a correction to allow for losses to FU. In addition, the sample size planned and the actual sample size randomized (N) according to the CONSORT flow diagram were extracted. If there was no statement or CONSORT flow diagram reporting the number of patients randomized, the data was extracted from implicit information (i.e., "enrolled" or "included"). When articles reported the sample size calculation, a discrepancy between the planned sample size and the number of participants randomized was examined. Moreover, funding status of each RCT was extracted in order to study the influence of this on sample size reporting. Data extraction was independently performed by two reviewers.

Statistical Methods

Descriptive statistics are presented as medians and I and III quartiles of IQR (Inter Quartile Range), or percentages as appropriate. The non-parametric matched-pairs Wilcoxon signed-rank test was used to compare the planned versus the actual sample sizes; univariable logistic regression was

employed to examine the association of sample size calculation reporting with funding status, and the trend for improvement in reporting over time. For hypothesis testing, a probability level lower than 0.05 was considered to be statistically significant. All statistical tests were two-sided. Stata software was used for all statistical analyses Stata-IC [61].

Results and limitations

Study selection

Fourteen relevant Cochrane SRs in the Cochrane Library were identified [62-75]. Sixty out of 301 RCTs included in these 14 SRs were excluded because they were duplicates or multiple publications of the same RCT, 7 were excluded as their full text could not be retrieved, and 12 were excluded because they did not satisfy the language criterion. A final total of 222 RCTs was included in our review, Appendix 1.

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 the country of publication, with the three top 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 (60%, n=132) by European researchers. The period of RCTs publication was from 1968 to 2013. The characteristics of the RCTs are reported in Table 1.

Table 1. General characteristics of RCTs for the reporting of sample size calculation in rehabilitation interventions for mechanical LBP.

	Frequency (No.)	(%)
No. of countries	32	
USA	42	18.9
UK	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		
Spine	50	22.5
Journal of Manipulative and Physiological Therapeutics	10	4.5
Pain; British Medical Journal; Archives of Physical Medicine and Rehabilitation	9	4.1
Clinical Journal of Pain	8	3.6
No funding reported, no. (%)	97	43.7
	median	**IQR**
No. of authors, median (IQR)	5	1-12
Year of publication of trial report, median (IQR)	2000	1968-2013

Sample size calculation

Reporting

Only 80 (36%) of the 222 RCTs reported a sample size calculation. We found a significant positive trend for improvement over time (p<0.001). Beginning in 2005, the majority of trials reported a sample size estimation, figure 1. Furthermore, we found that trials that reported a funding source were four times more likely to report sample size calculation than trials that reported not having received support (odds ratio [OR] 3.91, 95%CI 2.12–7.22; p<0.001).
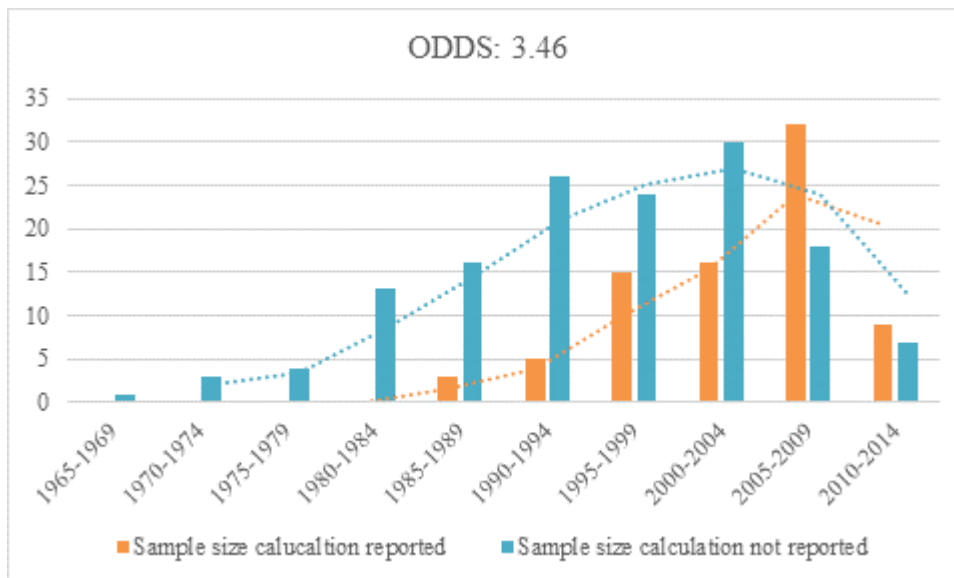
Figure 1. Reporting of sample size over time in rehabilitation interventions of LBP.

Complete description of sample size calculation

Thirteen (16%) of the 80 RCTs reporting sample size calculation gave an adequate description of the a priori sample size calculation, with all six elements provided in compliance with CONSORT guidelines. Half of the RCTs reported at least four out of six elements as reported in figure 2.
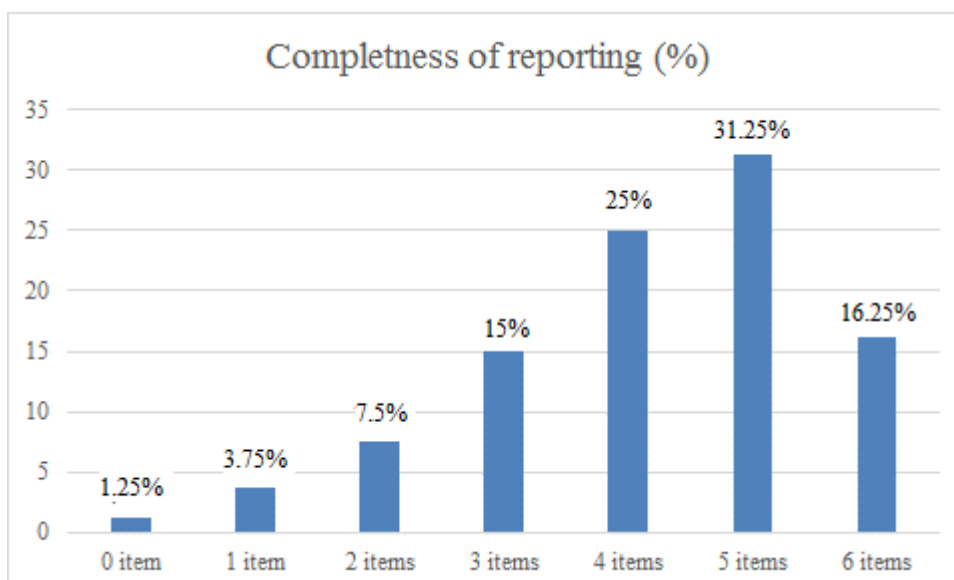
Figure 2. Completeness of reporting items for sample size calculation in rehabilitation interventions for mechanical LBP.

Of the six CONSORT components required for sample size calculation, the three most frequently reported were the power (91.25%, n=73), followed by the assumption concerning the expected treatment effect of the intervention (86.25%, n=69), and the alpha error or type I error (85%, n=68). Correction for losses to FU was the least frequently reported element (32.5%, n=26), see figure 3.
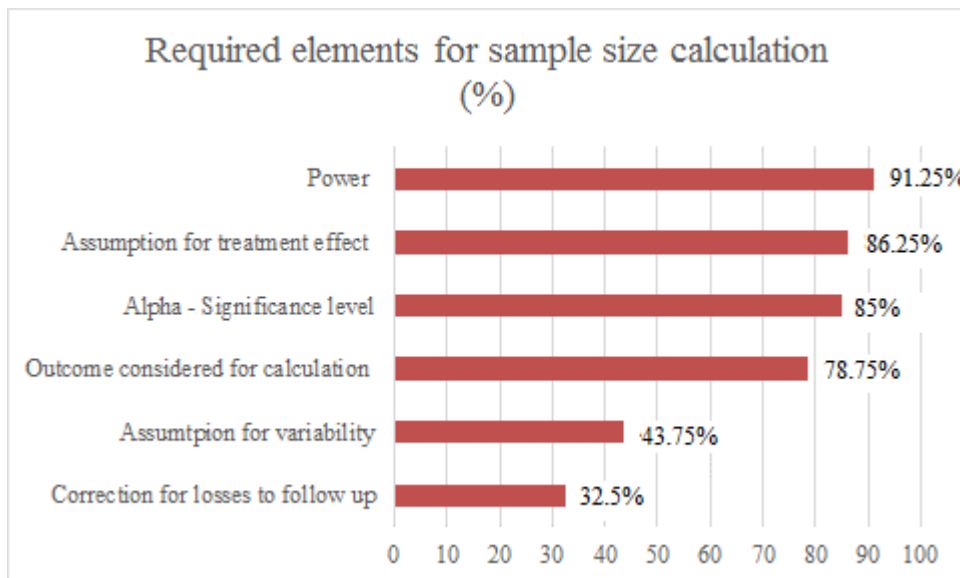
Figure 3. Number of trials reporting of required elements for sample size calculation in rehabilitation interventions for mechanical LBP.

Characteristics of each element reported

Each element could be expressed in a different way; common expressions for elements are presented in Table 2. Power was usually defined as $1 - \beta$ (82.5%, n=66). The MID was the assumed value for the detection of treatment effect most often reported in the 80 trials (46.25%, 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 of the sample size calculation in rehabilitation interventions for mechanical LBP.

Sample size calculation elements	No. (%)
Level of significance	
Alpha (type I error)	68 (85)
Power	
Beta (type II error)	10 (12.5)
Power (1 – Beta)	66 (82.5)
Total	73 (91.25)
Assumption for treatment effect	
MID	37 (46.25)
ES	9 (11.25)
Other (i.e., reduction in %)	24 (30)
Total	69 (86.25)
Assumption for variability	
SD	28 (35)
Other (i.e., variance)	7 (8.75)
Total	35 (43.75)
Correction for losses to FU	26 (32.5)
Outcome considered for sample calculation	
Disability	34 (42.5)
Pain	18 (22.5)
Other (i.e., recovery rate, work days)	19 (23.75
Total	63 (78.75)

Discrepancy between planned and randomized sample size

Planned sample size was reported in 74 RCTs (33.3%); in the remaining six RCTs the sample size calculation was reported but not the size of the sample. The median number of participants needed to prove sufficient power was 120 (range: 17–2000), whereas the median of the actual number of participants randomized among these 74 RCTs was 133 (IQR 15–741). The actual number of participants was lower than the number of those planned in 19 RCTs (25.67%), equal in 13 (17.56%), and higher in 49 (66.21%).

Limitations

This study focused only on the reporting of sample size calculation and its components as described in the Methods section of RCTs. It would have been interesting to compare the final publication with the published protocol in order 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.

3. Completeness of outcomes description in low-back pain rehabilitation interventions

The aim was to evaluate the completeness of reporting of outcomes that are most commonly used in RCTs examining interventions for LBP. To pursue this goal, we first determined the type and frequency of outcomes used. The completeness of the reporting of the four most commonly used outcomes was then determined and examined for a relationship with the year of publication of the trial: we hypothesized that outcomes would be reported more thoroughly in recently published trials supported by various initiatives promoting reporting, such as the CONSORT PRO and SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) [76,57,48]. Finally, we examined the completeness of description of the blinding of the outcome assessment.

Methods

Registered protocol

The study protocol was registered in the COMET database [77] in agreement with the COMET (Core Outcome Measures in Effectiveness Trials) Initiative [78].

Search strategy, eligibility criteria and study selection

Search strategy and study selection process are identical to those previously described in chapter 2 (page 26). However, this search strategy was run up to May 2013.

Data collection and Definitions

An outcome extraction form was designed, which was refined after the first 60 trials based on the problems identified. DistillerSR, a web-based password-protected database for data extraction [60] was used. Six pairs of independent researchers trained in SR methodology extracted study characteristics such as: information concerning the study population, intervention, control, sample size, number of reported outcomes and their assessment, and funding from the included RCT full-texts. Whether or not each RCT distinguished between primary and secondary outcomes was also recorded.

The primary outcome was defined as adequately reported when only one outcome, even if composite, was indicated as primary in the methods section or had been used in the sample size calculation. If the primary outcome was not clearly indicated (i.e. more than one outcome defined as primary, no indication that the outcome was used to calculate sample size in the presence of multiple outcomes), it was considered as not adequately planned.

After determining the four most frequently reported outcomes, the completeness of their reporting was assessed with an 8-item-checklist developed for this project; the items are represented in figure 4. The items were selected from established opinion of what aspects of the methodology should be reported [57,79,52,80-82]. The methods and results section of each trial were reviewed for judge if each of the eight items was reported or not reported. An outcome was considered to be fully reported if all of the items were presented. Changes in the completeness of reporting for each outcome over time were analysed. Finally, since blinding is one of the most important procedures to protect against bias in an RCT [83], its reporting was investigated. The frequency of blinding used across all included RCTs was determined. A trial was considered as blinded, unblinded, or unclear based on the information provided in the article [83]. When blinding was reported, the level was extracted: participants, trials investigators, assessors and data analysis.
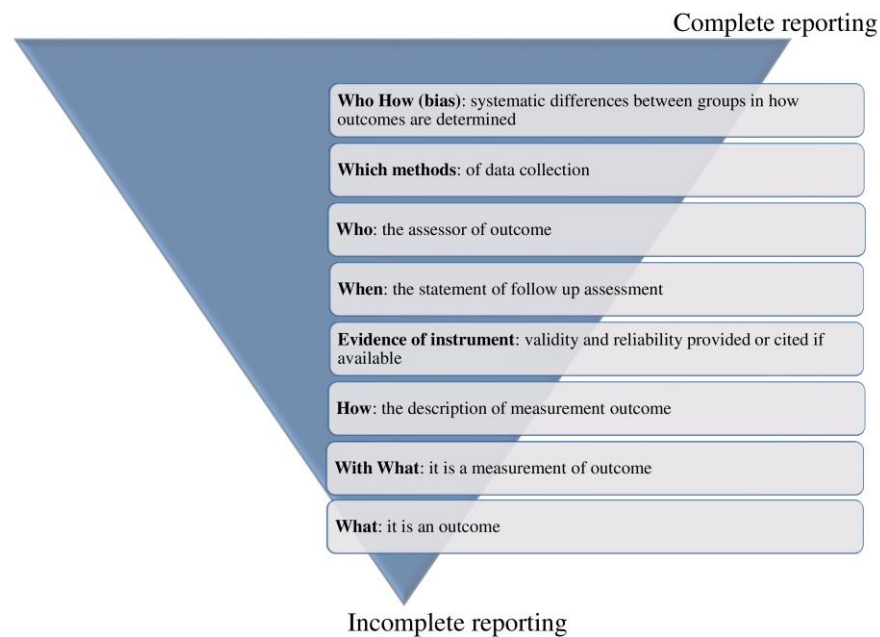
Figure 4. Eight aspects of outcome assessment necessary for its replication have been selected and used to evaluate the level of completeness of its reporting.

Items to be reported can be permutated from one level to the other.

Legend:
1- **What:** e.g., pain;
2- **With what:** e.g., VAS (Visual Analogue Scale) and 5 points scale are measurement instruments of the same outcome, the pain;
3- **How:** e.g., VAS was considered from 0 to 100;
4- **Evidence of instrument**: e.g., VAS (Boonstra AM et al., Int J Rehabil Res. 2008);
5- **When:** e.g., VAS was administrated before intervention and at 10 week of training exercise;
6- **Who:** e.g., Physical therapist conducted the outcomes assessment;
7-**Which methods**: e.g., paper, telephone, electronic, other;
8- **Who How (bias)**: e.g., blinding of outcome assessment.

Statistical Analysis

Completeness of reporting for the four most frequent outcomes was described, for every item in the checklist, by the proportion of RCTs adequately reporting the item. For every outcome, univariate logistic regression models were used to investigate the impact on each item (dependent binary variable) of publication year (continuous independent variable). The proper functional form of year was modelled using polynomial terms. For items with a significant quadratic term – representing a decreasing and then increasing proportion of adequately reported RCTs with

publication year –the linear effect of publication year for the most recent time period was estimated, fitting a new model, including just the linear term, only on the studies published after the curvature point. Results of the logistic regressions are presented graphically and as 10-year OR (i.e. relative increase or decrease of the probability that a study will report the item for any 10-year increment of publication year), and their corresponding 95%CIs. All tests were performed 2-sided with a significance level of 0.05. All analyses were performed with R software [84].

Results and limitations

Studies selection

After screening, from 11 Cochrane SRs [85,86,64,87-92,62,93] 220 RCTs were included. Excluding duplicates, not retrieved and non-English or French or Spanish articles, 185 RCTs constituted the study sample. The flow chart of study selection is present in Appendix 1.

How many outcomes and measurements are present in the published RCTs?

Overall thirty-six outcomes were reported more than one time but more than other 100 outcomes were reported only once time by the trialists. The outcomes most commonly reported were pain (89.2%), disability (63.8%), range of motion (38.9%), and quality of life (24.3%) (Table 3) measured respectively by 70, 43, 41 and 19 different measurement instruments. Disability and quality of life were assessed through self-reported measures, range of motion always by clinical assessment and pain with both type of measurements (e.g., the pressure pain thresholds were measured with a commercial device and the pain index was measured by a self-reported scale).

Did the authors specify primary and secondary outcomes?

Forty out of 185 RCTs (21.6%) distinguished between the primary and secondary outcomes in the methods section. Thirty-one (77%) of these 40 trials adequately identified the primary outcome. Table 3 provides additional information on the characteristics of each outcome. The adequate

reporting of the primary outcome appeared to have improved over time: from 0% (before 1994 no study reported the primary outcome) to 9% (3/35) between 1995-1999, 27% (11/41) between 2000-2004, and 45% (17/38) between 2005-2010.

Table 3. Characteristics of the most reported outcomes in rehabilitation interventions for mechanical LBP.

| | Most reported outcomes | | | |
	Pain	**Disability**	**Range of motion**	**Quality of life**
Number of studies	164	119	70	45
(%)	(88.6)	(61.6)	(37.8)	(24.3)
Number of studies citing it as primary outcome *	13	19	0	1
Number of instruments	70	43	41	19
Number of self-reported instruments	62	43	0	19

*among 40 studies that identified the primary outcome, we considered only the 31 trials reporting only one outcome as the "primary outcome" in the methods and/or using it for sample size calculation, including the 9 trials using combined outcomes and 7 trials were the primary outcome was not within the four most reported outcomes.

Completeness of outcome reporting

Which aspects of outcome description were most reported?

The completeness of reporting was evaluated for the 4 most frequently used outcomes: pain, disability, range of motion and quality of life. The most often adequately described item for "pain", "disability" and "quality of life" outcomes was the *instrument to measure the outcome;* for the "range of motion" it was the reporting of the *timeline and FU measurement.* The most often poorly described item for "pain", "disability" and "range of motion" was the absence of reporting related to the *methods used for the data collection;* for the "quality of life" it was the reporting on the *methods used during the process to protect against bias (who how-bias).* For the most 4 frequent outcomes, Figure 5 presents for each item included in the checklist the proportion of RCTs satisfactory reporting it.
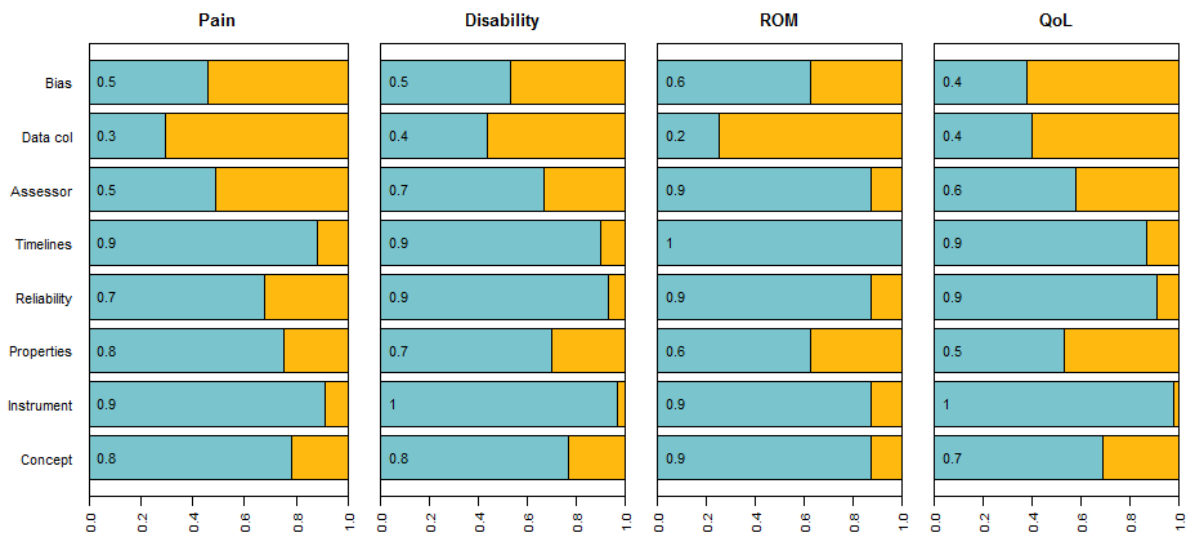
Figure 5. Outcome reporting of all items for the four most reported outcomes in rehabilitation interventions for mechanical LBP.

How many trials have a complete reporting of outcome assessment?

For the 4 most frequent outcomes, only few trials satisfactory reported all items: 10.3% for pain (17/165), 10.2% for disability (12/118), 5.5% for range of motion (4/72), and 3.7% for quality of life (3/45). The large majority of the RCTs insufficiently described the four most frequent outcomes to allow their assessment under the same conditions in future trials.

Did outcomes reporting improve over time?

For the "pain*"* outcome, only four items -*instrument, proprieties, reliability* and *data collection*- had an improvement in outcome reporting that was statistically significant over time, approximately doubling from one decade to the following. Figure 6 presents the relationship between item reporting and calendar year, and the OR of each item being reported vs. not reported for a 10 unit increase of calendar year.

Similarly, for the "disability*"* outcome, only four items -*instrument, reliability*, *data collection* and *method* - had a statistically significant improvement in outcome reporting over time. For the "range of motion" outcome, only the *instrument* item was statistically significant, and for "quality

of life*"* no item improved over the time. Graphs illustrating trends over time for *disability*, *range of motion* and *quality of life* are available in Figures 7-8-9.
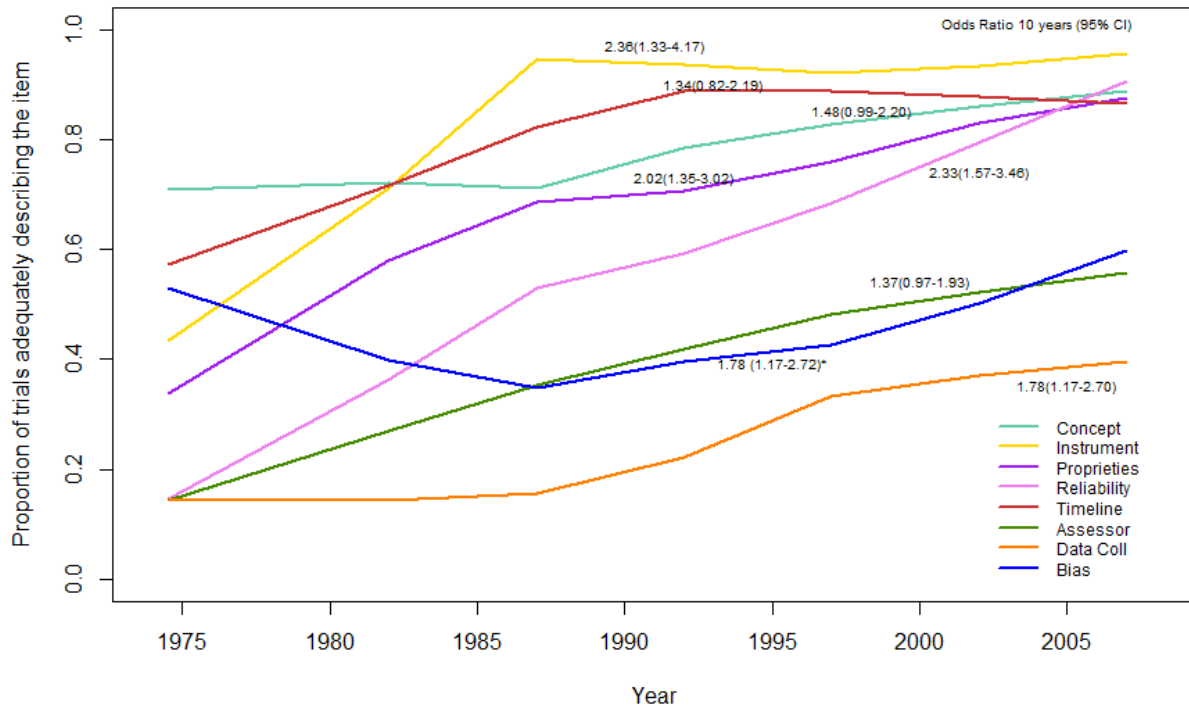
Figure 6. For each of the eight items of the outcome assessment checklist, the graph shows the proportion of RCTs adequately reporting it over the time for the outcome "pain".

The 10-years OR from the univariate model is also reported.

Note
* The coefficient of the quadratic term for year was statistically significant. Regression has been then split into before and after 1980. The reported OR refers to the period ≥ 1980.
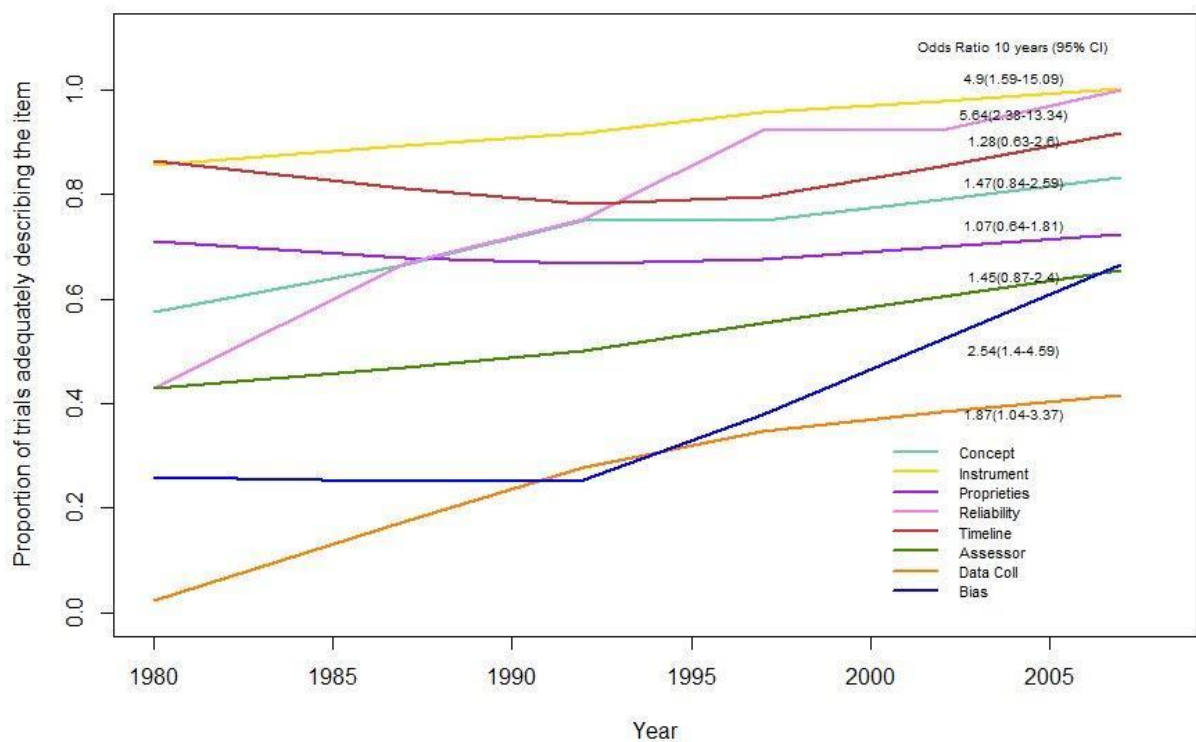
Figure 7. For each of the eight items of the outcome assessment checklist, the graph shows the proportion of RCTs adequately reporting it over time for the outcome "disability".

The 10-years OR from the univariate model is also reported.
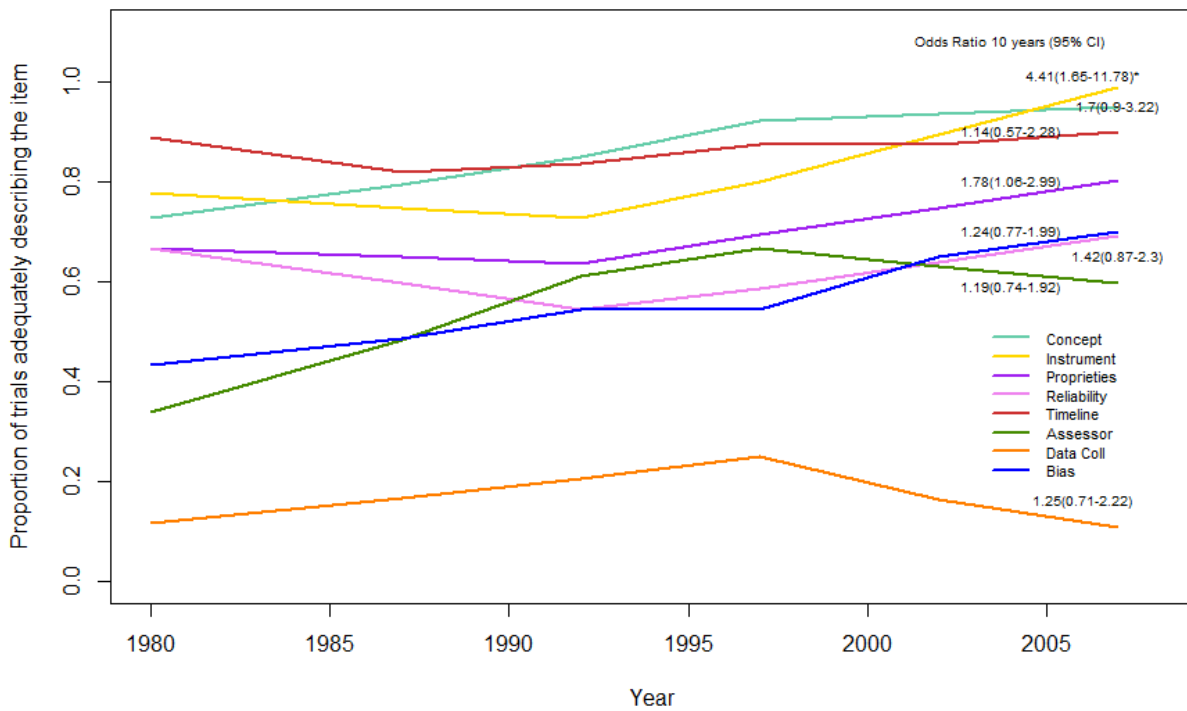
Figure 8. For each of the eight items of the outcome assessment checklist, the graph shows the proportion of RCTs adequately reporting it over time for the outcome "range of motion".

The 10-years OR from the univariate model is also reported.

Note
* The coefficient of the quadratic term for year was statistically significant. Regression has been then split into before and after 1980. The reported OR refers to the period ≥ 1980.
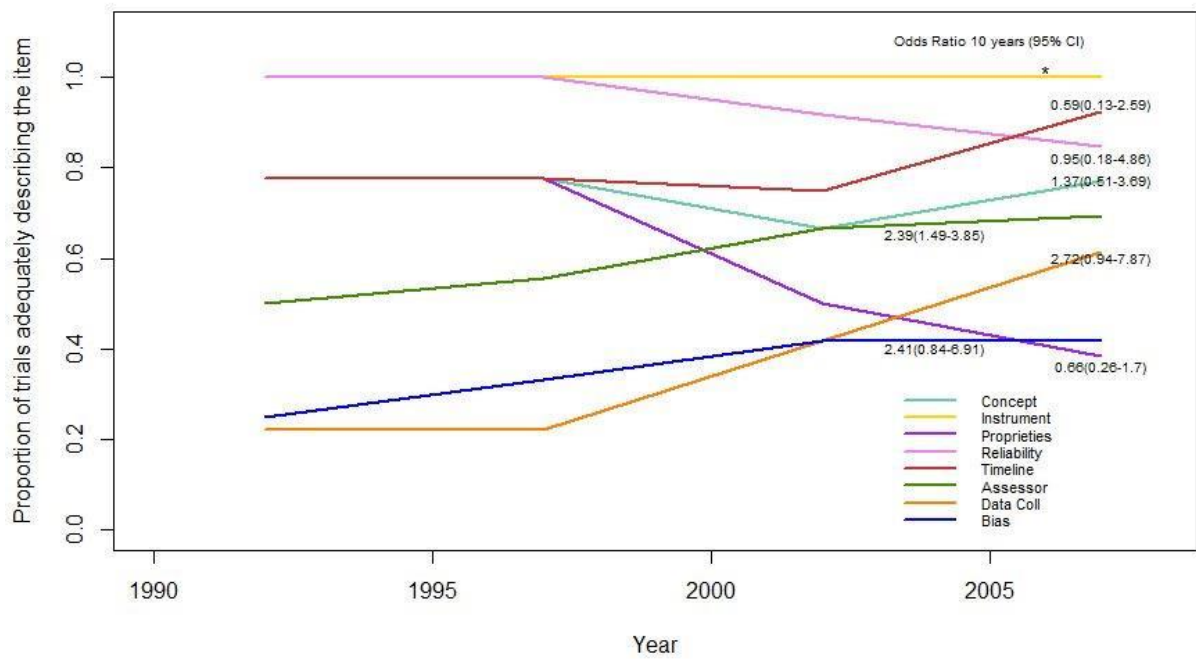
Figure 9. For each of the eight items of the outcome assessment checklist, the graph shows the proportion of RCTs adequately reporting it over time for the outcome "quality of life".

How many trials reported the use of blinding for outcome assessment?

Table 4 details the reporting of the use and level of blinding for outcome assessment. For the four outcomes (pain, disability, range of motion, and quality of life) the use of blinded assessment was: adequately reported in more or less an half of RCTs and unclearly reported in a percentage ranging from 33% (range of motion) to 44% (disability) of trials. The percentage of trials explicitly reporting no blinding varied between 7% for the "range of motion" outcome and 16% for the "quality of life" outcome. For the four outcomes (pain, disability, range of motion, quality of life) the blinding was more frequently employed for the outcome assessor (84.1%, 75.5%, 86.0%,72.7%), followed by participants (29.3%, 28.3%, 25.6%, 27.3%), trial investigator (24.4%, 26.4%,25.6%,18.2%) and data analyst (11.0%,13.2%,9.3%,18.2%).

Table 4. Reporting of the use and level of blinding for the most reported outcomes.

	Most reported outcomes			
	Pain n° of RCTs (%)	**Disability** n° of RCTs (%)	**Range of motion** n° of RCTs (%)	**Quality of life** n° of RCTs (%)
n° of RCTs assessing outcome*	165 (89.2)	118 (63.8)	72 (38.9)	45 (24.3)
Unclear Blinding	69 (41.8)	52 (44.1)	24 (33.3)	16 (35.6)
No Blinding	14 (8.5)	13 (11.0)	5 (6.9)	7 (15.6)
Blinding	82 (49.7)	53 (44.9)	43 (59.7)	22 (48.9)
Type of blinding n° of RCTs (%)**				
Participants	24 (29.3)	15 (28.3)	11 (25.6)	6 (27.3)
Trials investigators	20 (24.4)	14 (26.4)	11 (25.6)	4 (18.2)
Outcome assessors	69 (84.1)	40 (75.5)	37 (86.0)	16 (72.7)
Data analyst	9 (11.0)	7 (13.2)	4 (9.3)	4 (18.2)

*the percentage refers to the total number of RCTs investigating the single outcome (n=185)
** the percentage refers to the total number of RCTs reporting a blinded assessment (above line). A trial could adopt one or more type of blinding (e.g., both trials investigators and assessors were blinded).

Limitations

To determine whether the overall reporting was satisfactory, the highest possible threshold for adequate reporting was selected (i.e., all items of the checklists). Lower thresholds would have increased the number of compliant records; however, we judged that the fulfillment of all items was necessary for the true replication of a study. It may be possible that completeness of reporting was influenced not only by the dimension of the outcome (e.g. disability) or the measure used (e.g. Roland and Morris Questionnaire [RMQ]) but also by other merits and limitations (e.g. binary versus continuous, interpretability, relevance, statistical significance). Moreover, the implications of poor reporting in the dissemination of results of RCTs to patients were not explored. To capture the selective outcome reporting bias, it would have been necessary to study discrepancies between the registered protocol and its corresponding full text. Since the sample starts from 1970, it is difficult to detect potential bias as the widespread registration of protocols only began within the last ten years. Finally, when an outcome was assessed through a multidimensional scale (e.g. Oswestry Disability Index [ODI] which encompasses pain and disability), only the most inclusive dimension construct (e.g., disability) was arbitrarily retained.

4. Quality of reporting of rehabilitation interventions in low back pain rehabilitation interventions

The aim was to assess the quality of the reporting of rehabilitation interventions for LBP in RCTs included in Cochrane SRs. Furthermore, the relationship between the quality of reporting of rehabilitation interventions for LBP and the year of publication, presence of funding, and the continent in which the study was conducted were evaluated.

Methods

Strategy search, eligibility criteria and study selection

Search strategy and study selection process are identical to those previously described in chapter 2 (page 20). However, this search strategy was run up to May 2013.

Data Extraction and Analysis

The following general characteristics from each included RCT were extracted: name of journal, year of publication, country of affiliation of the corresponding author, total number of authors, and reporting of funding. To rate the completeness of intervention reporting, the checklist proposed by Schroter et al. was adopted [47]. This checklist outlines the items that should always be reported in an RCT investigating a rehabilitation intervention and largely overlaps with the recently developed TIDieR checklist, a template for intervention description and replication across all medical fields [51].

The checklist by Schroter et al. includes the following seven items: 1) *setting*: where the treatment was delivered; 2) *provider*: who delivered the treatment; 3) *recipient*: who received the treatment; 4) *procedure*: details about how to perform the treatment, including the sequencing of the technique; 5) *materials*: a description of the physical or informational materials used; 6) *intensity*: the dose/duration of individual treatment sessions; and 7) *schedule*: the interval, frequency, duration, or timing of the treatment.

The number of intervention items that were reported in an RCT was assessed ('intervention completeness'). The reporting was considered incomplete if one or more elements were not reported. DistillerSR, a web-based database, was used for data extraction and management [60]. Five pairs of reviewers, all actively practicing physiotherapists trained in the methodology of clinical trials, pilot tested the screening and data extraction process; they then independently extracted the general characteristics of included studies as well as the description of the interventions used. Disagreements or uncertainties were resolved by discussion among the reviewers.

Statistical analysis

Percentages were used to describe the 'intervention completeness' (i.e., proportion of items in the checklist that were reported). Median and IQR were used to describe the number of adequately reported item per RCT. A multivariable logistic regression model, adjusting for funding and continent, to investigate the impact of calendar year on each of the seven items was performed. The proper functional form for calendar year was explored. Results of the models were expressed as OR, and the corresponding 95%CIs, of the item being reported vs. not reported for a 10 unit increase of calendar year. All tests were two-sided with a significance level of 0.05. All analyses were performed with the Distiller SR [60] and R software [84].

Results and limitations

Studies selection

Eleven Cochrane SRs from the Cochrane Library were identified [63,62,64-72], comprising a total of 220 RCTs. Of these, 24 articles were excluded because they were duplicates of the same article or multiple publications of the same RCTs, 7 because they did not fulfill language inclusion criteria, and 4 because unable to retrieve the full text of the studies. The remaining 185 RCTs were included (Appendix 2).

General characteristics

Table 5 reports the descriptive characteristics of included RCTs. The 185 identified RCTs were published across 74 journals. The top journals for the number of published articles were: *Spine* (23.2%, n=43), *Journal of Manipulative and Physiological Therapeutics* (4.8%, n=9), *Pain* and *Archives of Physical Medicine and Rehabilitation* (each 4.3%, n=8), the *British Medical Journal* (3.7%, n=7), and *Physical Therapy* (3.2%, n=6). Over half of the RCTs reported information about funding sources (56.2 %, n=104). The median number of authors included in the studies was 4 (IQR, 3-6). The median year of publication was 1998 (IQR, 1990-2004); only 8 studies were published from 1968 to 1979 (4.3%). The majority of corresponding authors came from Europe (55.6%, n=103), followed by the North and South America (27.6%, n=51).

Table 5. Descriptive characteristics of the included RCTs in rehabilitation interventions for mechanical LBP.

	Number of RCTs	Percentage of RCTs	Median N° of adequately reported item (IQR)
RCTs involved per Continent			
Europe	104	56.2	5.0 (3.0-6.0)
North and South America	51	27.6	5.0 (3.0-6.0)
Asia	19	10.3	5.0 (3.0-6.0)
Oceania	11	5.9	6.0 (4.0-6.5)
Africa	0	-	-
RCTs published into the first 5 most frequent journals **Total number of involved journal n=74**			
Spine	43	23.2	6.0 (5.0-7.0)
Journal of Manipulative and Physiological Therapeutics	9	4.8	4.0 (3.0-5.0)
Pain	8	4.3	5.0 (3.8-4.5)
Archives of Physical Medicine and Rehabilitation	8	4.3	5.5 (4.3-7.0)
British Medical Journal	7	3.7	3.0 (1.0-3.0)
Physical Therapy	6	3.2	3.5 (3.0-5.5)
Other Journals	104	56.2	5.0 (3.0-6.0)
Funding			
RCTs reporting funding	104	56.2	5.0 (3.0-6.0)
RCTs not reporting funding	81	43.8	5.0 (3.0-6.0)

Completeness of intervention description

How many items were satisfactorily reported?

Figure 10 shows the distribution of the total number of items that were satisfactorily reported in each RCT. Across all RCTs, the median number of satisfactorily reported items was 5 (IQR, 3-6). The full replication of the intervention evaluated as possible in 33 RCTs (17.8%) that fulfilled all seven items in the checklist. Three RCTs did not satisfy the reporting of any item (1.6%). Only five RCTs reported online additional materials.
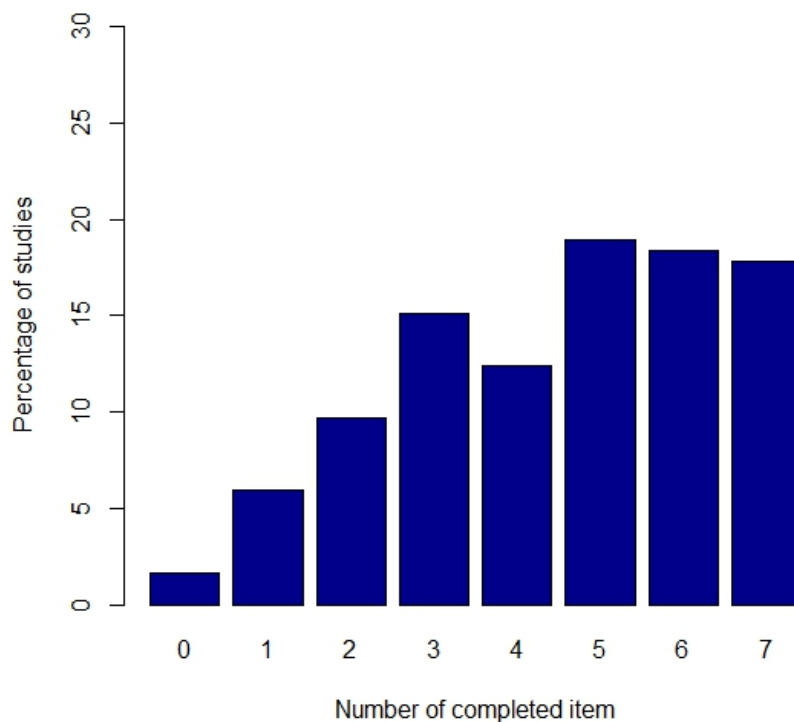
Figure 10. Overall completeness of rehabilitation interventions reporting in mechanical LBP.

Relative frequency distribution of the number of items (out of seven total in the checklist) that were satisfactorily reported in each RCT.

Which items were most satisfied?

Figure 11 reports the percentage of RCTs satisfactorily reporting each of the seven items in the checklist. The most frequently completed items were: *recipient* (91.3%), *provider,* (81.1%), and *intervention schedule* (69.7%). The least frequently completed items were: *procedure,* (43%), *the physical or informational materials* used (48.1%), and the *setting* where the intervention was delivered (53%).
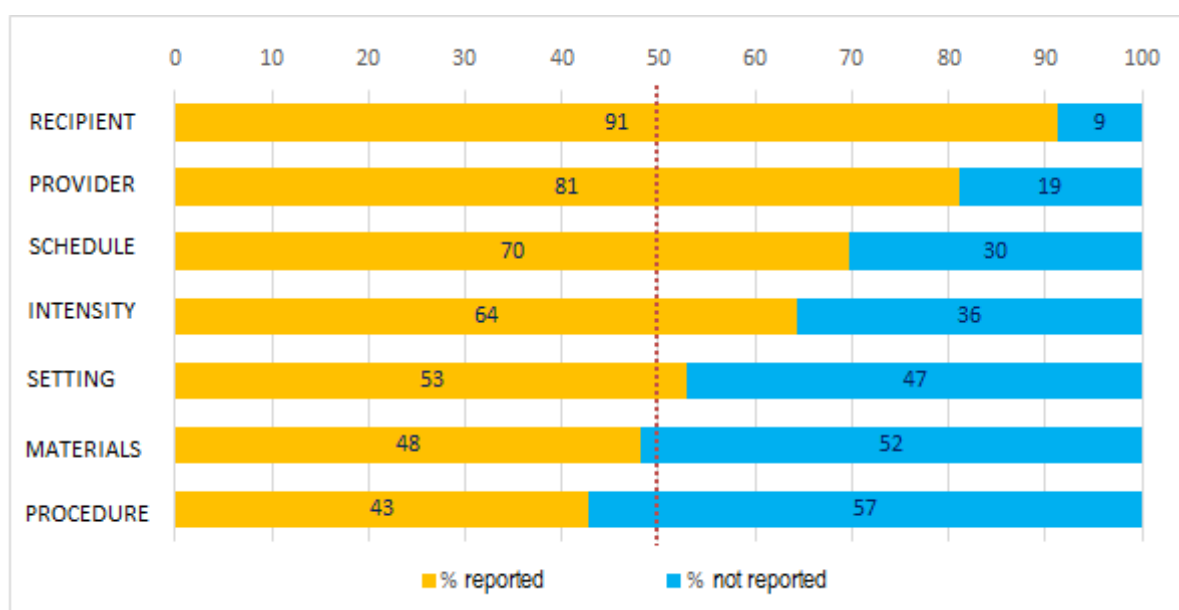
Figure 11. Percentage of studies providing reported information or not reported information in each intervention description item**.**

Legend:
Orange bar= reported information
Blu bar= not reported information

Did RCTs and items improved over time?

The percentage of trials that completely satisfied the reporting of the intervention (i.e., all seven items in the checklist) improved over time, from 14% (7 studies) in the decade 1971-1980 to

20% (75 studies) in the last decade 2001-2010. Figure 12 shows, for each item, the proportion of RCTs that satisfied the reporting over time.

For two items, *recipient* and *intensity*, the improvement in reporting was statistically significant: the reporting of *recipient* doubled from one decade to the following (OR for 10 years 2.06; 95%CI 1.11-3.83), while for *intensity* the improvement was less extensive (OR for 10 years 1.60; 95%CI 1.10-2.32).

A considerable reduction in the percentage of studies adequately describing *intensity, schedule*, and *materials* was found in the last five years examined, interrupting the positive trend over time.
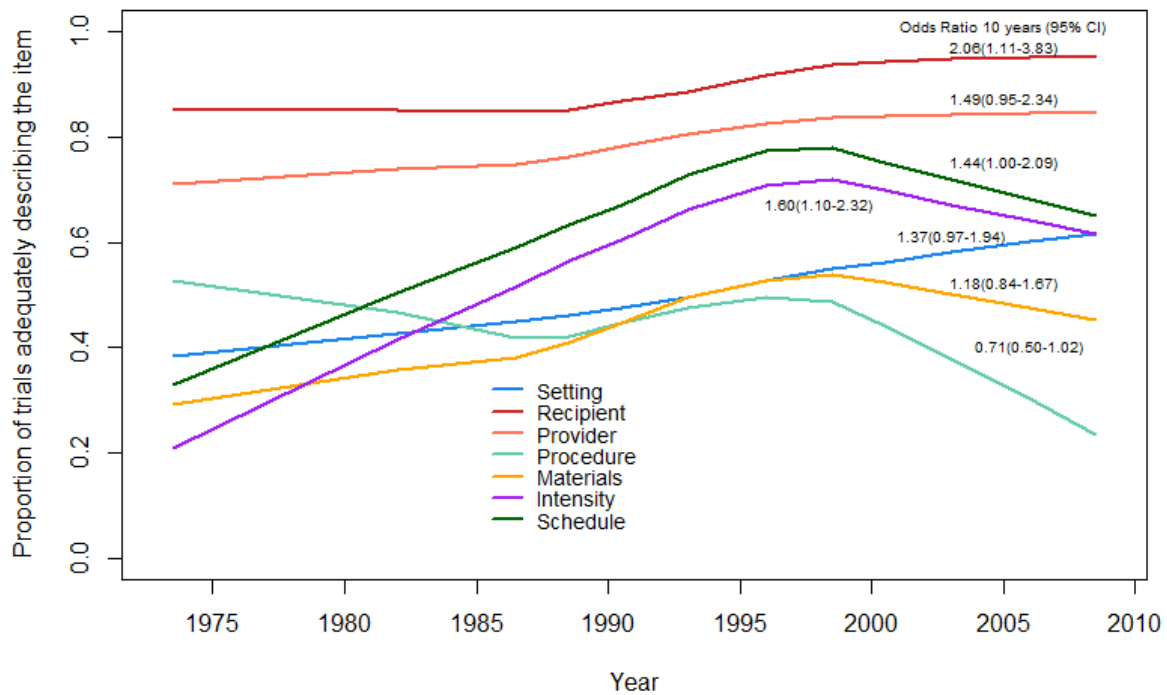
Figure 12. Proportion of RCTs adequately describing each of the seven items of rehabilitation interventions over the years.

The odd ratio for the item being adequately reported vs. not for a 10-year increment, and its 95% confidence interval, are reported.

Is a satisfactory reporting associated with country and funding?

Table 6 shows detailed results for continent and funding. Approximately 25% of the studies from Asia and Oceania and one-sixth of the studies from America and Europe met all reporting criteria. For each continent, *setting* and m*aterials* were completely reported in about half of the RCTs (min-max, 47.4%-58.3% and 45.6%-50.0%, respectively), and *schedule* was reported in slightly more than two-thirds of the RCTs (min-max, 67.0%-78.9%). *Intensity* was less frequently reported in European trials (55.3%) compared to the other continents (min-max, 74.0%-79.0%). Overall, *recipient* was the most reported item (min-max, 82.0%-100%) whereas *procedure* was the least reported (min-max, 37.3%-63.2%). *Provider* was most frequently reported in Oceania (100% of RCTs). The reporting of the person providing care (*provider*) and the person receiving care (*recipient*) were significantly influenced by the continent in the logistic regression model (p=0.01, p=0.04, respectively).

More than half of the 185 RCTs (56.2%) reported sufficient information about funding. Of these trials (n=104), only 20.2% reported all seven items of the intervention reporting. Among the RCTs providing funding information, the most reported item was *provider* (90.4%). We did not find any significant association between a checklist item and the reported funding in the regression models.

Table 6. Number of RCTs per continent with a complete reporting of rehabilitation interventions in mechanical LBP.

	North &Sud America 51(27.6%)	Europe 103(55.7%)	Asia 19(10.3%)	Oceania 12(6.5%)
Provider	39(76.5%)	87(84.5%)	12(63.2%)	12(100%)
Recipient	42 (82.4%)	97(94.2%)	18 (94.7%)	12(100%)
Schedule	36(70.6%)	69(67%)	15(78.9%)	9(75%)
Intensity	38(74.5%)	57(55.3%)	15(78.9%)	9(75%)
Materials	25(49%)	47(45.6%)	11(57.9%)	6(50%)
Procedure	19(37.3%)	43(41.7%)	12(63.2%)	5(41.7%)
Setting	39(76.5%)	87 (84.5%)	12(63.2%)	12(100%)
All 7 items completed	9(17.6%)	16(15.5%)	5(26.3%)	3(25%)

Limitations

First, only rehabilitative interventions for non-specific LBP were explored, excluding conditions such as pregnancy as well as treatments that were non-therapeutic (e.g., orthosis). Second, the studies spans across several decades and the appropriateness of examining old RCTs may be questioned.

5. Discussion of methodological assessment of reporting in LPB interventions

We have investigated the reporting of sample size calculations, interventions and outcome assessment in LBP rehabilitation, along the line suggested by the CONSORT Statement.

Reporting of sample size calculation is often incomplete, only a minority gave a complete description of the elements used. Then we found that numerous trials 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 done correctly.

Also the reporting of interventions and outcomes were largely incomplete were scarce. We found that only a minority of RCTs adequately described all elements of the intervention (18%) and all elements of the outcome assessment (3.67%- 10.3%). The low frequency of description of *materials* and *procedure* for interventions is highly consistent with previous studies [94,95]. Omitting from a trial the description of the *procedure,* if previously unpublished, could yield decision-makers to adopt uncorrected practices that may ultimately cause harm, increase adverse events, or prevent treatment of life-threatening disorders.

Moreover, a large number of outcome measures with a myriad of measurement instruments was used across all RCTs. Difficulties caused by heterogeneity in outcome measurement could be addressed through the development and use of an agreed standardized collection of outcomes, known as the core outcome set (COS), which should be measured and reported [96,97] in all rehabilitation RCTs for a specific condition. A COS is a scientifically agreed set of outcomes that have to be reported as a minimum in all studies conducted within a specific area of clinical practice, audit or research [82]. Standardisation of the used clinical outcome measures is necessary for comparison across studies [56] and registries [98]. Inconsistent choice of outcome measures in clinical trials means that many MAs are unable to include data from all the relevant studies [97].

Improving the reporting of outcome assessment must be pursue in order to avoid bias such as selection outcome or detection bias, and to increase applicability of research results into practice.

Therefore the question is: are rehabilitation researchers generally poor investigators or are rehabilitation journal editors /reviewers not allowing good reporting? There are several possible explanations. On one hand even if most journals' *Instructions to Authors* recommend the use of the specific standards for reporting, only a minority of these require them as mandatory [99]. Additionally, some journals do not encourage authors to provide supplemental materials to enhance reporting. On the other hand, guidelines for transparent reporting have been introduced only in recent years by the EQUATOR initiative and the CONSORT Statement but some time is required to actually see their effect in practice. Also, it could be claimed that sometimes problems of description are not identified by peer reviewers and editors, and that even when problems are detected only about two-thirds are fixed before publication (32.6%) [47]. Furthermore, it is not guarantee that the researchers actually performing the intervention for the clinical trial will be the same that wrote the publication. Maybe trials are well conducted but worse reported.

What could be done to improve the reporting in rehabilitation? Journals can help to improve the problem of incomplete reporting by providing specific instructions to authors, requiring editors and peer reviewers to verify the compliance with the instructions, detecting missing details before publication. Ideally, the full intervention description should be published with the primary article, but this often is not feasible, for example, with manual procedures or extensive training materials. Since describing such study materials could add significantly to the length of papers, we suggest that editors encourage the use of web extras and/or links to study materials on authors' or funders' institutional websites.

Authors are invited to be compliant with the guidelines for reporting. For example, an higher adherence to protocol stage, as recommended by SPIRIT initiative [76], and to the CONSORT statement [36]. In particular we call for adopting the new TidieR checklist for intervention reporting [51], the more explicit set of elements for adequate reporting of sample size recently published [100] and a core outcome set for LBP [101,102].

Therefore, if all rehabilitation journals endorsed the use of standardised checklists for reporting, the rehabilitation community could reach the goal to give a universal definition and exact replication of rehabilitation intervention. It would also promote the dissemination and replication of the most effective intervention in the way proposed in the original RCTs, allowing the same level of efficacy.

This might also be an opportunity to reduce costs for clinicians and patients, giving to non-drug therapies an alternative treatment over more expensive pharmacological therapies whenever possible.

Funders, authors, journals and research users should all be concerned with this issue and work together to promote the improving of intervention, outcome assessment and sample size calculation. Transparent and accurate reporting is a crucial step to facilitate the transferring into practice of research findings for community rehabilitation readers that have to use the information.

PART 2. Assessing clinical relevance in rehabilitation interventions for LBP

1. Introduction

Clinical relevance versus statistical significance

Clinical studies aim to show differences between two or more groups of patients undergoing different interventions in terms of an outcome, usually over time.

In common parlance a difference is called significant if the change is important or meaningful. In the world of statistics, a significant difference is simply a difference that is unlikely to have been caused by chance [103]. However, an observed statistically significant difference in terms of outcome between two interventions does not necessarily imply that this difference is clinically important [16] or that changes were clinically relevant for patients [17]. For example, a study demonstrated a statistically significant difference for acupuncture in reducing spasticity after stroke, however these improvements are not clinically relevant changes for the patients [104]. Often, also in the rehabilitation field, statistical significance is obtained even for a small clinical effects, because of a small variability in the sample or a huge sample size. However statistical significance by itself provides little information about the clinical meaningfulness of a treatment [105]. The concept of ''clinically important difference'' evolved to overcome the shortcomings of the ''statistically significant difference'' that considered outside the clinical contest [103].

Interpretation of patient-reported outcomes measure for clinical relevance

The concept of MID is particularly helpful in the evaluation of PROs because it could be offered as the new standard for determining effectiveness and describing patient satisfaction with a given treatment [103].

Clinical trials evaluating medical treatments and health interventions increasingly incorporate patients' self-reported measures.

As discussed in part I, PROs are very common in the rehabilitation field. For example, the effectiveness of LBP rehabilitation treatment in terms of pain relief is currently measured by the visual analogue scale (VAS), a patient-reported outcome. As described by Ostelo in 2008, there are many other self-reported measures used for the evaluation of LBP interventions [106]. In order to compare results between studies using different PROs and to enable data pooling in SRs, an international group of researchers recommended the use of a standardized COS of measures in 1998 [56], which was revised in 2000 [107]. Since then no update was performed. As previously shown in chapter 2, pain, disability and quality of life are the most used PROs in LBP population.

PROs require the patient to assign a response to questions about their perceptions or activities (e.g., symptoms). These responses are typically combined in some way to create summary scores that can be used to measure an outcome. Demonstrating the ability to detect responsiveness to meaningful change is necessary but not sufficient for estimating the smallest change in score that can be regarded as important [108]. Responsiveness has been defined as the ability of a questionnaire to detect clinically important changes over time [109], whereas the MID denotes the smallest score or change in score that would likely be important from the patient's or clinician's perspective. Because responsiveness and MID depend on population and contextual characteristics, there is not necessarily a single MID value for a PRO instrument across all applications and patient samples. There is often a range in MID estimates that varies across patient population and clinical study context [108].

The MID was defined in 1989 as ''the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management''[110]. A treatment effect reaching the MID would convey clinical relevance to study results and might justify clinicians' decision to treat [111]. Actually, research in the field of MID has mainly focused on the side of patient's benefit and most commonly considers the MID to be ''the smallest change that is important to patients''[112]. The MID is a threshold value for such a change and any amount of change greater than the MID threshold is considered to be clinically meaningful or important. Several measures, other than the MID, have been defined to capture the concept of a difference in outcome that is relevant for the health professional and/or the patient, most notably the MCD (minimal clinical difference), MCID (minimal clinical important difference) and the MCSD (minimal clinically significant difference)[113].

Methods to estimate the MID

Several studies have presented a clear overview of the different methods to assess a MID and provided some priorities for future research [103,114,115]. However, there is no consensus in the literature on the most appropriate technique for determining the MID and different methods may result in different values [116,17] This has led to some confusion about what is the minimum change that is clinically important for commonly used back pain outcome measures [106]. Nevertheless, four general approaches have been used to determine a MID: anchor-based methods (e.g., comparing VAS pain scores to a non-subjective measurement such as abilities or drug's reduction), distribution-based methods (e.g., built on the variability of the pain scores), panel of experts and previous clinical trial experience [108]. All approaches measure a

quantifiable change in outcomes, but they can produce different values for the change considered as clinical meaningful [103].

Anchor- based approaches

The anchor-based approach is the method most frequently used in PRO measures. It is based on an external criterion, either a different clinical variable for which a recognized clinical difference already exist (i.e., physiological measures) or patient-based indicator (i.e., actual changes in PRO measures such as mean changes in the actual population). The external indicator is needed to assign subjects into several groups reflecting the amount of clinical or health status change in the target patient population [108]. The determination of a specific MID value will depend on the selected anchor being proven to be a valid indicator of small but important changes [111]. It is strongly recommended to use multiple independent anchors and to examine and confirm responsiveness across multiple samples [117,118]. Although the use of an external criterion is the common characteristic of all anchor-based approaches, many differences remain between approaches [103].

Typically, methods used to assess the MID from patients, are based on a retrospective judgment about whether they have improved, stayed the same, or worsened over some period of time. The method then establish a threshold based on the change in health related quality of life in patients who report minimal change, either for better or for worse. Another class of anchor-based methods involves longitudinal FU to determine whether it is possible to identify subgroups that have clinically different outcomes, such as re-hospitalization, relapse of cancer, Medical Research Council grading, or different interventions. Although these approaches clearly yield differences that are clinically important, it is not at all clear that they are, in any sense, minimal. The last approach identifies subpopulations with different levels of health (for example, patients on hypertensive therapy vs. blood pressure therapy) and then looks at the

differences in scores on a generic HRQL measure. Although these differences have external significance in terms of population differences, the link to clinical relevance, or to any estimate of minimal relevant difference at the individual level, is unclear [119].

A limitation of anchor based MIDs is that not all instruments of an outcome have an available anchor-based MID. In addition, an anchor-based method depends on the characteristic of the population on which it was constructed. For example, in a surgical ward the MID for LBP relief should be at least 4-5 points of VAS, whereas in chronic LBP a 1.5 points of MID is acceptable.

Distribution based approaches

Most of the traditional approaches to determine the MID fall under the distributional approach [115]. Distribution-based methods define the MID based on the distribution of observed scores in the population, estimated in a relevant sample [108]. Several distributional methods have been developed, some specific for PROs, and many of them by Guyatt [117]:

1. *Minimum detectable change*

Minimum detectable change was defined as $1,96\sqrt{2} \; x \; SE$, for a 95% confidence interval [114,120]. A related concept, the reliable index change is obtained by dividing the individual patient change score by the square root of the standard error of measurement (SEM). If the RCI is greater than 1.96, the change in the patient is considered to be a true change with 95% confidence [121,105].

2. *Effect size (ES)*

It is a standardize measure of change obtained by dividing the difference in score from the baseline to post-treatment by the SD of the baseline score. It represents the number of SDs by which the score has changed from baseline to post-treatment. Cohen defined 0.2, 0.5, 0.8 as a small, moderate and large ESs, respectively [122].

3. *Half a standard deviation, and standard error of measurement*

'Half a standard deviation' (0.5 SD) was defined as half of the SD of the baseline scores[119,123]. SEM was defined as $SD \sqrt{1-r}$ where SD is the standard deviation of the baseline scores and r is the test-retest reliability coefficient [110]. A change score below the value of the SEM may reflect the imprecision of the measurement rather than a true change. The main critique to distribution-based methods consists on considering them as a ''purely'' statistical argument to support the choice of a MID [124].

Also, each method produces a different MID values. Finally, MID definitions do not take into account the cost of treatment to the patient, and the change in PRO scores depends on the patient initial baseline status [103].

Panel/opinion based approaches (Delphi method)

These are heuristic approaches, based on adequate surveys of experts' opinion, formally pointed out using a clinician's global assessment. Their authority is predicated on the foundations of experience, knowledge, data, and anecdote. In the end, this model is based on the belief system of the clinician (and importantly for some measures — the belief system and preferences of patients). This form of evidence is intuitive and often is the only evidence that is readily available [125].

This approach establishes the MID on the opinion of expert pairs, connoted by a subjective nature inclination.

Previous clinical trial experience

Since the literature on PROs applied in clinical trials increases, it is increasingly possible to understand responsiveness and MID for different PRO instruments based on demonstrated differences between two or more active treatments in clinical trials. Their results in PROs measures can provide insight into observed effects based on treatment comparisons and should be used to help determine MID [108].

Open questions and aims in LBP rehabilitation interventions

Since not always statistically significant difference means clinical relevance, we firstly would interpret the results of mechanical LBP rehabilitation interventions in published RCTs in terms of both statistical significance and clinical relevance. Secondarily, the goal is to synthetize the evidence according to the concept of clinical relevance in MA. Therefore, we would interpret MAs in MIDs and not only in a classical way of statistical significance. We will propose an application of the concept of MID to MAs to ameliorate the clinical interpretation of the results.

2. Low back pain Rehabilitation interventions: are they statistical significant and clinically relevant?

The aim was to interpret the results of published trials in mechanical LBP rehabilitation interventions in terms of clinical relevance and statistical significance.

We re-analyzed results from RCTs applying a priori planned MID, the one used in the sample size calculation, in order to verify if findings were achieved in terms of clinical relevance. The primary outcome was to identify the number of RCTs classified as statistically significant and clinically relevant, statistically significant but not clinically relevant, clinically relevant but not statistically significant, and both not statistically and clinically significant. The secondary outcome was the same categorization for all compared interventions.

Methods
Search strategy

Cochrane Database for SRs published from 1995 to April 2017 was searched using the terms "back pain" and "rehabilitation" in title, abstract and keywords. Only Cochrane SRs were selected because they represented a gold standard for identifying all relevant RCTs in a field [57,58] through highly sensitive search strategies.

Eligibility criteria and selection of studies

Inclusion of SRs

A SR was included if nonspecific LBP and rehabilitation interventions were involved. Following the definition given by the National Library of Medicine, rehabilitation was defined as the 'return to the highest possible degree of human function in people suffering from a disease or have damage done by rehabilitation professionals' [59]. SRs focused on interventions

other than rehabilitation (eg, prevention) or subgroups the population (eg, pregnant women) have been excluded.

Inclusion of RCTs

Studies were included from all eligible SRs whenever they met the following criteria: (i) the study design was an RCT, (ii) the language of publication was English or Italian, (iii) the publication reported the sample size calculation for the primary outcome and the a priori planned MID.

Two authors independently searched and evaluated the Cochrane SRs and then examined the records of all potentially eligible RCTs of SRs, eliminating duplicates. Each disagreement was solved by consensus; if it was not possible to find an agreement, a third author was consulted.

Data extraction

An ad hoc data collection form in Excel was developed. Data extraction was performed by an author and verified by a second author, in case of disagreement a third author was consulted for a final decision.

General characteristics were extracted from each RCT (e.g., year of publication). After that, the following methodological features were collected: planned sample size, primary outcome and referred measurement instrument, details on measurement scoring, a priori planned MID, bibliographic reference and explanation of the nature of the MID (eg., anchor/distribution method or other methods), scheduled FU, number of randomized patients, number of patients evaluated at the time of FU, and dropouts. If the FU time was not specified by the authors, the first post-treatment FU was selected. To estimate the clinical relevance of the results, the following data were extracted: mean difference (MD) of the primary outcome measure between treatment and comparator with its 95%CI or SD, and statistical significance of the difference

reported as achieved or not achieved statistical significance at the declared alpha level. The clinical relevance of the intervention effect was then determined from the obtained data considering if it was achieved, i.e. the MD was equal or greater than the declared MID, or not achieved.

Data analysis

Data were described by medians and IQR for continuous variables, number and percentages for categorical variables. The number of RCTs falling in each of the following four categories was calculated: *statistically significant and clinically relevant, statistically significant but not clinically relevant, clinically relevant but not statistically significant, and both not statistically and clinically significant*. If a trial presented multiple arm comparisons, at least the statistically significant one was selected in order to be more conservative. The absolute frequency of results was reported from all multiple arm comparisons of interventions into the same four categories.

Results

Study selection

Sixty-one Cochrane SRs were identified in the Cochrane library. After removing duplicates, a total of 20 SRs were included in this study. One hundred-one RCTs were considered eligible, but only 42 of these met the inclusion criteria. The study selection process is shown in the flow chart of Appendix 3.

General characteristics

The 42 included RCTs were published in 18 journals. Most of these were published in Spine (31%, n = 13), in the British Medical Journal (12%, n = 5), and in the Clinical Journal of Pain (12%, n = 5). Fourteen countries have been designated as publishing countries, of which the most frequent are the United States (26.2%, n = 11) followed by the United Kingdom (19%,

n= 8), Norway and the Netherlands (9.5%, n = 4). The publication period of the 38 trials runs from 1996 to 2014 (median = 2006; IQR = 2003 - 2008). General characteristics are reported in table 7.

Table 7. General characteristics of the RCTs.

Characteristics		n° of RCT	%
N° of countries (n=14)			
	United states	11	26.2
	United Kingdom	9	21.4
	Norway	4	9.5
	Spain	1	2.4
	Finland	2	4.8
	Sweden	1	2.4
	Netherland	4	9.5
	Switzerland	1	2.4
	Italy	1	2.4
	Brazil	3	7.1
	Australia	3	7.1
	Thailand	1	2.4
	Taiwan	1	2.4
N° of journals (n=19)			
Most frequent journals			
	Spine	13	31.0
	Clinical Journal of Pain	5	11.9
	British Medical Journal	5	11.9
	Journal of Manipulative and Physiological Therapeutics	3	7.1
N° of reported funding		34	81.0

Sample size characteristics

To evaluate the interpretation of statistical significance and clinical relevance in the 42 trials, the sample size calculations of 60 outcomes were evaluated, accounting for a total of 81 comparisons of interventions.

The most frequently observed outcome used in sample size calculation was disability 67% (28/42), measured through the RMQ 61% (17/28) and the ODI 36% (10/28). The second most

used outcome was pain 32% (19/60) mainly described through the Visual Analogue Scale 63% (12/19) and followed by the Pain NRS (Numerical Rating Scale) 16% (3/19). Four studies did not report the number of patients obtained in the sample size calculation. For the other 38 studies, the median of the sample sizes planned a priori was n=125, while the median of the enrolled sample sizes was n=133.

Clinical relevance characteristics
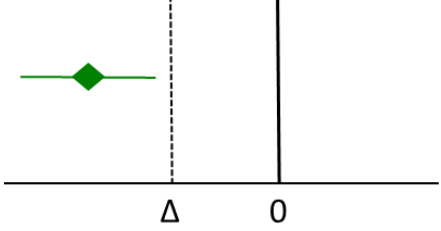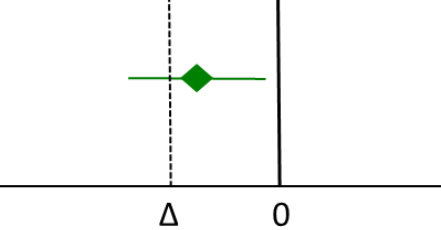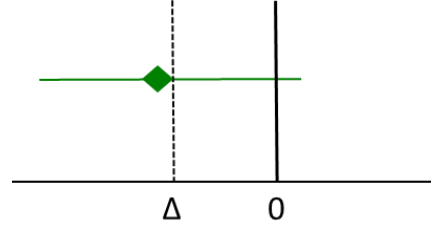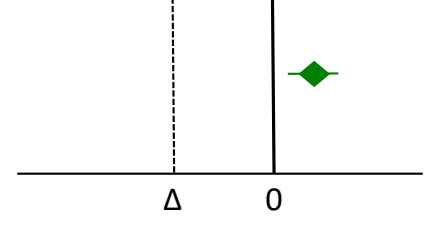
Most trials (n=37, 88.1%) reported MID as an absolute value, while 5 trials (11.9%) reported it as a percentage of improvement over the baseline.

Half of the studies (n = 20, 47.6%) presented the bibliographic reference of the source used to calculated the MID. Eliminating duplicates 16 different sources were found and examined. Of those, 6 (37.5%) used an anchor based method to estimate the MID, 1 (6.3%) a distribution based one, 1 (6.3%) an Expert panel, 3 (18.8%) cited other articles, 3 (18.8%) were not clear in methodology, while 2 (12.5%) were not found.

Is the effect always clinically relevant?

Table 8 shown the main findings for statically and clinically results. Overall, we found that more than one-third of RCTs (37.5%, n=15) were both statistically and clinically significant whereas few (23.8 %, n=10) were statistically significant but not clinically relevant. Taking into account all comparisons of multiple arm trials (n=81) the scenario was slightly different: less than one-fifth of the compared interventions were neither statistically nor clinically significant (24.7%, n=20/81). Among these, 70% (n=14/20) considered placebo or usual care as comparator while only 30% (n=6/20) explored head to head comparisons.

Table 8. Statistically significance and clinically relevance on continuous outcomes of LBP. Δ is the MID. Negative values means greater pain reduction in the treatment vs. control group.

Scenario	N° of trials (total=42), Frequency (%)	N° of comparisons (total=81), frequency (%)
a) statistically significant and clinically relevant	15 (37.5%)	20 (24.7%)
b) statistically significant but not clinically relevant	10 (23.8%)	22 (27.2%)
c) probably clinically relevant but not statistically significant	1 (2.4%)	2 (2.5%)
d) not statistically or clinically significant	15 (35.7%)	34 (42%)

3. Re-analysis of a meta-analysis in minimal important difference units in low back pain interventions: are the effects still significant?

Published as Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: the need to present minimal important differences units in meta-analyses.
Gianola S, Andreano A, Castellini G, Moja L, Valsecchi MG.
Health Qual Life Outcomes. 2018 May 15;16(1):91.

The aim is to explore how performing the MA in MID unit, instead of MD, changes the interpretation of the summary estimate and the conclusion of a SR focused on rehabilitation intervention for LBP; we will then discuss the practical problems encountered applying thus methodology. Data from a published Cochrane review of "Multidisciplinary biopsychosocial rehabilitation for chronic LBP" [74] were used.

Methods

Standard methods for meta-analyses of a continuous patient-reported outcome

Pain is a PRO score typically treated as continuous variable. In MA of continuous data, the MD is the measure of the absolute difference between the mean outcome values measured in each arm of a parallel group clinical trial. When outcome measurements in all trials are made on the same scale, a well-established inverse variance MA method can be used to combine results across i trials and obtain a pooled MD estimate [126] from the MD of the individual study (MD_i) and their variance $Var(MD_i)$.

$$MD_{i=}\ m_{1i} - m_{2i} \qquad\qquad Var(MD_i) = \frac{sd_{1i}^2}{n_{1i}} + \frac{sd_{2i}^2}{n_{2i}}$$

A pooled estimate can be obtained by either a fixed or random-effect model. Under the fixed-effect model we assume that all studies in the MA share a common (true) ES. In our case the true MD in the population of patients with LBP. The weight (W_i) assigned to each study in the

fixed-effect MA is the inverse of variance of the within study variance assessed to be normally distributed for a continuous outcomes:

$$W_i = \frac{1}{Var(MD_i)}$$

The weighed mean (MD) is then computed as:

$$MD = \frac{\left(\sum_{i=1}^{k} W_i MD_i\right)}{\sum_{i=1}^{k} W_i}$$

The variance of the summary effect is estimated as the reciprocal of the sum of the W_i weight, or:

$$V_{MD} = \frac{1}{\sum_{i=1}^{k} W_i}$$

However, in many SRs the fixed-effect assumption is implausible. Studies will differ in the clinical and demographical characteristics of participants and in the implementations of interventions, consequently the true ESs can be different in different studies. A solution is to use a random effect model, which assume a normal distribution of the true effect across studies with a variance τ^2. To compute a study's variance under the random-effects model, we need to know both within-study variance $Var(MD_i)$ and τ^2. The classic method for estimating τ^2 is the following [126]:

$$\tau^2 = \frac{Q - df}{C} \qquad \text{with } Q \text{ estimated as:}$$

$$Q = \sum_{i=1}^{k} W_i MD_i^2 - \frac{\left(\sum_{i=1}^{k} W_i MD_i\right)^2}{\sum_{i=1}^{k} W_i}$$

where k is the number of studies, df the degree of freedom (df=k-1) and C the denominator of τ^2, estimated as:

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

However, when different scales are used to measure the same conceptual outcome (e.g. pain, disability), SMD is adopted as the summary measure instead of MD. This because is necessary

to standardize the different outcome measures before they can be compared and combined in a MA to obtain a pooled SMD, using the inverse variance method either by fixed-effect or random-effects models [126]. In this case the study outcome measure SMD_i and its variance $Var(SMD_i)$ are calculated as:

$$SMD_i = \frac{m_{1i}-m_{2i}}{s_i}\left(1-\frac{3}{4N_i-9}\right) \quad ,with \qquad s_i = \sqrt{\frac{(n_{1i}-1)SD_{1i}^2+(n_{2i}-1)SD_{2i}^2}{N_i-2}}$$

$$Var(SMD_i) = \frac{N_i}{n_{1i}n_{2i}} + \frac{SMD_i^2}{2(N_i-3.94)}$$

The SMD expresses the intervention effect in SD units rather than the original scales, with its value depending on both the size of the effect (the difference between means) and its variability. This approach has two important limitations. First, the same effect will appear different if population heterogeneity across eligible trials differs; second, health professionals and decision makers will not have an intuitive sense of the clinical importance of the effect reported in SD units.

MID units method for meta-analysis of different pain instruments

A potential solution to the limitations of SMDs pooling is to substitute it with a MA in MID units. MID units, (MU_{ij}) are define as MD divided by the MID that was established for the j instruments used in the trial. If available from the literature, the MID should be anchor based [127].

$$MU_{ij} = \frac{MD_i}{MID_j} \qquad for\ j\ from\ 1\ to\ J \tag{1}$$

The variance of MU_{ij} is [127]:

$$Var\left(MU_{ij}\right) = \frac{Var\ (MD)_i}{MID_{ij}^2}$$

Defining the trial weights as $w_{ij}=1/Var(MU_{ij})$, we can use the inverse variance method to pool MID units using the following formula:

$$MU = \frac{\left(\sum_{i=1}^{k} w_i MU_i\right)}{\sum_{i=1}^{k} w_i}$$

In the fixed effect model, the variance of MU can be calculated using the following formula:

$$Var(MU) = \frac{1}{\sum_{i=1}^{k} w_i}$$

and confidence intervals can be subsequently derived. Pooling of MID units is naturally extended to the random-effects model using weights equal to $w_i = 1/(Var_{MU} + \tau^2)$.

The imputation of MID for instruments without an established MID

When an anchor-based MID for an instrument is not present in the literature, one option is to choose a distribution-based approach to calculate it. We used the method proposed by Johnston et al. [128] to derive a MID for these instruments.

The following measures were calculated: the "standard deviation ratio" (SDR_i), the anchor-based MID divided by the baseline standard deviation (SD_i), for the control group or, if not reported, the SD at the end-of-treatment for the same group.

$$SDR_{ik} = MID_k/SD_{ik} \tag{2}$$

where $k = 1, \dots K$ instruments with an anchor based MID

When several $SDRs$ are available from trials using instruments with an established anchor based MID, we can calculate their median. To estimate the MID for instruments without an anchor-based one, we multiplied the $median\ SDR$ of the studies with an anchor based MID for their SDs (baseline or endpoint).

$$MID_{ig} = SD_{ig} * median\ SDR_{ik} \tag{3}$$

where $g = 1, \dots G$ instruments without an anchor based MID

The calculated distribution-based MID was used for the instruments without an anchor based one to compute MID units for that trial (eq.1).

Case-study: application of the method

Investigated comparisons

The MA was firstly performed in SMD and subsequently in MID units, interpreting the obtained summary "pain relief" effects for the following comparison: "MBR versus usual care in chronic populations" in short, medium and long terms of FU studies.

The three MA included 13 trials of which 10 employed two widely used disease-specific pain instruments: the pain NRS and the VAS. Both instruments have demonstrated a validity and responsiveness in various setting [129]. Another valid pain instrument [130] reported in two trials [131,132] was the SF-36 bodily pain (Medical Outcomes Study 36-Item Short Form Health Survey). One primary study did not report the employed instrument [133].

Definition of the anchor based MID for instrument measuring pain relief in low back pain

An extensive literature search was performed in order to define an established MID for all instruments usually used to measure perceived back pain in studies for people with nonspecific LBP excluding specific causes for back pain (e.g., such as cauda equina syndrome). The search strategy adopted from Ferreira 2012 [134] was updated to May 2015, which was conducted on MEDLINE up to May 2011 and of CINAHL, LILACS, and EMBASE up to August 2010 without restriction for language.

Fourteen studies were reported having an anchor based MID for NRS, 5 for VAS and only 1 for three other different instruments (pain self-efficacy questionnaire, patient-specific functional scale and 11-Face Faces Pain Scale). The MIDs established by Ostelo et al. [106] for the *VAS 0-100 mm* and for the *NRS* were adopted in our MAs, since this study is a consensus expert panel that revised anchor based MIDs and its results are consistent with all other retrieved studies. The proposed MID values are 15 for VAS and 2 for NRS.

For the other instruments presented in our MA (SF-36 body pain), a widely accepted anchor-based MID was not find and the distribution-based method, above described (page 74), to calculate their MID (eq.1-3) was adopted. The same approach was used for the study not reporting the employed instrument.

Results

Descriptive characteristics

Short term MBR versus usual care for reduction of back pain includes 9 studies; the medium term comparison counts 6 studies and the last-term one counts 7 studies. Table 9 summarizes the number of studies reporting the different instruments for the three analysed comparisons.

Table 9. Number of studies using each instrument for the considered comparisons.

	Instruments			
Comparison: **MBR vs Usual Care**	NRS a.b. MID=2	VAS a.b. MID=15	SF36-body pain d.b. MID	Not reported d.b. MID
Short term(n=9)	2	4	2*	1**
Medium term (n=6)	2	3	1***	0
Long term (n=7)	3	4	0	0

Legend:
a.b.=Anchor based, d.b.= Distribution based
* Tavafian 2008, MID= 24,8; Tavafian 2011, MID=19,5
** Moix 2003, MID= 2.6
*** Tavafian 2011, MID=20,4

Results of the meta-analyses in SMD

Data from the Cochrane SR [74] were used to calculate the SMDs and their 95% Confidence Intervals (Table 10). The MD was calculated as control minus treatment in all MAs, i.e., a positive SMD favors MBR over usual care.

Table 10. Reported outcomes of the studies included in the MAs.

Study	Intervention			control		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**
Short term						
Abbassi 2012	2.6	2.0	12	3.2	1.6	11
Basler 1997	4.1	2.1	36	4.2	1.4	40
Lambeek 2010	3.9	2.5	60	5.5	2.3	62
Moix 2003	14.5	3.2	13	14.9	3.2	15
Morone 2011	4.5	2.3	41	7.6	2.1	29
Morone 2012	5.0	2.2	25	8.0	2.2	25
Tavafian 2008	-71.5	16.2	44	-56.6	30.0	47
Tavafian 2011	-65.8	22.6	92	-56.3	23.6	97
Von Korff 2005	4.9	2.0	110	5.3	1.9	120
Medium term						
Bendix 1996/98	5.7	2.1	45	6.9	2.1	49
Lambeek 2010	3.6	2.6	58	4.8	2.4	60
Morone 2011	4.4	2.5	41	6.5	1.9	29
Morone 2012	4.0	2.2	25	7.0	2.2	25
Tavafian 2011	-72.3	22.8	92	-60.3	25.8	96
Von Korff 2005	4.2	2.0	110	4.7	2.2	110
Long term						
Abbassi 2012	3.7	2.5	12	4.3	1.4	11
Bendix 1996/98	6.0	2.2	50	6.5	2.2	49
Lambeek 2010	4.2	2.7	59	4.5	2.7	60
Linton 2005	2.9	2.0	61	4.1	2.6	47
Lukinmaa 1989	47.3	20.5	86	44.6	20.5	72
Strand 2001	37.2	20.5	81	42.5	20.5	36
Von Korff 2005	4	2.3	99	4.7	2.1	98

Short term multidisciplinary biopsychosocial rehabilitation versus usual care for back pain relief

Nine studies, involving 879 patients, investigated the effect of a MBR intervention versus usual care for pain relief. The point estimate for the pooled differences is 0.56 SMD (95%CI, 0.29, 0.85) in the short term (Figure 13). The 95%CI did not include the zero, indicating a statistically significant effect in favor of MBR over usual care in terms of pain relief for LBP patients.

Study	Instrument	Weight	SMD 95% CI
Abbassi 2012	VAS	2.70%	0.32 [-0.51 , 1.14]
Basler 1997	NRS	9.04%	0.06 [-0.39 , 0.51]
Lambeek 2010	VAS	13.80%	0.66 [0.30 , 1.03]
Moix 2003	NA	3.32%	0.12 [-0.62 , 0.86]
Morone 2011	VAS	6.58%	1.38 [0.85 , 1.91]
Morone 2012	VAS	4.87%	1.34 [0.73 , 1.96]
Tavafian 2008	SF-36	10.37%	0.61 [0.19 , 1.03]
Tavafian 2011	SF-36	22.07%	0.41 [0.12 , 0.70]
Von Korff 2005	NRS	27.25%	0.20 [-0.05 , 0.46]
			0.56 [0.28 , 0.83]

favour control   favour experiment

Test of (residual) heterogeneity: Cochran's Q-test = 29.1 (df=8) ; p-value= 3e-04
$I^2 = 72.51$
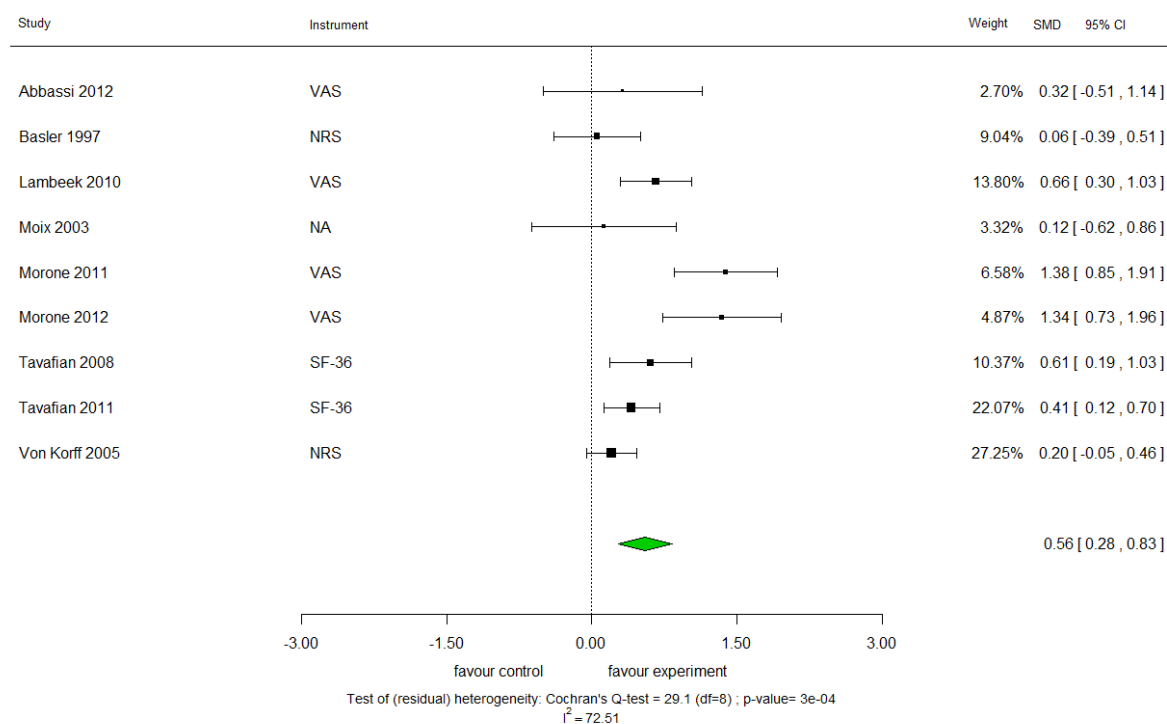
Figure 13. MA of SMD for "Multidisciplinary biopsychosocial rehabilitation versus usual care for back pain in short term".

Medium term multidisciplinary biopsychosocial rehabilitation versus usual care for back pain relief

Six studies, involving 740 patients, investigated the effect of a MBR intervention versus usual care for pain relief. The point estimate for the pooled between group difference is 0.60 SMD (95%CI 0.34, 0.87) in the medium term (Figure 14). The 95%CI did not include the zero, indicating a statistically significant effect in favor of MBR over usual care in terms of pain relief for LBP patients.
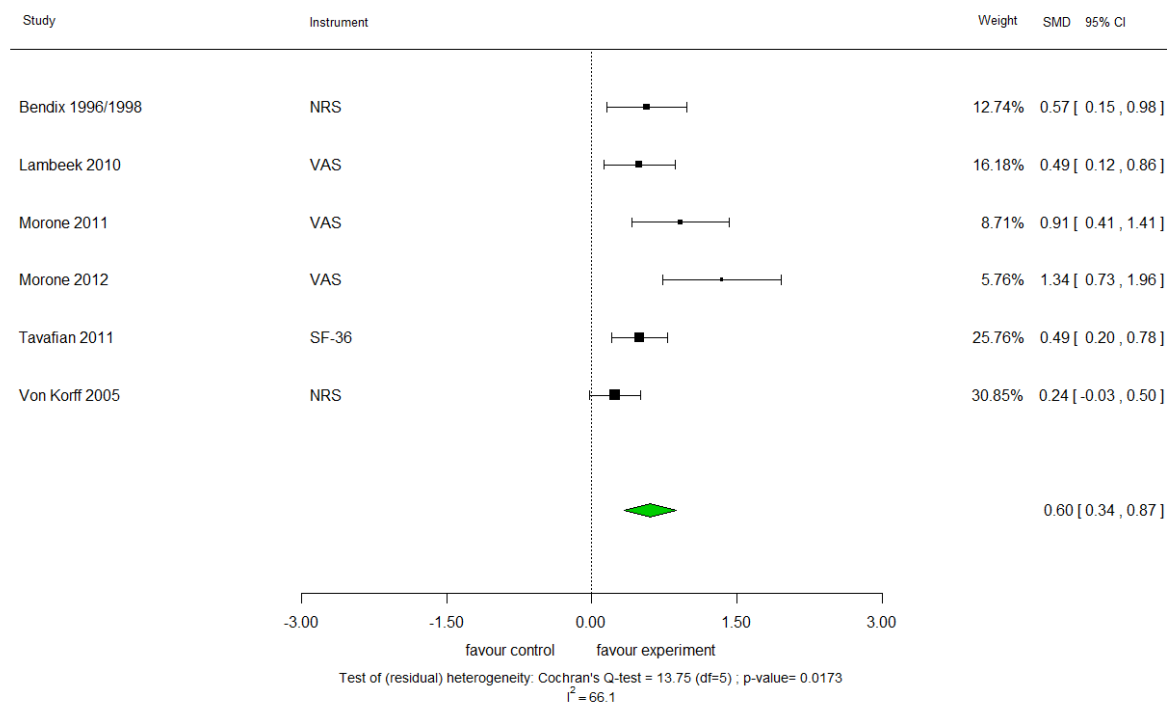
Figure 14. MA of SMD for "Multidisciplinary biopsychosocial rehabilitation versus usual care for back pain in medium term".

Long term multidisciplinary biopsychosocial rehabilitation versus usual care for back pain relief

Seven studies, involving 821 patients, investigated the effect of a MBR intervention versus usual care for pain relief. The point estimate for the pooled between group differences is 0.21 SMD (95%CI 0.04, 0.37) in the long term (Figure 15). The 95%CI did not include the zero, indicating a statistically significant effect in favor of MBR over usual care for pain relief in LBP patients.

Study	Instrument	Weight	SMD 95% CI
Abbassi 2012	VAS	2.87%	0.28 [-0.54 , 1.10]
Bendix 1996/1998	NRS	12.41%	0.23 [-0.17 , 0.62]
Lambeek 2010	VAS	14.99%	0.11 [-0.24 , 0.47]
Linton 2005	NRS	12.96%	0.52 [0.14 , 0.91]
Lukinmaa 1989	VAS	19.74%	-0.13 [-0.44 , 0.18]
Strand 2001	VAS	12.49%	0.26 [-0.14 , 0.65]
Von Korff 2005	NRS	24.55%	0.32 [0.04 , 0.60]
			0.21 [0.04 , 0.37]

favour control   favour experiment

Test of (residual) heterogeneity: Cochran's Q-test = 7.96 (df=6) ; p-value= 0.2407
$I^2 = 24.67$

Figure 15. MA of SMD for "Multidisciplinary biopsychosocial rehabilitation versus usual care for back pain in long term".

Results of meta-analyses in MID units

We firstly used the described standard deviation ratio method (SDR) (eq 2) to calculate the

MIDs (eq 3) for instruments without an anchor based one. The distribution of the SDR for

each instrument is described in table 11.

Table 11. Distribution of SDR for each instrument having an anchor based MID in the three comparisons.

Comparison: MBR vs Usual Care	Median SDR	Minimum of SDR	Maximum of
Short term (studies with established MID, n=6/9)			
NRS (n=2)	1.256	1.053	1.460
VAS 0-10 cm (n=4)	0.698	0.638	0.938
Overall (n=6)	0.826	0.638	1.460
Medium term (studies with established MID, n=5/6)			
NRS (n=2)	0.931	0.909	0.952
VAS 0-10 cm (n=3)	0.682	0.622	0.789
Overall (n=5)	0.789	0.622	0.952
Long term(studies with established MID, n=7/7)			
NRS (n=3)	0.909	0.769	0.952
VAS 0-10 cm (n=4)	0.732	0.560	1.071
Overall(n=7)	0.769	0.560	1.071

Short term multidisciplinary biopsychosocial rehabilitation versus usual care for back pain relief

The median SDR for VAS was 0.69 (range 0.64-0.94), whereas the median SDR for NRS was 1.25 (range 1.05-1.46). Combining both the overall median SDR was 0.82 (Table 3) and it was used to calculate the distribution based MIDs for the three studies using an instrument without an established MID. The pooled estimate of the effect was 0.75 MID units (95%CI; 0.27, 1.24; Figure 16). For clinicians this result become easy to translate into clinical practice. For example, since the MID for NRS is 2 points, the summary measure of 0.75 MID units means 1.5 scores of improvement in NRS scale.
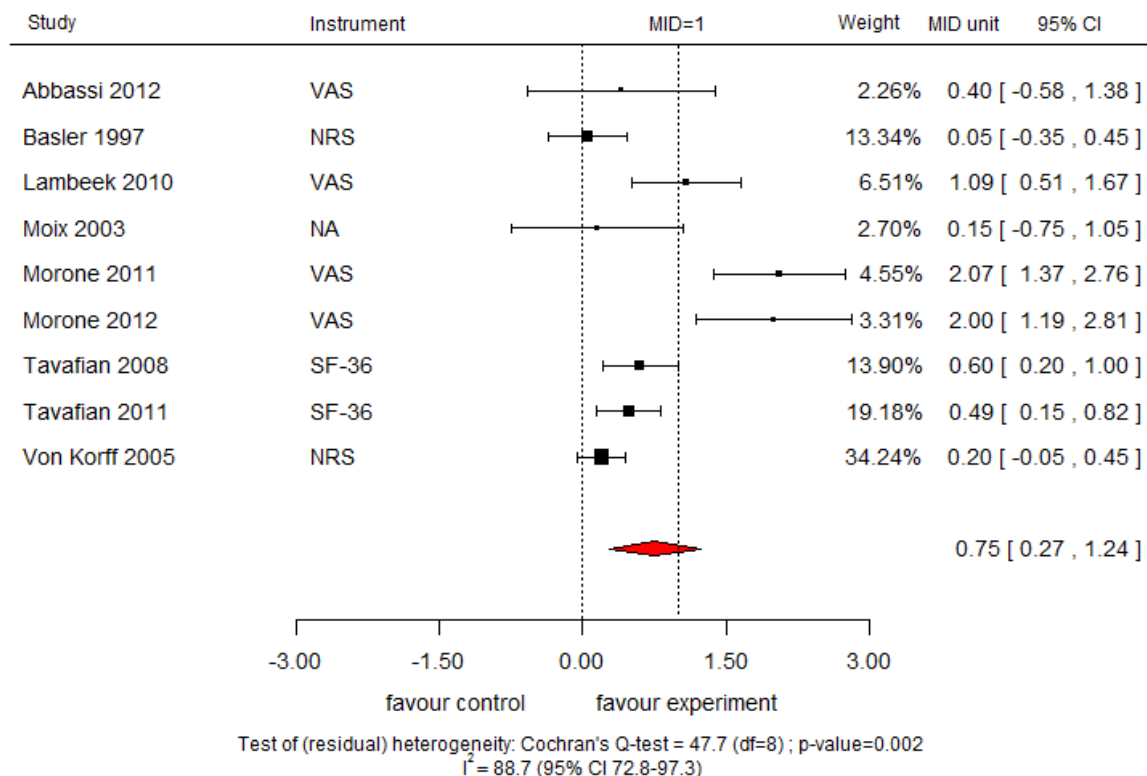
Figure 16. MAs in MID units for "Multidisciplinary biopsychosocial rehabilitation versus usual care for back pain in short term".

Medium term multidisciplinary biopsychosocial rehabilitation versus usual care for back pain relief

The median SDR for VAS was 0.68 (range 0.62-0.79), whereas the median SDR for NRS was 0.93 (range 0.91-0.95). Combining both, the overall median SDR was 0.79 (Table 3) and it was used to calculate the distribution based MIDs for one study using an instrument without an established MID. The pooled estimate of the effect was 0.86 MID units (95%CI; 0.39, 1.33; Figure 17). For clinicians this result become easy to translate into clinical practice. For example, since the MID for NRS is 2 points, the summary measure of 0.86 MID units means 1.7 scores of improvement in NRS scale.
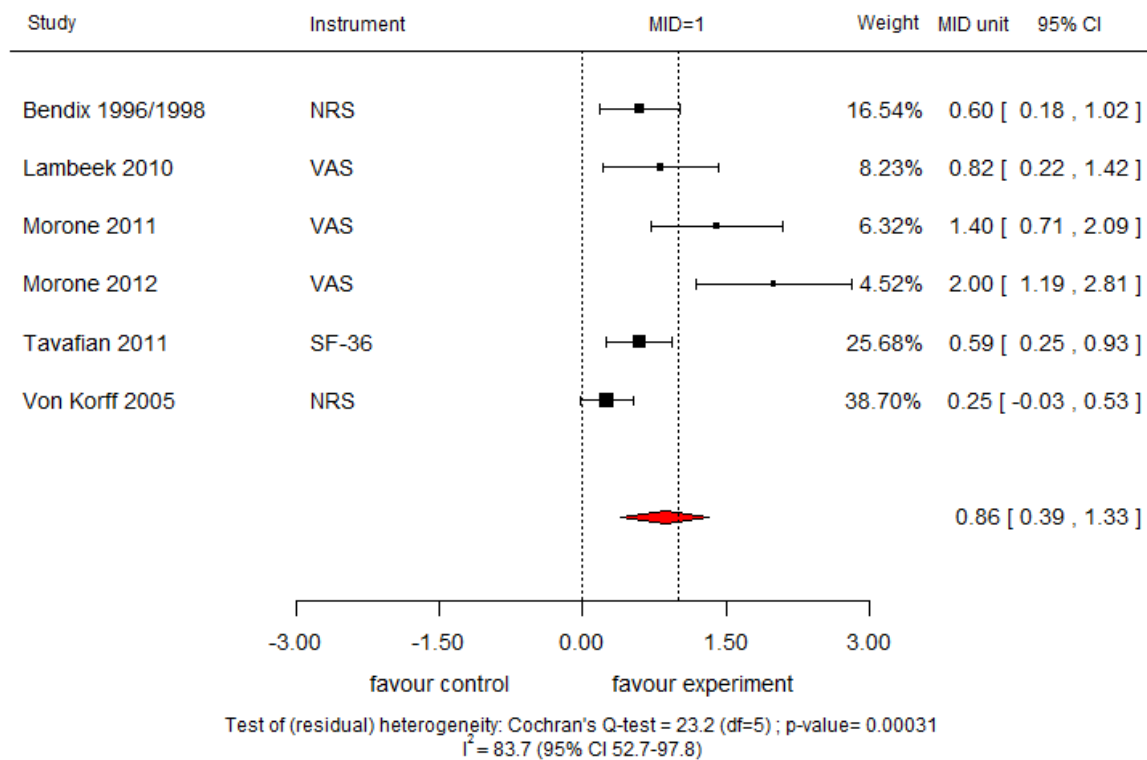
Figure 17. MA of MID units for "Multidisciplinary biopsychosocial rehabilitation versus usual care for back pain in medium term".

Long term multidisciplinary biopsychosocial rehabilitation versus usual care for back pain relief

The median SDR for VAS was 0.73 (range 0.56-1.07), whereas the median SDR for NRS was 0.91 (range 0.77-0.95). Combining both, the overall median SDR was 0.77 (Table 11). The pooled estimate of the effect was 0.27 MID units (95%CI; 0.07, 0.48; Figure 18). For clinicians this result become easy to translate into clinical practice. For example, since the MID for NRS is 2 points, the summary measure of 0.27 MID units means 0.5 scores of improvement in NRS scale.
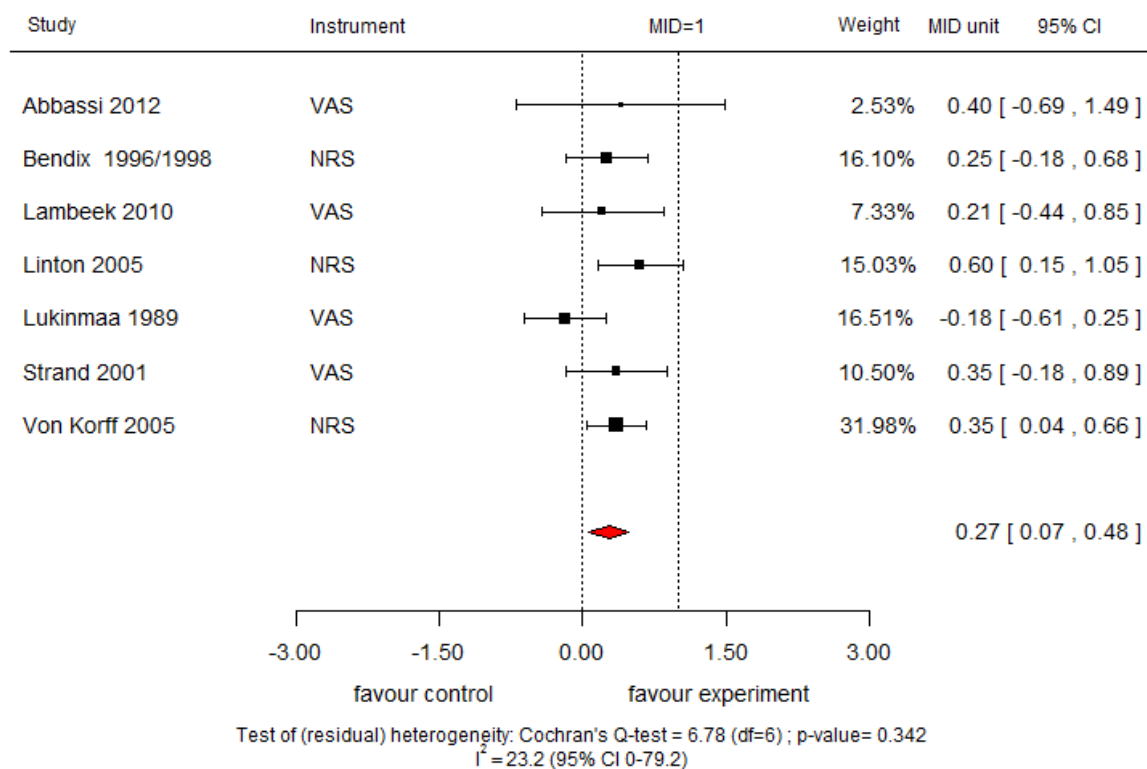
Figure 18. MA of MID units for "Multidisciplinary biopsychosocial rehabilitation versus usual care for back pain in long term".

Comparisons of results in SMD and MID units and clinical interpretation of results in MID units

To better compare results of the MAs coming from the two different units of measure, the summary estimates in SMD and MID units are reported in Table 12.

Table 12. Contrast of the results from the MAs in SMD and MID units for each considered comparison.

Comparisons	SMD (95% IC)	MID units (95% IC)
Short term MBR vs Usual Care	0.56 (0.28-0.83)	0.75 (0.27-1.24)
Medium term MBR vs Usual Care	0.60 (0.34-0.87)	0.86 (0.39-1.33)
Long term MBR vs Usual Care	0.21 (0.04-0.27)	0.27 (0.07-0.48)

The three overall estimates are statistically significant both in SMD and MID units MA. As expected, the results are consistent across units of measurement. Therefore, in conclusion there is a statistically significant difference between the effects of MBR vs. usual care at all FU times.

However, the overall estimates in SMD do not allow to immediately appreciate the clinical impact of an intervention, whereas those in MID units reflect also the clinical relevance. Concerning the interpretation of the results in MID units, Jonhstone et al. proposed the following guide: "if the pooled estimate is greater than 1 MID unit, and one accepts that the estimate of effect is accurate, many patients will gain clinically important benefits from treatment. If the estimate of effect lies between 0.5 and 1.0, the treatment may still benefit an appreciable number of patients. As the pooled estimate falls below 0.5 MID units it becomes

progressively less likely that an appreciable number of patients will achieve important benefits from treatment" [127].

In our MAs, all summary estimates in MID units are below 1, i.e. the average change from baseline to end of FU is smaller than the MID. However, two comparisons (short and medium term) are clinically relevant for a sub-group of the subjects, because the upper limit of the 95%CI is greater than one MID unit. On the contrary, the 95%CI of the third comparison never gets above one MID unit. We can conclude that the long-term effect is not clinically relevant (Figure 19).
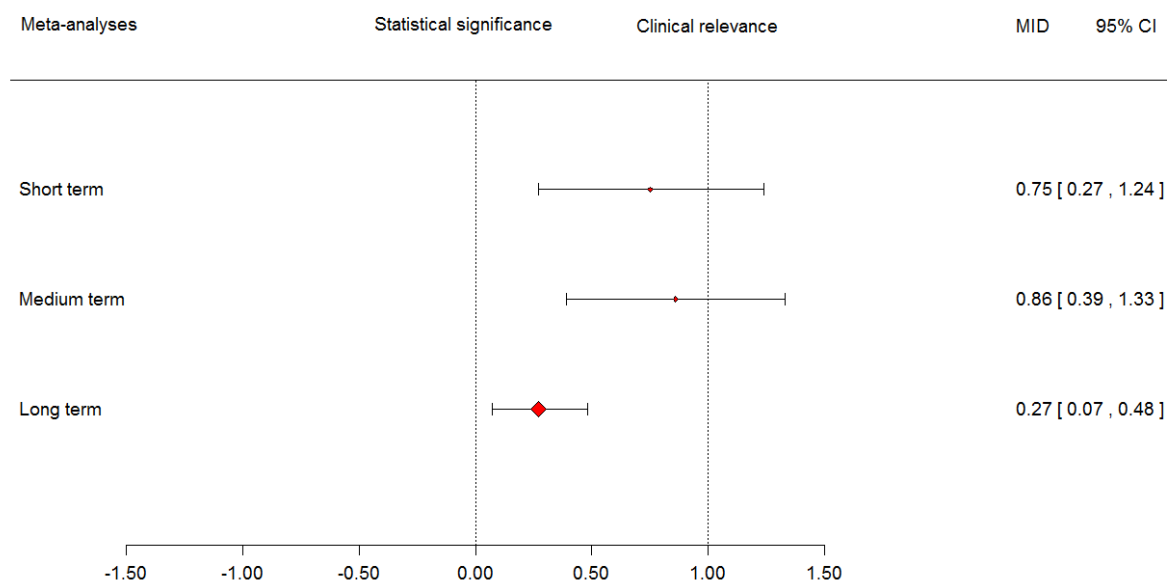
Figure 19. Clinical interpretation of the results of the MAs in MID units.

4. Discussion of clinical relevance in LPB interventions

Clinical relevance is not yet seriously contemplated in the efficacy consideration. In fact, only an half of the trials reported the source reference for the adopted planned MID. Despite the amount of efforts in conducting RCTs in LBP rehabilitation, just few of them reached both statistically and clinically significant results and the majority of their multi arm comparisons were compared against open-and-shut interventions. This finding assume a great importance because the efficacy of results in the RCTs usually is represented by the statistical significance. If authors of trials reported results in term of clinical relevance the threshold for efficacy will change and a considerable part of the interventions statistically significant become not efficacious. This concept was better investigated and developed by the analysis reported in the case study review of MBR for chronic pain.

Results of MAs in SMD and MID units change. Comparing MBR versus usual care for short and medium term pain relief suggest a statistically significant but clinically modest effect for patients, as the pooled estimate is slightly lower than 1 MID, which is by definition the minimal important clinical difference. In the long term comparison, the pooled estimated is only one third of the MID, suggesting that on average the benefit, although statistically significant, is not clinically relevant.

The MID unit approach maintains the advantage of SMD, i.e standardizing the outcome measures allowing comparisons across different scales, but it increases the interpretability of the results for clinicians and patients, because the improvement is measured in a clinically meaningful way. The use of MID units avoids the problems associated with heterogeneity of between study variances as a result of using the SD to calculate the SMD. The approach

described by [128] also suggests how to include studies without an established MID. This protect against the bias due to selection of the studies.

However, even when anchor based MIDs are available, application of the method in particular instances presents some limitations. The use of MID units requires previous studies having reported an estimate of the MID (possibly an anchor based MID) for several trials. Not all instruments used to asses an outcome have an established MID, in fact this is possible for only a little number of outcome measures. Specifically, in a large cohort of trials (n=185) focused on LBP rehabilitation we previously found 70 different instruments to measure pain [135] but only 5 of these have an anchor based MID.

Secondly, in the concept of MID is intrinsic that the clinical effect is related only to one intervention. The MID is informative only about the comparisons investigating the treatment versus control treatment (i.e., usual care, placebo etc.); if we compare two different treatments the MID value should be changed to account for the effect already provided by the control treatment. For example, in the comparison of MBR versus pharmacological treatment, we can't apply the same MID that we used against usual care (that we assumed having no effect) because pharmacological treatment will already have an effect on pain releif and the MID for the difference between experimental and control treatments should be adjusted to be a clinically meaningful increase beyond the effect of the control.

Finally, MAs in MID units are vulnerable to naïve, oversimplified interpretation as we estimate an average effect while the actual effect is different from patient to patient. However, this is not unique to the MID approach, but it is a common problem when we deal with means and consequently in all MAs, where the interpretation of the results is based on the average effect in the population. Also we have to consider that, when we define a MID, we choose a single value while in reality the MID is subjective, i.e. the clinical relevance of a change in the outcome may differ from patient to patient.

Despite its limitations, the approach to perform MA in MID units shows various advantages in terms of clinical interpretability and in avoiding selection bias. These characteristics suggest its use in MAs of continuous outcomes measured with different instruments, such as pain, to express results in readily interpretable way for rehabilitation stakeholders.

PART 3. Identifying the best available treatment for low back pain: a network meta-analysis

1. Introduction

Systematic reviews and meta-analysis in rehabilitation

The Cochrane Collaboration define a SR as "a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies. MA refers to the use of statistical techniques in a SR to integrate the results of included studies'' [136].

Also in the professionals allied to medicine MA rise as an efficient tool of keeping up-to-date with the accumulation of evidence in clinical content areas and as a background document for clinical practice guidelines is now gaining momentum [137]. The rehabilitation field is not an exception. The first SRs relevant to the rehabilitation field were published in the early 1980s [138]. Since then, the number of SRs relevant to rehabilitation field has substantially increased [138]. A recent survey confirms this secular trend. In the 1999, 59 SRs of physiotherapy were published [138] whereas in only nine months of 2011 the new publications targeted as 'systematic reviews' in Medline were 263 [23]. Two third of these publications were not clinical evidence synthesis but qualitative evaluations of key methodological dimensions, aiming at limiting the bias and improving the reliability and accuracy of results. Around one third of reviews included at least a MA and considered the consistency (heterogeneity) of results across studies. Only one quarter described the flow of information throughout the review process, reasons for exclusion of studies and discussed potential publication bias, as methodologically required [23].

In the rehabilitation field often many different treatments and interventions have been proposed and evaluated with RCTs for a single condition. In this setting, classical pairwise MA permit to synthesize evidence of the relative effect of two treatment at a time and not all treatments have been directly compared. To allow for an overall evaluation of many different treatment MA methodology has been extended to perform NMA.

Network meta-analysis: multiple comparisons

Network meta-analysis is an extension of standard pairwise MA by including multiple comparisons across a range of interventions, addressing the comparative effectiveness of multiple treatment alternatives [139].

Clinicians, patients, and health-policy makers often need to decide which treatment is "best" based on all relevant evidence. Unfortunately, robustly designed RCTs that simultaneously compare all interventions of interest are almost never available. As an alternative, indirect treatment comparisons can provide useful evidence [140].

In NMA, evidence on the relative treatment effect of two treatments is generated by considering the entire evidence in the network, coming from both direct and indirect treatment comparisons. This provides an advantage over the more traditional pairwise analyses. Indirect comparisons can be performed if studies in a SR provide information on three or more competing interventions. For example, suppose there are studies directly comparing 'chiropractic' (A) with 'manual therapist' (B) in providing therapy for restoration of human functions, and studies comparing 'chiropractic' with 'physiotherapist' (C). Suppose further that these have been combined in standard, pairwise MAs to separately derive direct estimates of AB and AC intervention effects, measured as MD in restoration of human function. If there are no head-to-head trials directly comparing interventions 'manual therapist' (B) and 'physiotherapist' (C) it

can derive an indirect estimate of the relative effect of BC by combining the two summary

estimates AB and AC (Figure 20).
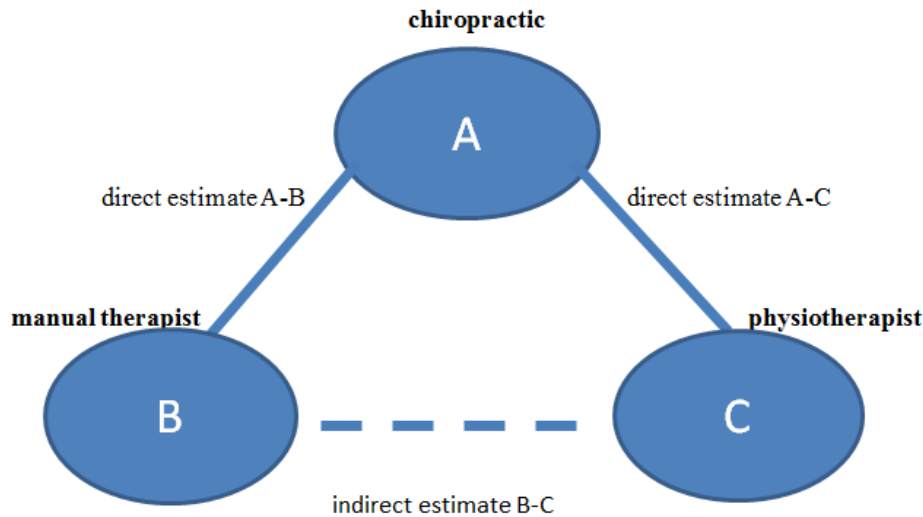
Figure 20. Example of a SR with studies comparing the effectiveness of 'chiropractic' (A), 'manual therapist' (B) and 'physiotherapist' (C) in providing therapy for restoration of human functions.

Anyway, indirect estimates can be derived via many other routes. The only requirement is that

two interventions are 'connected' and not necessarily via a common comparator. An example

of this situation is provided in Figure 21, where 'manual therapist' (B) and 'physician' (D) do

not have a common comparator, but we can compare them indirectly via the route 'manual

therapist' (B) – 'chiropractic' (A) – 'physiotherapist' (C) – 'physician (D).
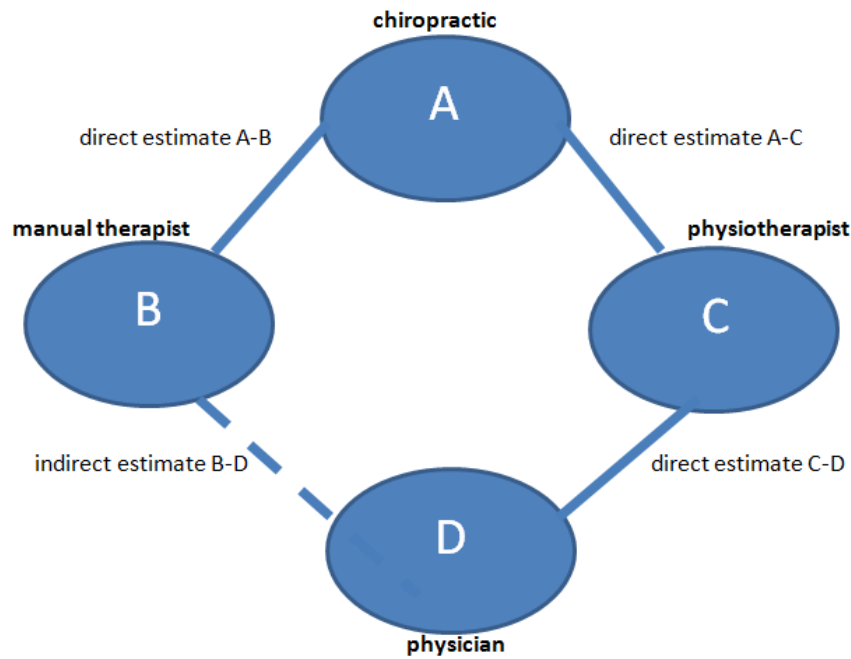
Figure 21. Example of a SR with studies comparing the effectiveness of 'chiropractic' (A), 'manual therapist' (B), 'physiotherapist' (C) and 'physician' (D) in providing therapy for restoration of human functions.

Assumptions of network meta-analysis

NMA requires some assumptions to estimate indirect effects: *similarity or transitivity, consistency and heterogeneity*. First, *similarity or transitivity* means that studies should be combined only if they are clinically and methodologically similar [141]. Of course, studies comparing different interventions are likely to differ in a wide range of characteristics. Sometimes these characteristics are associated with the outcome of interest in the sense that different levels of a particular characteristic may influence the effect of an intervention. If the AB and AC trials differ with respect to such characteristics, also called 'effect modifiers', they are not appropriate for an indirect comparison [142]. The underlying assumption of indirect comparisons is that the common comparator intervention A allows a transitive relationship

between the AB and AC effects. In words, this means that we can compare interventions B and C via A.

Transitivity requires that intervention A is similar when it appears in studies AB and studies AC with respect to clinical characteristics that may affect the two relative effects [143]. Anyway, similarity is not required for all characteristics of trials and patients because, even if studies are dissimilar, not all characteristics are effect modifiers [144].

The secondarily assumption is *consistency,* that is the statistical expression of transitivity, that implies that different sources of evidence (direct and indirect) should be in agreement and underlies any mixed estimate [145,146]. A mixed estimate is an inverse variance weighted average of the direct and indirect summary estimates. The consistency assumption is expressed mathematically by the consistency equations, which suggest that the true direct and indirect intervention effects for a specific comparison are similar:

$$'true'\ indirect\ MD(BvsC)='true'\ direct\ MD(AvsC)-'true'\ direct\ MD(AvsB)$$

It can be evaluated only when a loop in the evidence network exists, that is, when there is a direct and indirect evidence for a particular comparison of interventions. In addition of consistency, which is specific of NMA, we have to consider the heterogeneity of the treatment effect as we do in pairwise MA. In NMA we usually assume that each study estimates a study-specific treatment effect, or that there is heterogeneity in each pairwise comparison, and that the effects have a normal distribution around a common mean with variance τ^2. Consequently we perform a random-effects analysis [147]. The third assumption which is often made in NMA, and that it will be used in our analysis, is that *all pairwise-comparison treatment effects share the same heterogeneity τ^2.*

In practice, one must check these assumptions, to the extent possible. Particularly, similarity or transitivity can be investigated exploring covariates distribution between studies, and

different tests are available for consistency [141]. Clinical and methodological diversity, as well as variation in ES, will be verified in order to decide the inclusion of studies and their relative interventions for the contribution of network [141]. Agreement is seldom perfect, and both statistical and clinical judgment may be required (e.g., re-examining information in the reports on some trials, calculating direct and indirect estimates separately before proceeding to a NMA). Anyway, some differences on study-level characteristics can be controlled by adjustments via meta-regression, but these are unlikely to overcome substantial disparities among the studies. Interpretations of results should acknowledge this limitation [141].

Open questions and aims in LBP rehabilitation interventions

There are many different therapeutic interventions for acute non-specific LBP, including pharmacological and physiotherapy treatments, but none of them is universally accepted [6]. According to a recent SR published by the Ontario Protocol for Traffic Injury Management Collaboration [148], investigating high quality guidelines, patients with acute LBP should be encouraged to return to activity and may benefit from paracetamol, non-steroid anti-inflammatory drugs (NSAIDs), or spinal manipulation. However, the uncertainty about the most effective treatment may be due to the absence of multiple direct comparison of the different treatments available. In fact, the majority of the available studies contrasts only two interventions at a time. It would be helpful for clinicians, patients and all stakeholders to know the relative efficacy of all available treatments for acute LBP in terms, to better choose among the different options on the basis of evidence and not only according to expert's opinion.

2. Case example of NMA: Effectiveness of treatments for acute and sub-acute mechanical non-specific low back pain

Aim

To assess the effectiveness of currently available interventions used to treat acute non-specific LBP. Particularly, to compare different treatment options for relieving pain.

Methods

The PRISMA-NMA extension statement was used to structure the contents of the actual SR and NMA [149]. Additional sections were considered according to Chaimani et al.[150].

Eligibility criteria

Types of studies

RCTs were included only if authors had explicitly stated that it was randomised [151]. Quasi-randomised trials and cross-over trials were be excluded.

Participants

Trials that involved participants aged older than 18 years, both male and female, experiencing pain until 12 weeks of non-specific LBP were included. The LBP population was classified based on timing onset of pain: acute (less than six weeks) or subacute (six to 12 weeks) [4]. Accordingly, trials for pain duration were selected regardless of the definition of population declared for a study (e.g., chronic patients having pain less than 12 weeks). When the recruitment criteria for duration of pain exceeded 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 in the absence of clarification, the study was excluded. According to the definition of aspecific LBP, studies focusing on specific pathological entities (e.g., spondylolisthesis) and subgroups of patients (e.g., pregnant women) were excluded. 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 cannot be identified, or patients present with both conditions, were excluded.

Interventions

All conservative rehabilitation or pharmacological treatments provided by health professionals, such as general medical practitioners or physiotherapists, aiming at relieving pain and reducing physical disability were considered. Any modality (e.g. physical, pharmacological), treatment extent, frequency or intensity was took into consideration. RCTs or arms of RCTs including non-conservative treatments (e.g., surgical approaches) or alternative-medicine treatments as herbal medicine and homeopathy were excluded. Acupuncture and dry needling were comprised because they are largely used and scientifically studied in the Western countries [152].

Comparators

All comparators which have been used in the included trials were encompassed. However, for the purpose of the NMA, a common comparator defined as "placebo" was defined, including no treatment, sham intervention and placebo therapies.

Outcomes and study time-points

Primary outcomes were pain intensity (e.g., measured by NRS, visual analogue scale, McGill Pain Questionnaire or, box scale, other validated quantitative measures). The effects were evaluated at short-term (within 1 month from the end of treatment), medium-term (around 6 months) and long-term (around 12 months) of FU assessment.

Information sources

The following electronic databases since the inception date up to 29 November 2017 were investigated: MEDLINE (PubMed), CENTRAL, EMBASE (Elsevier, EMBASE.com).

Appropriate thesaurus and free-text terms were applied to perform the search strategy. Investigators and relevant trial authors seeking information about unpublished data were contacted. The reference lists of all studies and any interesting SR or MA identified during the search process were examined.

No restriction on language or publication period was applied. Non-English studies for which a translation could not be obtained were classed as potentially eligible awaiting classification but not considered into full review.

Study selection

Two independent authors screened the abstracts of all publications that are obtained by the search strategy. Then, they independently assessed the full text of potentially relevant studies for inclusion. All studies that did not fulfil the above inclusion criteria were discarded. Then the full text of the remaining articles was obtained. Disagreements were resolved through discussion. Anyway, a third author was consulted if disagreement persisted. Coevidence software [153] was employed to manage the study selection phase.

Data extraction

A data extraction form was specifically designed and piloted for 20 random trials using an Excel spreadsheet (Microsoft Inc.). Two researchers independently extracted characteristics and outcome data of the included studies. Disagreements were resolved through discussion or with assistance from a third author if necessary.

From each included study the following general characteristics were extracted: name of the first author, year of publication, setting, number of centers, number of randomized participants, the interventions compared for pain outcome. All relevant arm level outcome data were extracted (e.g. mean and SD of pain and number of reported patients at each FU).

The following characteristics were extracted as potential effect modifiers: age, gender, stage of disease duration (acute, subacute or both LBP), year of publication, duration of the treatment, duration of FU assessment and risk of bias (RoB) assessment.

It was assumed per transitivity that any patient that meets the inclusion criteria is, in principle, equally likely to be randomized to any of the eligible LBP intervention.

Risk of bias within individual studies

Two pairs of review authors independently assessed the RoB of the included studies. Disagreements were resolved through discussion or arbitration of a third review author when consensus could not be reached. The RoB was assessed for each included study using the Cochrane Collaboration criteria [151]: 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 any information was reported. To summarize the RoB overall 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 were met; 'high risk of

bias' when at least one criterion was unmet; and 'moderate risk of bias' in the remaining cases. Allocation concealment, blinding of outcome assessment, and incomplete outcome data were not expected to vary in importance across the pain intensity outcome, and therefore we summarized the RoB of each study.

Measurement of treatment effects and missing data

The ES of continuous pain data was based on SMD. Whenever possible, the final treatment mean score was used for calculation of the ES. Pain was analyzed with the total number of randomly assigned participants as denominators for final values.

If a small number of studies provided insufficient information SD final values, these may be substituted with the mean values of SD obtained from all other reported studies investigating the same outcome measurement [154]. Anyway, if all the required data cannot be extrapolated from the published article, attempts were made to contact the authors for additional information. Where this was not possible, studies were only described and NMA performed only for available data.

Assessment of transitivity and geometry of the network

The assumption of transitivity was assessed comparing the distribution of the above potential effect modifiers across the various pairwise comparisons. Then, the network of treatments based on the characteristics of the available studies was presented and evaluated graphically. In particular, the following aspects were evaluated: if there is a sufficient number of comparisons in the network with available direct data; if there is a high number of comparisons based on a single study; if there is any "closed loop" which allows testing agreement between direct and indirect estimates for comparison in the network; if any key treatment is missing;

and if the possible lumping of treatments is minimizing the clinical relevance of the review. Those information/consideration were used to assess the feasibility of a NMA.

Multi-arm trials comparing three or more interventions were included. All the comparisons in which an intervention presented multiple co-interventions for the experimental group (e.g., mixed treatment: manipulation plus exercise versus waiting list controls) or for the control group (e.g., physiotherapy defined as, laser therapy, some physical exercise plus drugs taken as needed) were not considered to avoid inconsistency across trials. For these reasons the node *usual care* was admitted only if it was defined as usual management of the general medical practitioners (minimal intervention such as advice and/or drugs taken as needed).

Data synthesis

Direct estimates

For every pairwise comparison between interventions, the SMD was calculated as the ES for pain scale measurements with its 95%CI.

To evaluate the direct comparisons, we firstly performed conventional pairwise MAs for pain reduction using a random-effects model to incorporate the assumption that the different studies were estimating different, yet related, treatment effects [126].

Statistical heterogeneity was investigated for each pairwise comparison by visual inspection of the relative forest plot as well as the I^2 statistic, which provides an estimate of the percentage of variability due to the heterogeneity rather than a sampling error (Higgins 2009 handbook). For comparison with more than one study, we rated the presence of statistical heterogeneity as follows: 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 [155].

Indirect and mixed estimates

To estimate indirect and mixed comparisons, a NMA was performed using a random-effects model within a frequentist setting fitting a multivariate normal model [156] and summarizing the results using ESs and their credible intervals (CrI). This model was initially developed to synthesize jointly multiple outcomes and in NMA was adopted handling different comparisons as different outcomes.

Assuming a common estimate for the heterogeneity variance across the different comparisons, correlations induced by multi-arm studies were accounted [157,158].

Anyway, these models were enable us to estimate the probability that each intervention will be at each possible rank for each outcome, given the relative ESs as estimated in the NMA. The ranking probabilities for all treatments at each possible rank for each intervention for pain outcome, given the relative ESs estimated in the NMA, were estimated using the *mvmeta* command setting up to 8780 draws. Hierarchy using the Surface Under the Cumulative Ranking Curve (SUCRA) and mean ranks were obtained. SUCRA can also be expressed as a percentage of a treatment that can be ranked first without uncertainty [159].

Assumptions when heterogeneity is estimated

Since a few studies were expected to be included in each direct comparison (maximum four), a common heterogeneity variance for all direct comparisons was assumed in standard pairwise MAs. Then, a common estimate for the heterogeneity variance across different comparisons was also assumed in NMA.

The presence of heterogeneity was statistically assessed for all direct pairwise comparisons using common τ^2 and I^2 statistics.

Assessment of statistical heterogeneity in the entire network was based on the magnitude of the heterogeneity variance parameter (τ^2) estimated by using NMA models [160].

Inconsistency assessment

Any patient who met the inclusion criteria, in principle, was equally randomized to any of the eligible interventions. The model with and without the consistency assumptions were fitted [161]. Inference about the presence of inconsistency from any source in the entire network was based on the design-by-treatment model proposed by Higgins [162]. This method accounts for different sources of inconsistency that can occur when studies with different designs (two-arm trials versus three-arm trials) give different results and when there is disagreement between direct and indirect evidence. Using this approach we first tested inconsistency globally, testing all the inconsistency parameters using a global Wald test statistic, which under consistency follows a χ^2 distribution [156]. Inconsistency and heterogeneity are interwoven; to distinguish between these two sources of variability, I^2 for inconsistency was employed to measure the percentage of variability that cannot be attributed to random error or heterogeneity (Jackson 2014).

To evaluate the presence of local inconsistency, we then used a loop-specific approach. This method evaluates the consistency assumption in each closed loop of the network separately, as the difference between direct and indirect estimates for a specific comparison (inconsistency factor) [163]. Then, the magnitude of the inconsistency factors and their 95%CI were used to make inference about the presence of inconsistency in each loop. Assuming a common heterogeneity estimate within each loop, results were presented graphically in a forest plot using the *ifplot* command.

However, to separates direct' and 'indirect' trial-level evidence for a particular comparison (or node), a *node-splitting* method was applied and the results were compared [161]. To estimate the inconsistency, the difference between indirect and direct estimates were calculated whenever indirect estimates could be constructed with a single common comparator [143]. Inconsistency was defined as disagreement between direct and indirect evidence with a 95%CI of the difference excluding 0.

In case of significant inconsistency, the distribution of clinical and methodological variables that were suspected might be potential sources of either heterogeneity or inconsistency in every comparison-specific group of trials was investigated.

Finally, the comparative efficacies between all treatments versus "placebo" as reference were estimated. All analysis were performed in Stata-IC [164] using the *network* and *intervalplot* commands [165-167,156]. The graphic method proposed by Salanti et al. was adopted for the interpretation of the results from the NMA [159].

Preliminary results

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 545 records of which 512 did not meet the inclusion criteria. Within this group, 252 records included a different population (e.g, chronic pain), 138 different interventions (e.g., mixed treatments) or comparisons (e..g, exercise versus exercise), 21 different outcomes (e.g., cost-effectiveness), 40 non-RCT studies, 14 were further duplicates and 25 studies are still awaiting assessment (not able to retrieve full-text or awaiting translation). Finally 36 studies were included. For a further description of screening process and reasons for exclusion of trials, see the study flow diagram in Appendix 4.

General characteristics

A total of 36 trials were included accounting for 7111 participants. The sample size varied between 87 and 219 participants (median 117). The majority of studies involved acute (n=19 trials) whereas a minority only subacute (n=9) or both acute and subacute patients (n=8). Twenty trials were multi-centric trials and 16 were single-center trials. Twenty-one of the trials

recruited participants from a Hospital or Clinic, 5 from an academic hospital, 4 from private practice, 1 in mixed setting and in 5 it was not stated. The median year of publication of the trials was 2003 (IQR 2000-2010). Table 13 summarized the characteristics of included studies. Because of the paucity of trials reporting findings in medium and long-terms, we analyzed only the network related to the short- term FU of treatments effect.

Table 13. General characteristics of included trials.

StudyID	Author	Year	Setting	Type of trial	Population- stage of disease
1	Casale[168]	1988	Hospital or Clinic	Monocenter	Acute LBP (less than 6 weeks)
2	Eken[169]	2014	Accademic hospiatal	Monocenter	Acute LBP (less than 6 weeks)
3	Ketenci[170]	2005	Accademic hospiatal	Monocenter	Acute LBP (less than 6 weeks)
4	Hagen[171]	2010	Hospital or Clinic	Monocenter	Subacute LBP (6-12 weeks)
5	Hindle[172]	1972	Not stated	Monocenter	Acute LBP (less than 6 weeks)
6	Hsieh[173]	2002	Hospital or Clinic	Monocenter	Mixed (acute and subacute)
7	Berry[174]	1988	Not stated	Monocenter	Acute LBP (less than 6 weeks)
8	Jallema[175]	2005	Hospital or Clinic	Multicenter	Mixed (acute and subacute)
9	Li[176]	2008	Hospital or Clinic	Multicenter	Subacute LBP (6-12 weeks)
10	Linton[177]	2000	Hospital or Clinic	Monocenter	Subacute LBP (6-12 weeks)
11	Little[178]	2001	Hospital or Clinic	Multicenter	Subacute LBP (6-12 weeks)
12	Machado[179]	2010	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
13	Malmivaara[180]	1995	Mixed	Multicenter	Acute LBP (less than 6 weeks)
14	Mayer[181]	2005	Hospital or Clinic	Multicenter	Mixed (acute and subacute)
15	Moffett[182]	1999	Hospital or Clinic	Multicenter	Subacute LBP (6-12 weeks)
16	Nadler[183]	2002	Hospital or Clinic	Multicenter	Subacute LBP (6-12 weeks)
17	Nadler[184]	2003	Hospital or Clinic	Multicenter	Subacute LBP (6-12 weeks)
18	Postacchini[185]	1988	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
19	Pozo-Cruz[186]	2012	University	Multicenter	Subacute LBP (6-12 weeks)
20	Rabin[187]	2014	Private practice	Multicenter	Mixed (acute and subacute)
21	Ralph[188]	2008	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
22	Roberts[189]	2002	Private practice	Multicenter	Acute LBP (less than 6 weeks)
23	Schneider[190]	2015	Accademic hospiatal	Monocenter	Mixed (acute and subacute)
24	Seferlis[191]	1998	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
25	Shin[192]	2013	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
26	Staal[193]	2004	Private practice	Monocenter	Mixed (acute and subacute)
27	Storheim[194]	2003	Not stated	Monocenter	Subacute LBP (6-12 weeks)
28	Szpalski[195]	1994	Not stated	Monocenter	Acute LBP (less than 6 weeks)
29	Takamoto[196]	2015	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
30	Wand[197]	2004	Hospital or Clinic	Monocenter	Acute LBP (less than 6 weeks)
31	Whitfill[198]	2010	Not stated	Monocenter	Mixed (acute and subacute)
32	Lindstrom[199]	1995	Private practice	Monocenter	Mixed (acute and subacute)
33	Dreiser[200]	2003	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
34	Hasegawa[201]	2014	Accademic hospiatal	Monocenter	Acute LBP (less than 6 weeks)
35	Williams[202]	2014	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)
36	Tuzun[203]	2003	Hospital or Clinic	Multicenter	Acute LBP (less than 6 weeks)

Risk of bias in included studies

Figure 22 and Figure 23 summarized the RoB assessments. Regarding the overall RoB across studies (n=36), only 7 trials were at low RoB (19.5%). We categorized 55.5% of the studies as at unclear RoB (n=20) and 25% at high RoB (n=9).
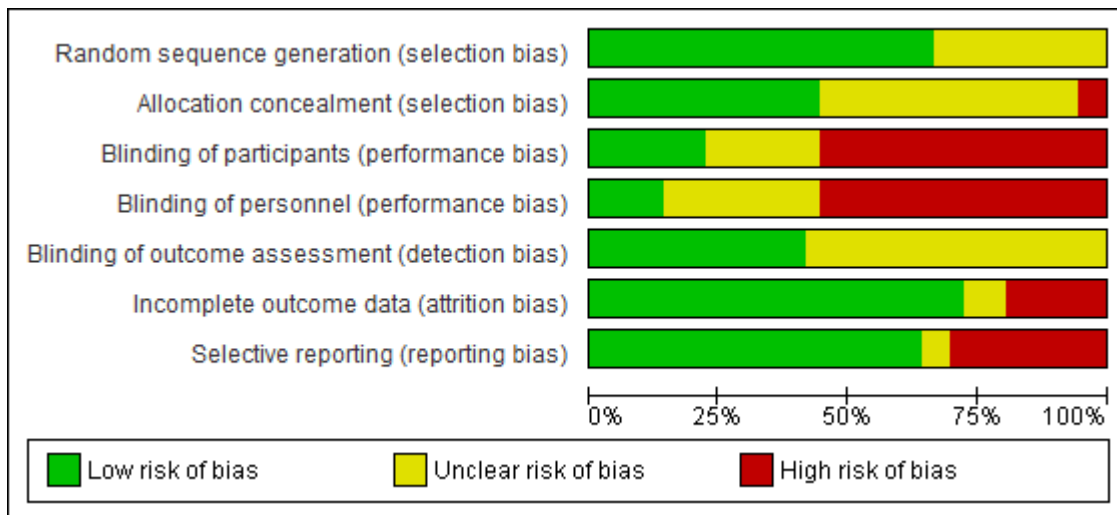
Figure 22. Percentage of studies for the overall RoB.

Study	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	?	?	?	?	?	+	−
Roberts 2002	+	−	−	−	?	+	−
Schneider 2015	+	+	−	−	+	+	−
Seferlis 1998	?	?	−	−	?	−	+
Shin 2013	+	+	−	−	+	+	+
Staal 2004	+	+	−	−	+	+	+
Storheim 2003	+	+	−	−	+	−	+
Szpalski 1994	?	?	?	?	?	+	+
Takamoto 2015	+	?	−	−	+	−	−
Tuzun 2003	+	+	+	?	+	+	+
Wand 2004	+	+	−	−	+	−	+
Whitfill 2010	?	?	?	?	?	?	+
Williams 2014	+	+	+	+	+	+	+

Figure 23. RoB assessment for each study included in the NMA.

Effect of interventions on pain intensity at short term FU

Twenty-one RCTs reported the outcome at short term FU and included all necessary data. The majority of patients were defined as acute LBP except for 4 studies having mixed population (acute and subacute) and 4 exclusively in subacute phase. Median mean age was 38 years (min-max: 34-41) and the average median proportion of men was 40%. Those figures were very similar across all interventions.

Many different treatments were evaluated in the included studies. We included in a single node cognitive exercise and *exercise*, educational interventions (such as provided booklets) were included in the *usual care* node. We defined *placebo,* including both pharmacological placebo as well as sham therapies, as the common comparator against all active treatments.

The duration of treatment ranged from 1 day to 7 weeks. The considered FU assessment was always the first after the end of treatment. However, because of the different duration of therapies, it encompassed different time-points. Six out of 21 studies (29%) were at low RoB, 4 (19%) at high risk and 11 (52%) at unclear RoB.

Among information collected from the trials, the variables considered as potential effect modifiers are summarized in table 14 (Year, Stage of LBP, Length of treatment, Week of FU, Pain instrument, ROB). In general, similarity of trials characteristics was guaranteed in terms of clinical and methodological features. Influence of the week of FU assessment was then investigated in the sensitivity analysis.

Table 14. Variables investigated as effect modifiers for transitivity assessment for included

arms.

ID	Author	Year	Stage of LBP	Treatments	length of treatment	Week of FU	Pain instrument	N	RoB
1	Eken[169]	2014	Acute LBP (less than 6 weeks)	1.oppioid 2.paracetamol 3.fans	1 day	1 day	VAS 0-100	137	Unclear
2*	Hsieh[173]	2002	Mixed (acute and subacute)	1.manual therapy 2.manual therapy 3.exercise	3 weeks	3 weeks	VAS 0-10	148	Unclear
3	Li[176]	2008	Subacute LBP (6-12 weeks)	1.myorelaxant drug 2.oppioid	1 week	1 week	NRS-10	220	Unclear
4	Machado[179]	2010	Acute LBP (less than 6 weeks)	1.usual care 2.exercise	3 weeks	3 weeks	VAS 0-10	146	Low
5	Malmivaara[180]	1995	Acute LBP (less than 6 weeks)	1.exercise 2.usual care	Not stated	3 weeks	VAS 0-10	119	Low
6	Mayer[181]	2005	Mixed (acute and subacute)	1.heat wrap 2.exercise 3.education	5 days	1 week	6 points scale	76	Unclear
7	Moffett[182]	1999	Subacute LBP (6-12 weeks)	1.exercise 2.usual care	4 weeks	6 weeks	Aberdeen back pain scale 0-100	187	Unclear
8	Nadler[183]	2002	Subacute LBP (6-12 weeks)	1.paracetamol 2.NSAIDs	2 days	4 days	6 points scale	371	Unclear
9	Nadler[184]	2003	Subacute LBP (6-12 weeks)	1.heat wrap 2.placebo	3 days	5 days	6 points scale	191	High
10	Rabin[187]	2014	Mixed (acute and subacute)	1.exercise 2.manual therapy	7weeks	8 weeks	NRS 0-10	105	High
11*	Schneider[190]	2015	Mixed (acute and subacute)	1.manual therapy 2.manual therapy 3.usual care	4 weeks	4 weeks	VAS 0-10	114	Low
12	Szpalski[195]	1994	Acute LBP (less than 6 weeks)	1.NSAIDs 2.placebo	2 weeks	2 weeks	VAS 0-10	73	Unclear
13*	Takamoto[196]	2015	Acute LBP (less than 6 weeks)	1.manual therapy 2.placebo 3.massage	2 weeks	1 month	VAS 0-100	63	High
14	Wand[197]	2004	Acute LBP (less than 6 weeks)	1.exercise 2.usual care	6 weeks	6 weeks	VAS 0-10	102	High

15*	Williams[202]	2014	Acute LBP (less than 6 weeks)	1.paracetamol 2.paracetamol 3.placebo	4 weeks	4 weeks	VAS 0-10	1641	Low
16	Tuzun[203]	2005	Acute LBP (less than 6 weeks)	1.myorelaxant 2.placebo	5 days	5 days	VAS 0-100	143	Low
17*	Dreiser[200]	2003	Acute LBP (less than 6 weeks)	1. NSAIDs 2. NSAIDs 3. placebo	1 week	1 week	VAS 0-100	369	Unclear
18*	Ketenci[170]	2005	Acute LBP (less than 6 weeks)	1.myorelaxant 2.myorelaxant 3.placebo	1 week	1 week	VAS 0-100	97	Unclear
19	Berry[174]	1988	Acute LBP (less than 6 weeks)	1.myorelaxant 2.placebo	1 week	1 week	VAS 0-100	113	Unclear
20	Hasegawa[201]	2014	Acute LBP (less than 6 weeks)	1.acupuncture 2.placebo	1 week	1 week	VAS 0-10	80	Unclear
21	Shin[192]	2013	Acute LBP (less than 6 weeks)	1.acupuncture 2.NSAIDs	1 week	2 weeks	NRS 0-10	58	Low

*Studies including at least 2 out of 3 arms with the same intervention have been managed following the Cochrane Handbook [154]. They contributed to the network accounting for a total number of 27 studies. Full data are available in Appendix 5.

Network plot

Figure 24 shows a graphical representation of the network. Nodes and edges were weighted according to the number of studies including the respective treatments and comparisons. Pain intensity data were available at 1 month of FU for an overall of 10 interventions.

The following treatments were identified: usual care, non-steroids anti-inflammatory drugs, acupuncture, exercise, heat wrap, manual therapy, muscle relaxant, opioid, paracetamol, placebo. The most studied comparison was the placebo (n=13), followed by exercise (n=8), manual therapy (n=7), usual care (n=7), NSAIDs (n=6), muscle relaxant drug (n=5), paracetamol (n=4), heat wrap (n=3), acupuncture and opioid (n=2 each one). Figure 25 shows the network according to the RoB assessment. There are two comparisons at low RoB (paracetamol vs placebo and manual therapy vs usual care) and two at high risk (manual therapy vs placebo and heat wrap vs placebo). All the remaining are unclear.

Figure 24. Network plot.

The size of the nodes is proportional to the number of studies evaluating each intervention, and the thickness of the edges is proportional to the precision (the inverse of the variance) of each direct comparison.
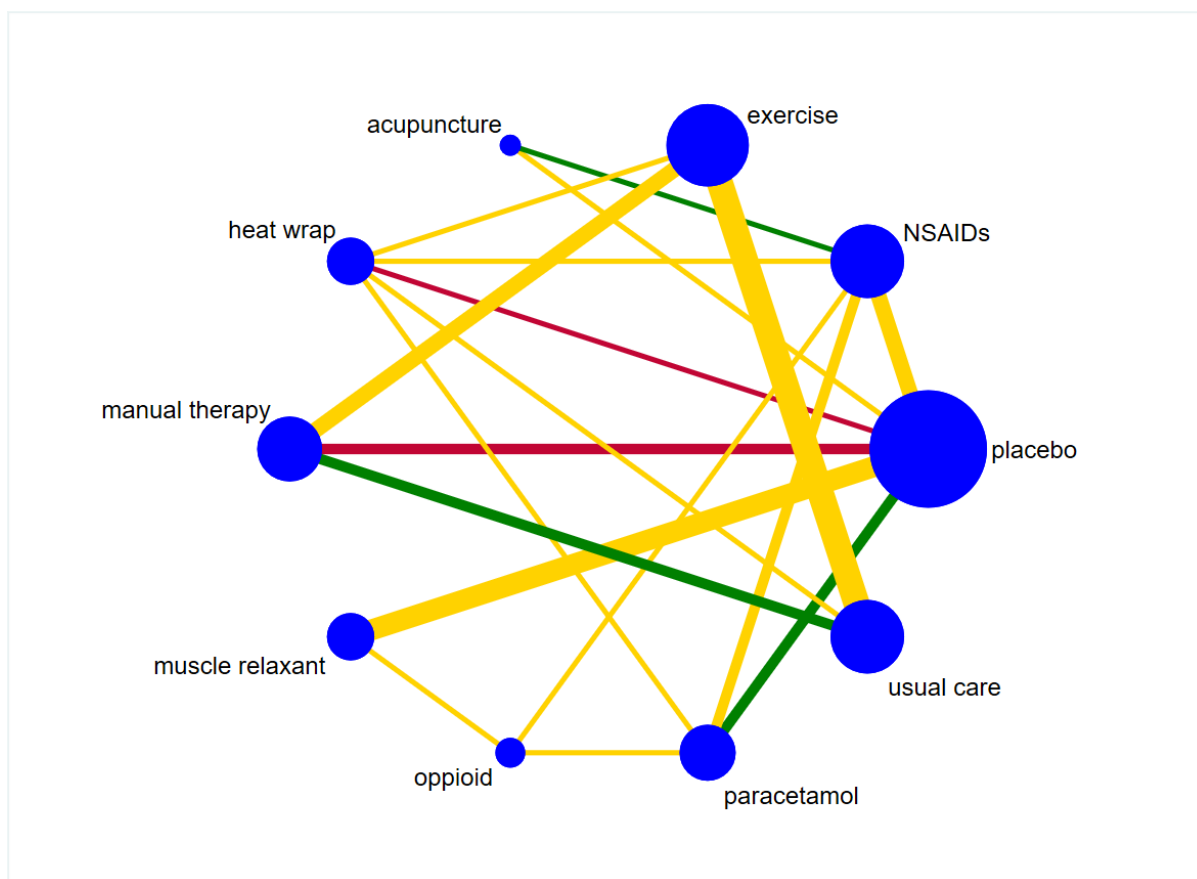
Figure 25. Network plot for RoB assessment.

Colored edges presented the RoB for each direct comparison in the network. Green, yellow and red colors are being used to denote pairwise MAs of low, unclear and high RoB.

Pairwise comparisons

The most studied comparison was the exercise versus usual care followed by muscle relaxant drug versus placebo. The only statistically significant pairwise comparisons were: manual therapy versus exercise (0.32; 95%CI 0.06, 0.58, I^2 =0), NSAIDs versus placebo (-0.46; 95%CI -0.66, -0.27, I^2=0), muscle relaxant drugs versus placebo (-1.06; 95%CI -1.88, -0.23, I^2=91) and manual therapy versus placebo (-1.29; 95%CI -1.85, -0.72, I^2=0). Figure 29 shows the available pairwise comparisons (Figure 26).

For most comparisons 95% confidence intervals for statistical heterogeneity were wide and point estimates included values suggesting either no or large heterogeneity, which reflects the small number of studies available for most pairwise comparisons.
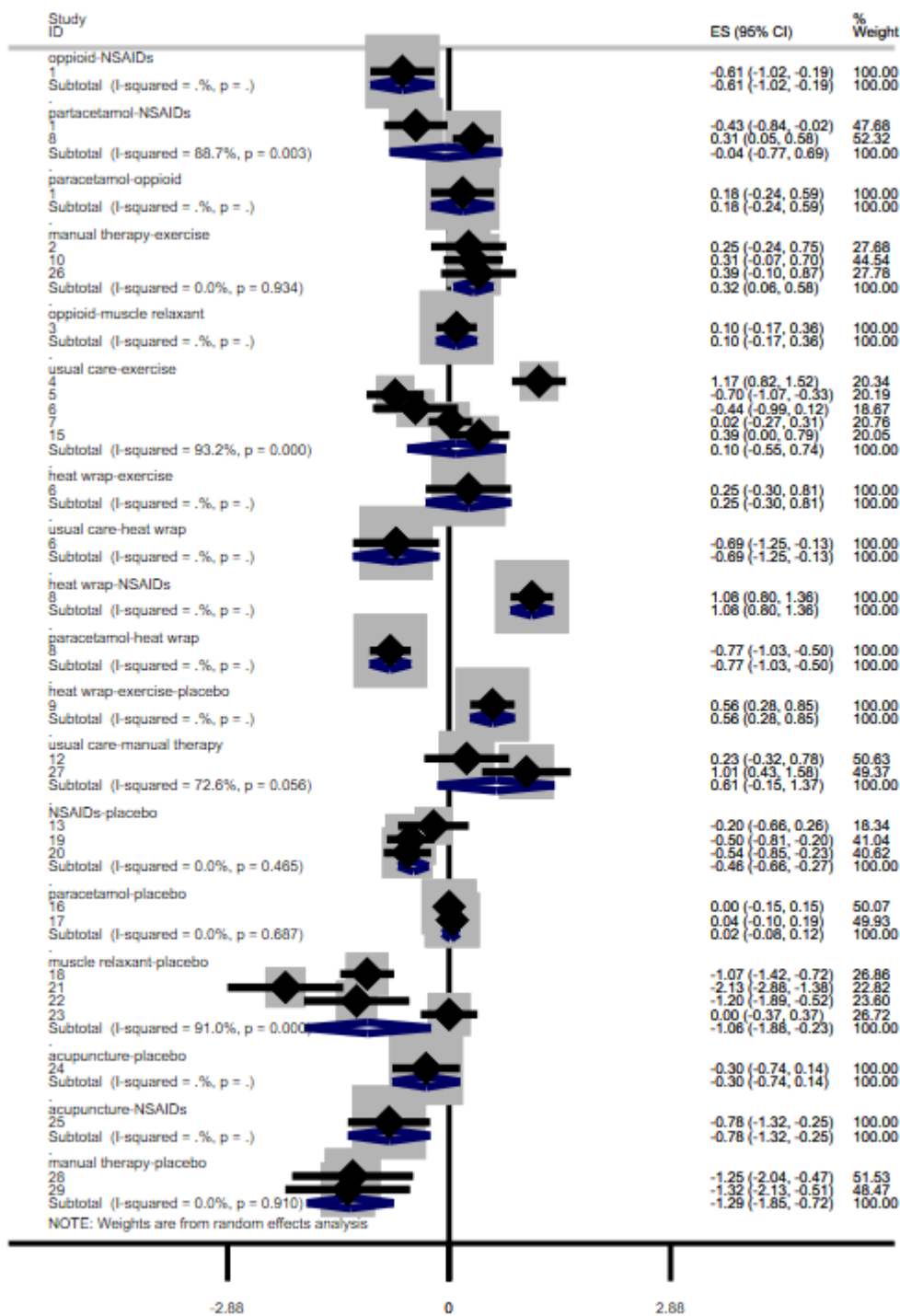
Figure 26. Pairwise comparisons in acute LBP trials.

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 (p=0.4181, χ^2=8.16 on 8 df). We then examined inconsistency for each loop. We found 8 triangular and 4 quadratic loops. Four out of 12 loops were inconsistent, with an inconsistency factor (IF) significantly greater than 0 (Figure 27). However, since any individual comparison may be involved in several loops, we 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 (Table 15). 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.
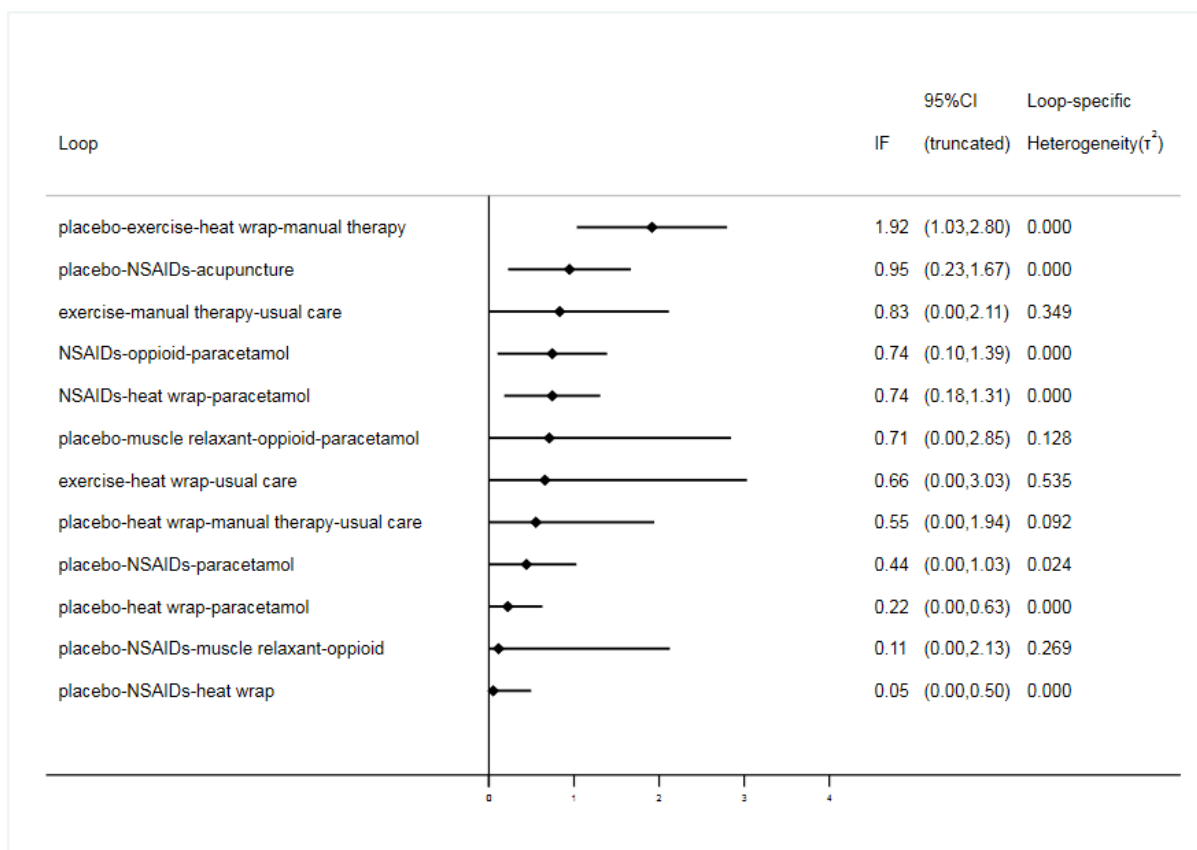
Figure 27. Consistency of loops.

Table 15. Analysis of inconsistency for every comparison.

		Direct effect		Indirect effect		Direct-Indirect effect			τ^2
		Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.	P	
placebo	NSAIDs	-0.42	0.33	-0.14	0.39	-0.28	0.51	0.58	0.54
placebo	acupuncture	-0.30	0.57	-1.16	0.65	0.87	0.86	0.32	0.53
placebo	heat wrap	0.56	0.56	0.24	0.44	0.33	0.71	0.65	0.54
placebo	manual therapy	-1.29	0.45	0.11	0.66	-1.40	0.80	0.08	0.50
placebo	muscle relaxant	-1.02	0.31	-0.73	0.79	-0.29	0.84	0.73	0.54
placebo	paracetamol	0.02	0.38	-0.51	0.41	0.53	0.56	0.34	0.53
NSAIDs	acupuncture	-0.78	0.59	0.08	0.63	-0.87	0.86	0.32	0.53
NSAIDs	heat wrap	1.08	0.55	0.33	0.50	0.75	0.74	0.31	0.53
NSAIDs	opioid	-0.61	0.58	-0.27	0.63	-0.34	0.86	0.69	0.54
NSAIDs	paracetamol	-0.05	0.40	0.28	0.49	-0.32	0.63	0.61	0.54
exercise	heat wrap	0.25	0.55	2.00	0.55	-1.75	0.78	0.02	0.47
exercise	manual therapy	0.32	0.31	-0.74	0.41	1.06	0.51	0.04	0.48
exercise	usual care	0.10	0.24	1.00	0.52	-0.89	0.57	0.12	0.51
heat wrap	paracetamol	-0.76	0.56	-0.41	0.54	-0.35	0.78	0.65	0.54
heat wrap	usual care	-0.69	0.61	-1.06	0.62	0.37	0.87	0.67	0.54
manual therapy	usual care	0.61	0.43	0.11	0.38	0.50	0.57	0.38	0.53
muscle relaxant	opioid	0.10	0.56	0.39	0.63	-0.29	0.84	0.73	0.54
opioid	paracetamol	0.18	0.57	1.00	0.65	-0.83	0.86	0.34	0.53

Network meta-analysis

As there was no evidence of global inconsistency, a consistency model was fitted. The statistical significant interventions for acute LBP were the muscle relaxant drugs against the placebo (-0.98; 95%CI -1.52, -0.43) followed by manual therapy (-0.84; 95%CI -1.61, -0.08). Opioid and acupuncture were borderline significant. Figure 28 summarizes the results of the NMA for the outcome measure of pain in the short-term for all trials. Table 16 presents the overall probabilities to be the best treatment. Provided hierarchies of ES on pain intensity. Interventions are ordered according to pain intensity ranking. For pain intensity, a SMD below 0 favor the column defining treatment. To obtain SMDs for comparisons in the opposite direction, negative values should be converted into positive values, and vice versa.
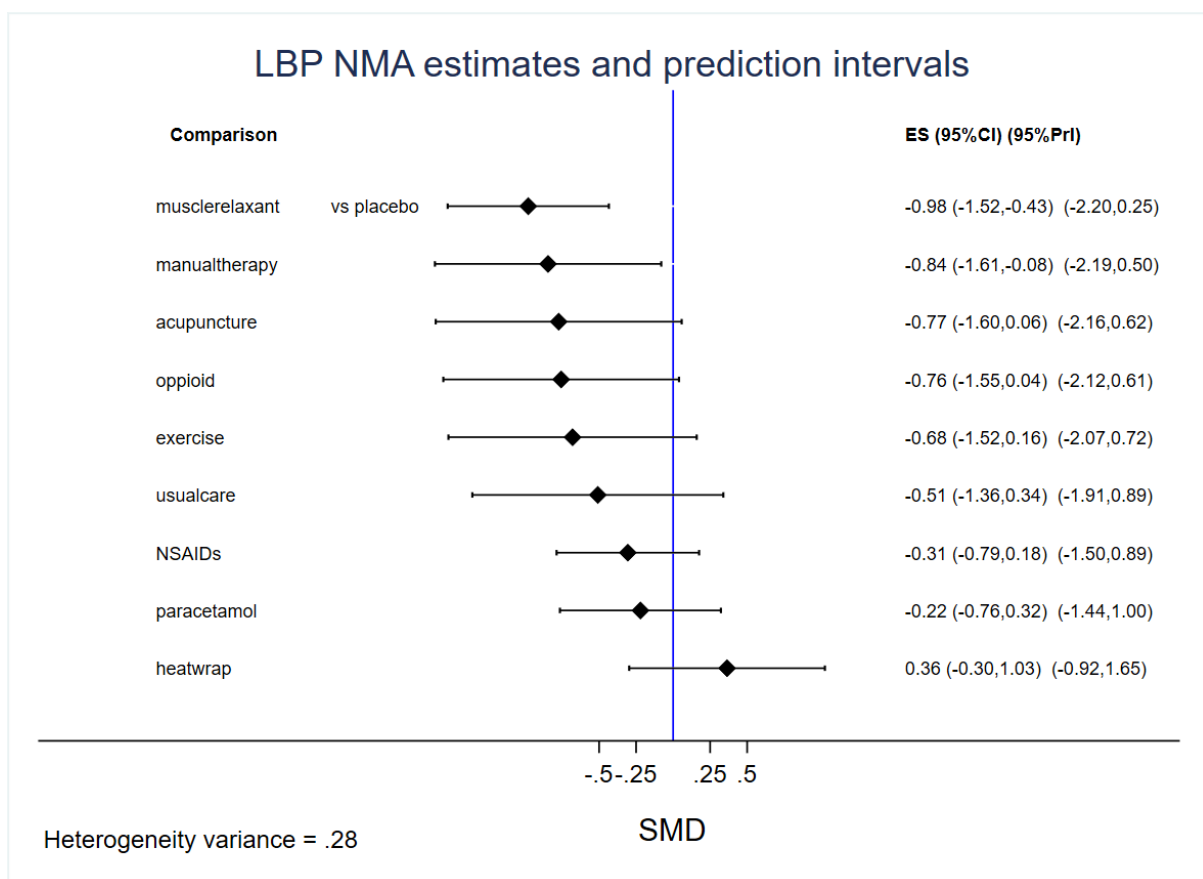
Figure 28. NMA for pain intensity at short-term of FU, all interventions against placebo.

Table 16. Multiple treatments meta-analysis, overall probabilities.

Comparisons between treatments 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.

usualcare	0.51 (-0.34,1.35)	0.29 (-0.67,1.24)	-0.25 (-1.39,0.89)	-0.47 (-1.47,0.53)	-0.34 (-0.89,0.22)	0.87 (0.04,1.71)	-0.26 (-0.71,0.18)	-0.17 (-1.35,1.00)	0.20 (-0.73,1.14)
-0.51 (-1.35,0.34)	placebo	-0.22 (-0.76,0.32)	-0.76 (-1.55,0.04)	-0.98 (-1.52,-0.43)	-0.84 (-1.61,-0.08)	0.36 (-0.30,1.03)	-0.77 (-1.60,0.06)	-0.68 (-1.52,0.16)	-0.31 (-0.79,0.18)
-0.29 (-1.24,0.67)	0.22 (-0.32,0.76)	paracetamol	-0.54 (-1.37,0.30)	-0.76 (-1.49,-0.03)	-0.62 (-1.52,0.27)	0.59 (-0.16,1.33)	-0.55 (-1.50,0.39)	-0.46 (-1.41,0.50)	-0.09 (-0.68,0.51)
0.25 (-0.89,1.39)	0.76 (-0.04,1.55)	0.54 (-0.30,1.37)	oppioid	-0.22 (-1.02,0.58)	-0.09 (-1.17,1.00)	1.12 (0.14,2.10)	-0.02 (-1.14,1.11)	0.08 (-1.04,1.19)	0.45 (-0.37,1.27)
0.47 (-0.53,1.47)	0.98 (0.43,1.52)	0.76 (0.03,1.49)	0.22 (-0.58,1.02)	muscle relaxant	0.13 (-0.80,1.07)	1.34 (0.50,2.18)	0.21 (-0.78,1.19)	0.30 (-0.69,1.29)	0.67 (-0.02,1.37)
0.34 (-0.22,0.89)	0.84 (0.08,1.61)	0.62 (-0.27,1.52)	0.09 (-1.00,1.17)	-0.13 (-1.07,0.80)	manual therapy	1.21 (0.39,2.03)	0.07 (-0.44,0.59)	0.16 (-0.96,1.29)	0.54 (-0.33,1.41)
-0.87 (-1.71,-0.04)	-0.36 (-1.03,0.30)	-0.59 (-1.33,0.16)	-1.12 (-2.10,-0.14)	-1.34 (-2.18,-0.50)	-1.21 (-2.03,-0.39)	heat wrap	-1.14 (-1.96,-0.31)	-1.04 (-2.08,-0.01)	-0.67 (-1.39,0.05)
0.26 (-0.18,0.71)	0.77 (-0.06,1.60)	0.55 (-0.39,1.50)	0.02 (-1.11,1.14)	-0.21 (-1.19,0.78)	-0.07 (-0.59,0.44)	1.14 (0.31,1.96)	exercise	0.09 (-1.07,1.26)	0.47 (-0.45,1.38)
0.17 (-1.00,1.35)	0.68 (-0.16,1.52)	0.46 (-0.50,1.41)	-0.08 (-1.19,1.04)	-0.30 (-1.29,0.69)	-0.16 (-1.29,0.96)	1.04 (0.01,2.08)	-0.09 (-1.26,1.07)	acupuncture	0.37 (-0.47,1.22)
-0.20 (-1.14,0.73)	0.31 (-0.18,0.79)	0.09 (-0.51,0.68)	-0.45 (-1.27,0.37)	-0.67 (-1.37,0.02)	-0.54 (-1.41,0.33)	0.67 (-0.05,1.39)	-0.47 (-1.38,0.45)	-0.37 (-1.22,0.47)	NSAIDs

Ranking of treatments

Rank probability indicating the possibility of each intervention being the best (1) and then the worst (0) are presented in table 17. Figure 29 shows the cumulative probability rank of the greatest likelihood of being the efficacious treatment for acute LBP. Muscle relaxant drugs (34.5%) ranked the first, the second was manual therapy (18.8%) and the last was the heat wrap (0%).

Table 17. Rank probabilities.

Treatment	SUCRA	PrBest	MeanRank
placebo	16	0	8.6
NSAIDs	38.2	0.1	6.6
exercise	70.7	13.5	3.6
acupuncture	63.2	15.4	4.3
heat wrap	3.3	0	9.7
manual therapy	76	18.8	3.2
muscle relaxant	83	34.5	2.5
oppioid	68.4	16	3.8
paracetamol	32.3	0.2	7.1
usual care	48.8	1.4	5.6
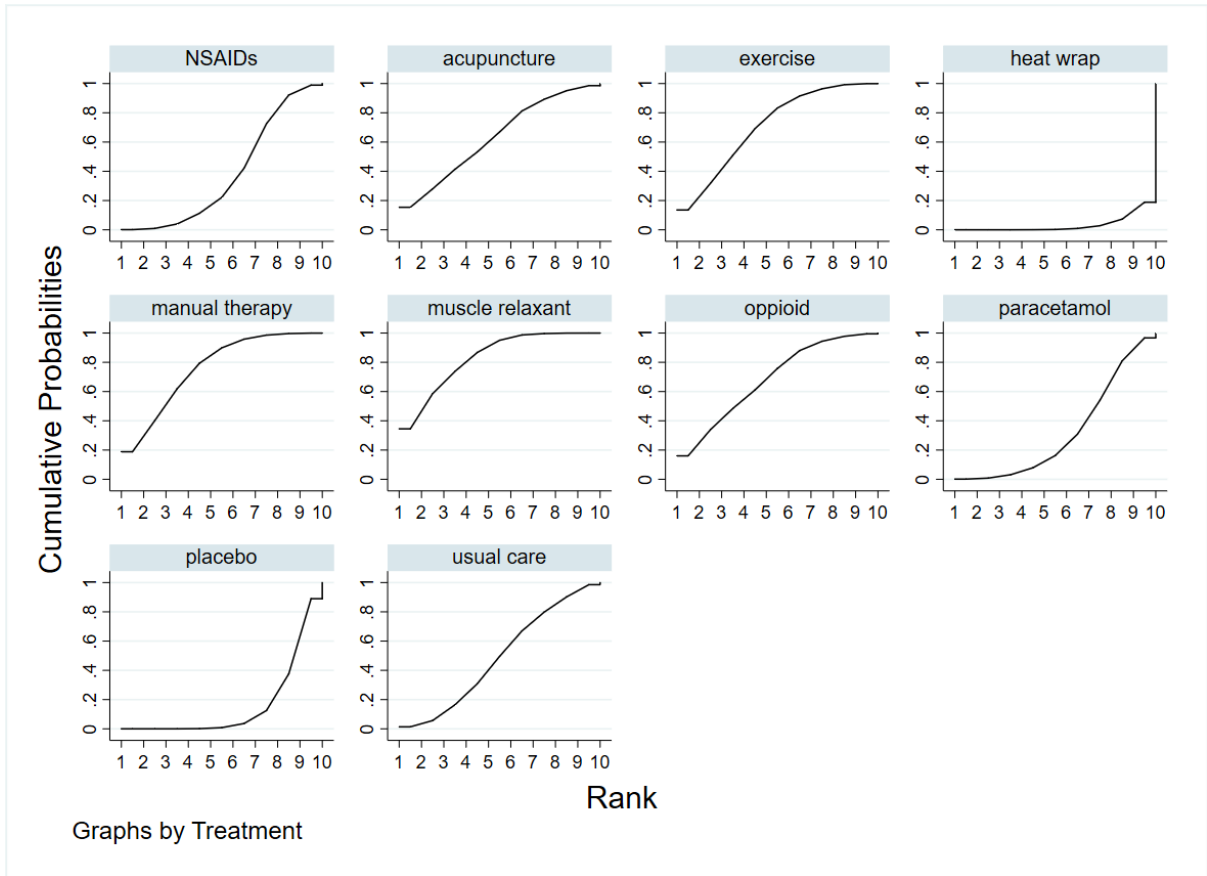
Figure 29. SUCRA Cumulative Probabilities.

Sensitivity analysis

The sensitivity analysis was performed by excluding studies with time point assessment < 1 week. Accordingly, we excluded 4 studies running the analyses only for treatments approximately ranging from 1 week to 1 month. Again, we did not found inconsistency. The effectiveness of treatments slightly change: the effect of acupuncture become significantly greater than placebo (-1.34; 95%CI -2.47, -0.21), together with the confirmed manual therapy (-1.28; 95%CI -2.25, -0.31) and muscle relaxant (-1.01; 95%CI -1.75, -0.27). Those interventions are the most efficacious interventions for acute and subacute LBP. Sensitivity analysis is shown in figure 30.
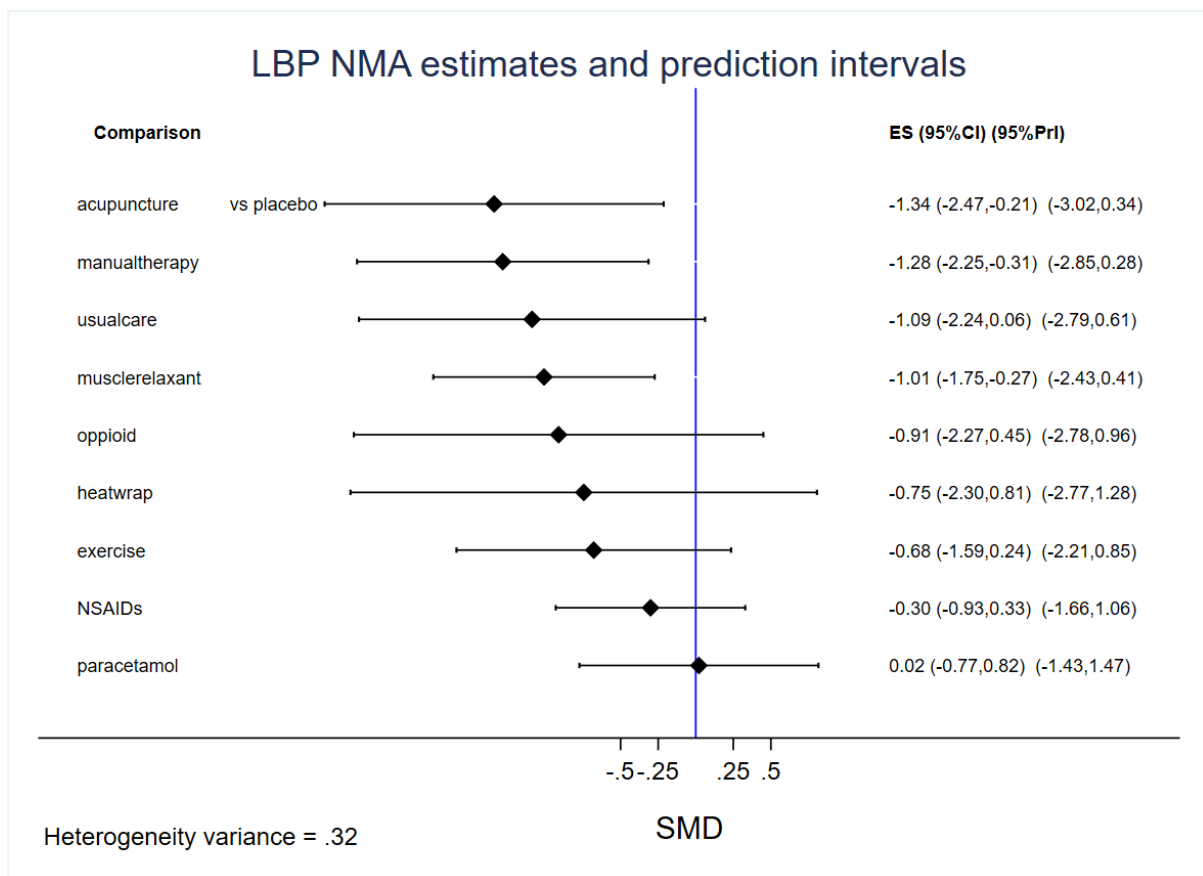
Figure 30. Sensitivity analysis.

Discussion of multiple interventions for acute LBP

In our NMA we compared all available evidence-based treatments for LBP and found that muscle relaxant drugs, followed by manual therapy, are the most efficacious treatments for reducing pain intensity at short term of FU. Muscle relaxant results are sustained from a recent SR studying their effects on LBP [204]. High quality evidence coming from five trials (n = 496 participants) declared muscle relaxants statistically significant against placebo for pain relief in the short-term (MD 21.3, 95%IC 29.0, 13.5) for acute LBP. However, authors called for caution with the interpretation of the findings as the evidence comes from specific muscle relaxant medicines (i.e., thiocolchicoside) [204]. We cannot compare our results for manual therapy with the recent published SRs [205] since interventions and comparator were not similar to ours. In particular, authors studied only spinal manipulation whereas in our network all manual therapy techniques were combined together for clinical as well as methodological reasons. In fact, in clinical setting often manipulations are administrated with manual therapy. In addition, lumping spinal manipulations and manual therapy techniques we increased the statistical power of manual therapy node.

For all other treatments, the 95% CI crossed the line of non-difference against placebo (or no treatment). This results was partially in contrast to the actual guidelines. Till now, eleven publish guidelines were found in LBP [206] and many inconsistencies were present between them [207]. For example, O'Connell et al. found that the Canadian guidelines advocate the use of tricyclic antidepressants and acetaminophen whereas the NICE guidelines only recommends to consider the use of NSAIDs, and if NSAIDs are ineffective, contraindicated or not tolerated, consider aweak opioid, with or without paracetamol for acute back pain. Furthermore, the actual American College of Physicians guideline published in 2017 [208] sustained non-pharmacologic treatment with superficial heat (moderate-quality evidence), massage,

acupuncture, or spinal manipulation (low-quality evidence). If pharmacologic treatment is desired, NSAIDs or muscle relaxants (moderate-quality evidence).

The difference between our investigation and the published guidelines could be found in the eligibility criteria of selected trials. We considered only patients with non-specific acute and sub-acute LBP (e.g., no sciatica) performing "pure" treatments, 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.

Unfortunately, in our study selection process we found that a large part of published trials comprised mixed interventions (both in pharmacological and non-pharmacological treatments) and they were excluded. For the remaining trials (n=36), the interventions were several and we grouped some of them into predefined nodes. However, for many direct and indirect comparisons evidence comes from a very limited number of studies (n=21) due to incomplete reporting of outcome data. Assessing the presence of clinical and methodological transitivity, we studied the pairwise comparisons and we found a certain amount of heterogeneity among studies within a direct comparisons. This is acceptable, as long it may be assumed that their treatment effects share a common mean, and it may even increase generalizability [141]. Since no global evidence of inconsistency between direct and indirect comparisons was found, we conclude that the muscle relaxant drugs have the 34.5% of probabilities to be the best treatment for acute and sub-acute LBP.

GENERAL CONCLUSION

Since the widespread use in clinical practice of treatments with no evidence of a beneficial effect wastes the limited health care resources and can harm patients, it is imperative that healthcare professionals involved in the care of people with non-specific LBP have access to up-to-date, evidence based information to assist them. At present, clinical practice treatment for back pain does not align with the available research evidence [209]. Even if abundant literature is produced around LBP interventions, only a small proportion is in line with EBM principles and often sound scientific evidences do not translate in clinical decision making. During the PhD, I investigated three fundamental aspects for KT.

First, the reporting of the essential elements was investigated (sample size, intervention and outcome assessment) in LBP rehabilitation trials and a poor reporting was found, preventing replication of many interventions in clinical practice. A further obstacle for translation research into practice is the interpretation of the results only in terms of statistical significance. In this thesis it was shown how results could change if they are presented in term of clinical relevance. A third issue is represented by a partial vision of the clinical problem when combining primary studies in systematic reviews and meta-analyses, which test only a pair of treatments at a time. Their conclusions on relative efficacy are limited to the pair of evaluated treatments whereas NMAs provide a more complete scenario of available treatments, allowing to establish the efficacy of an intervention relative to all other available treatments. Clinicians, researchers and all stakeholder in LBP rehabilitation field needs to be aware that improving reporting of RCTs, considering the produced clinical evidence in making clinical decision and valuing results coming by multiple comparisons meta-analysis will help to translate the (best) research evidence into practice.

Appendix

Appendix 1. Study selection of studies reporting sample size calculation in LBP rehabilitation interventions.

Appendix 2. Study selection of studies reporting intervention and outcome assessment in LBP rehabilitation interventions.

Identification

SRs identified through Cochrane Library searching (**n =75**)

Screening

SRs screened (**n = 28**)

SRs excluded with reasons:
-education and/or prevention (**n = 6**)
-not therapeutics (**n=1**)
-pregnancy (**n=1**)
-protocol (**n=8**)
-withdrawn (**n=1**)

SRs assessed for eligibility (**n = 11**)

Eligibility

Trials assessed for eligibility (**n = 220**)

Records removed after duplicates or related articles(**n =196**)

Full-text articles excluded, with reasons:
-not RCT (**n=0**)
-language (**n=7**)
-no full text retrieved (**n=4**)

Included

Studies included in qualitative synthesis (**n = 185**)

Appendix 3. Study selection of studies reporting clinical relevance in LBP rehabilitation interventions.

```
                    ┌─────────────────────────────┐
                    │ Cochrane SRs identified      │
                    │ through Cochrane Database     │
                    │ searching (n = 61)           │
                    └──────────────┬──────────────┘
                                   │
                                   ▼
                    ┌──────────────────┐      ┌─────────────────────────────┐
                    │ SRs screened      │─────▶│ Records excluded because     │
                    │ after duplicates  │      │ outside topics of the study  │
                    │ removed (n = 61)  │      │ (n= 43):                     │
                    └────────┬─────────┘      │ -Education or prevention      │
                             │                │ -Pregnancy                   │
                             │                │ -Diagnosis/prognosis         │
                             │                │ -Withdrawn                   │
                             │                │ -Surgery                     │
                             │                └─────────────────────────────┘
                             ▼
                    ┌──────────────────┐
                    │ SRs assessed for  │
                    │ eligibility       │
                    │ (n= 20)           │
                    └────────┬─────────┘
                             │
                             ▼
                    ┌──────────────────┐      ┌─────────────────────────────┐
                    │ RCTs assessed for │─────▶│ Full text excluded (n=63):   │
                    │ eligibility       │      │ -Language (n=2)              │
                    │ (n=105)           │      │ -Duplicates (n=4)            │
                    └────────┬─────────┘      │ -Irretrievable (n=13)        │
                             │                │ - MID and SS not reported    │
                             │                │ (n=27)                       │
                             │                │ -SS not reported (n=1)       │
                             │                │ -MID not reported (n=3)      │
                             │                │ -SD not MID (n=1)            │
                             │                │ -Effect size not MID (n=2)   │
                             │                │ -Graphics, impossible to find│
                             │                │ data (n=4)                   │
                             │                │ - Reported mean and          │
                             │                │ interquartile range (n=2)    │
                             │                │ - MID calculated after the   │
                             │                │ intervention (n=2)           │
                             │                │ -MID calculated intra-group  │
                             │                │ (n=1)                        │
                             │                │ -Regression mode (n=1)       │
                             ▼                └─────────────────────────────┘
                    ┌──────────────────┐
                    │ Studies included  │
                    │ in qualitative    │
                    │ synthesis (n=42)  │
                    └──────────────────┘
```

Cochrane SRs identified through Cochrane Database searching (n = 61)

SRs screened after duplicates removed (n = 61)

Records excluded because outside topics of the study (n= 43):
-Education or prevention
-Pregnancy
-Diagnosis/prognosis
-Withdrawn
-Surgery

SRs assessed for eligibility (n= 20)

Full text excluded (n=63):
-Language (n=2)
-Duplicates (n=4)
-Irretrievable (n=13)
- MID and SS not reported (n=27)
-SS not reported (n=1)
-MID not reported (n=3)
-SD not MID (n=1)
-Effect size not MID (n=2)
-Graphics, impossible to find data (n=4)
- Reported mean and interquartile range (n=2)
- MID calculated after the intervention (n=2)
-MID calculated intra-group (n=1)
-Regression mode (n=1)

RCTs assessed for eligibility (n=105)

Studies included in qualitative synthesis (n=42)

Appendix 4. Study selection of NMA in acute and subacute LBP.

Records identified
through MEDLINE
(PubMed) searching
(n = 3401)

Records identified
through EMBASE
searching
(n = 3901)

Records identified
through CENTRAL
searching
(n = 2114)

Records after Endnote and coevidence
duplicates removed
(**n =6964**)

Records screened
(**n =6964**)

Studies irrelevant
(**n =6419**)

Full-text articles assessed
for eligibility
(**n =545**)

Full-text articles excluded, with
reasons:
(**n =512**)
249 Wrong patient population
138 Wrong intervention
21 Wrong outcomes
25 Wrong study design
17 SRs/LG
9 protocol
10 poster conference
4 letter/author reply
14 duplicate
25 awaiting assessment

Included studies coming
from 11 SRs
(**n =3**)

Studies included in
qualitative synthesis
(**n = 33**)

Studies included in
qualitative synthesis
(**n = 36**)

Studies included in
quantitative
synthesis within 1 month
of follow-up
(**n =21**)

Appendix 5. Dataset of the network pain intensity in acute and subacute LBP.

study id	author	year	treatment	n	mean	sd
1	Eken	2014	paracetamol	46	19	22.4
1	Eken	2014	NSAIDs	46	27.6	20.4
1	Eken	2014	oppioid	45	15.5	16
2	Hsieh1	2002	manual therapy	49	2.58	1.93
26	Hsieh2	2002	manual therapy	51	2.78	1.82
2	Hsieh1	2002	exercise	24	2.13	1.28
3	Li	2008	myorelaxant	110	2.8	1.9
3	Li	2008	oppioid	110	3	2.2
4	Machado	2010	usual care	73	2.3	0.3
4	Machado	2010	exercise	73	2	0.2
5	Malmivaara	1995	exercise	52	3.1	1.7
5	Malmivaara	1995	usual care	67	1.9	1.7
6	Mayer	2005	heat wrap	25	2.3	1
6	Mayer	2005	exercise	25	1.96	1.5
6	Mayer	2005	usual care	26	1.37	1.44
7	Moffett	1999	exercise	89	16.35	9.9
7	Moffett	1999	usual care	98	16.53	9.9
8	Nadler	2002	heat wrap	113	2.61	0.87
8	Nadler	2002	paracetamol	113	1.95	0.87
8	Nadler	2002	NSAIDs	106	1.68	0.84
9	Nadler	2003	heat wrap	95	2.5	1.55
9	Nadler	2003	placebo	96	1.56	1.76
10	Rabin	2014	exercise	48	2.4	1.8
10	Rabin	2014	manual therapy	57	3.1	2.5
27	Schneider	2015	manual therapy	37	1.9	1.5
12	Schneider	2015	manual therapy	37	3.3	2
27	Schneider	2015	usual care	20	3.8	2.4
13	Szpalski	1994	NSAIDs	37	0.56	1.14
13	Szpalski	1994	placebo	36	0.79	1.09
15	Wand	2004	exercise	50	2.4	2
15	Wand	2004	usual care	52	3.3	2.5
16	Williams	2014	paracetamol	550	1.7	2.3
16	Williams	2014	placebo	272.5	1.7	2.3
17	Williams	2014	paracetamol	546	1.8	2.4
17	Williams	2014	placebo	272.5	1.7	2.3
18	Tuzun	2005	myorelaxant	74	25	21
18	Tuzun	2005	placebo	69	47	20
19	Dreiser	2003	NSAIDs	124	23.7	20.21*
20	Dreiser	2003	NSAIDs	122	22.9	20.21*
19	Dreiser	2003	placebo	61.5	33.9	20.21*
20	Dreiser	2003	placebo	61.5	33.9	20.21*
21	Ketenci	2005	myorelaxant	38	6.3	11.7
21	Ketenci	2005	placebo	13.5	47.3	27.9
22	Ketenci2	2005	myorelaxant	32	18.6	16.6

22	Ketenci2	2005	placebo	13.5	47.3	27.9
23	Berry	1988	myorelaxant	59	19	23.2
23	Berry	1988	placebo	54	19	22.9
24	Hasegawa	2014	acupuncture	40	2.58	2.3
24	Hasegawa	2014	placebo	40	3.3	2.47
25	Shin	2013	acupuncture	29	2.5	1.83*
25	Shin	2013	NSAIDs	29	3.95	1.83*
26	Hsieh2	2002	exercise	24	2.13	1.28
12	Schneider	2015	usual care	20	3.8	2.4
28	Takamoto	2015	manual therapy	11.5	5	20.21*
28	Takamoto	2015	placebo	21	31	20.21*
29	Takamoto2	2015	placebo	19	32.4	20.21*
29	Takamoto2	2015	manual therapy	11.5	5	20.21*

*Imputed SD

Bibliography

1.   Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet.*2016 Oct 10.
2.   Deyo RA, Weinstein JN. Low back pain. *The New England journal of medicine.*2001 Feb 01;344(5):363-370.
3.   Frymoyer JW. Back pain and sciatica. *The New England journal of medicine.*1988 Feb 04;318(5):291-300.
4.   van Tulder M, Becker A, Bekkering T, et al. Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. *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.*2006 Mar;15 Suppl 2:S169-191.
5.   Hoy D, March L, Brooks P, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Annals of the rheumatic diseases.*2014 Jun;73(6):968-974.
6.   Koes BW, van Tulder M, Lin CW, et al. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. *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.*2010 Dec;19(12):2075-2094.
7.   Casazza BA. Diagnosis and Treatment of Acute Low Back Pain. *American Family Physician.*2012;85(4):343-350.
8.   van Tulder MW, Koes BW, Bouter LM. A cost-of-illness study of back pain in The Netherlands. *Pain.*1995 Aug;62(2):233-240.
9.   Costa L, Maher CG, Hancock MJ, et al. The prognosis of acute and persistent low-back pain: a meta-analysis. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne.*2012 Aug 07;184(11):E613-624.
10.   Costa Lda C, Maher CG, McAuley JH, et al. Prognosis for patients with chronic low back pain: inception cohort study. *BMJ.*2009;339:b3829.
11.   MacDermid JC, Graham ID. Knowledge Translation: Putting the "Practice" in Evidence-Based Practice. *Hand Clinics.*2009 Feb;25(1):125-+.
12.   Williamson PR, Altman DG, Blazeby JM, et al. Developing core outcome sets for clinical trials: issues to consider. *Trials.*2012;13:132.
13.   Moher D, Simera I, Schulz KF, et al. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC medicine.*2008;6:13.
14.   Simera I, Altman DG, Moher D, et al. Guidelines for reporting health research: the EQUATOR network's survey of guideline authors. *PLoS medicine.*2008 Jun 24;5(6):e139.
15.   Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis and rheumatism.*2001 Aug;45(4):384-391.
16.   Wright JG. The minimal important difference: who's to say what is important? *Journal of clinical epidemiology.*1996 Nov;49(11):1221-1222.
17.   van der Roer N, Ostelo RW, Bekkering GE, et al. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine (Phila Pa 1976).*2006 Mar 1;31(5):578-582.
18.   Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ.*2005 Oct 15;331(7521):897-900.
19.   Cipriani A, Higgins JP, Geddes JR, et al. Conceptual and technical challenges in network meta-analysis. *Annals of internal medicine.*2013 Jul 16;159(2):130-137.
20.   Salanti G, Higgins JP, Ades AE, et al. Evaluation of networks of randomized trials. *Statistical methods in medical research.*2008 Jun;17(3):279-301.

21. Mills EJ, Ioannidis JP, Thorlund K, et al. How to use an article reporting a multiple treatment comparison meta-analysis. *Jama.*2012 Sep 26;308(12):1246-1253.
22. Riley RD, Abrams KR, Sutton AJ, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British journal of cancer.*2003 Apr 22;88(8):1191-1198.
23. Gianola S, Gasparini M, Agostini M, et al. Survey of the reporting characteristics of systematic reviews in rehabilitation. *Physical therapy.*2013 Nov;93(11):1456-1466.
24. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet.*2001 Apr 14;357(9263):1191-1194.
25. Moher D, Cook DJ, Eastwood S, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet.*1999 Nov 27;354(9193):1896-1900.
26. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine.*2009 Aug 18;151(4):264-269, W264.
27. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ.*2003 Jan 4;326(7379):41-44.
28. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet.*2007 Oct 20;370(9596):1453-1457.
29. Rennie D. CONSORT revised--improving the reporting of randomized trials. *Jama.*2001 Apr 18;285(15):2006-2007.
30. Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama.*1995 Feb 1;273(5):408-412.
31. DerSimonian R, Charette LJ, McPeek B, et al. Reporting on methods in clinical trials. *The New England journal of medicine.*1982 Jun 3;306(22):1332-1337.
32. A proposal for structured reporting of randomized controlled trials. The Standards of Reporting Trials Group. *Jama.*1994 Dec 28;272(24):1926-1931.
33. Call for comments on a proposal to improve reporting of clinical trials in the biomedical literature. Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. *Annals of internal medicine.*1994 Dec 1;121(11):894-895.
34. Rennie D. Reporting randomized controlled trials. An experiment and a call for responses from readers. *Jama.*1995 Apr 5;273(13):1054-1055.
35. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Jama.*1996 Aug 28;276(8):637-639.
36. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ.*2010;340:c332.
37. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of clinical epidemiology.*2010 Aug;63(8):e1-37.
38. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *The Cochrane database of systematic reviews.*2012;11:MR000030.
39. McKeown A GJ, 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. Reporting of Sample Size Calculations in Analgesic Clinical Trials: ACTTION Systematic Review. *The Journal of Pain.*2015;16(3 (March)):199-206.
40. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet.*2005 Apr 9-15;365(9467):1348-1353.

41.   Castellini G. GS, Banfi G., Bonovas S., Moja L. Mechanical low back pain: secular trend and intervention topics of randomized controlled trials. *Physiotherapy Canada.*2015;In press.

42.   Abdul Latif L, Daud Amadera JE, Pimentel D, et al. Sample size calculation in physical medicine and rehabilitation: a systematic review of reporting, characteristics, and results in randomized controlled trials. *Arch Phys Med Rehabil.*2011 Feb;92(2):306-315.

43.   Rutterford C, Taljaard M, Dixon S, et al. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *Journal of clinical epidemiology.*2014 Dec 15.

44.   Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ.*2010;340:c869.

45.   Antes G. The new CONSORT statement. *BMJ.*2010;340:c1432.

46.   Glasziou P, Meats E, Heneghan C, et al. What is missing from descriptions of treatment in trials and reviews? *BMJ.*2008 Jun 28;336(7659):1472-1474.

47.   Schroter S, Glasziou P, Heneghan C. Quality of descriptions of treatments: a review of published randomised controlled trials. *BMJ open.*2012;2(6).

48.   Boutron I, Moher D, Altman DG, et al. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Annals of internal medicine.*2008 Feb 19;148(4):295-309.

49.   Boutron I, Guittet L, Estellat C, et al. Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. *PLoS medicine.*2007 Feb;4(2):e61.

50.   Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Annals of internal medicine.*2002 Sep 17;137(6):511-520.

51.   Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ.*2014;348:g1687.

52.   Coster WJ. Making the best match: selecting outcome measures for clinical trials and outcome studies. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association.*2013 Mar-Apr;67(2):162-170.

53.   Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Archives of general psychiatry.*1995 Oct;52(10):867-871.

54.   Moja L. Clinical trials: trial registration cannot alone transform scientific conduct. *Nat Rev Urol.*2010 Jan;7(1):7-8.

55.   Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine (Phila Pa 1976).*2011 Oct 1;36(21 Suppl):S54-68.

56.   Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. *Spine (Phila Pa 1976).*1998 Sep 15;23(18):2003-2013.

57.   Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *Jama.*2013 Feb 27;309(8):814-822.

58.   Moher D, Tetzlaff J, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews. *PLoS medicine.*2007 Mar 27;4(3):e78.

59.   MeSH. http://www.ncbi.nlm.nih.gov/mesh. In: National Library Medicine controlled vocabulary NIoHN e, accessed in September 2013.

60.   DistillerSR. https://systematic-review.ca. *accessed on October 2013.*

61.   Stata-IC. Stata Statistical Software, v 13. College Station, TX: StataCorp LLC2013.

62.   Rubinstein SM, Terwee CB, Assendelft WJ, et al. Spinal manipulative therapy for acute low back pain: an update of the cochrane review. *Spine (Phila Pa 1976).*2013 Feb 1;38(3):E158-177.

63.   Yousefi-Nooraie R, Schonstein E, Heidari K, et al. Low level laser therapy for nonspecific low-back pain. *The Cochrane database of systematic reviews.*2008 (2):CD005107.

64.   Walker BF, French SD, Grant W, et al. Combined chiropractic interventions for low-back pain. *The Cochrane database of systematic reviews.*2010 (4):CD005427.

65.   Hayden JA, van Tulder MW, Malmivaara A, et al. Exercise therapy for treatment of non-specific low back pain. *The Cochrane database of systematic reviews.*2005 (3):CD000335.

66.   Heymans MW, van Tulder MW, Esmail R, et al. Back schools for non-specific low-back pain. *The Cochrane database of systematic reviews.*2004 (4):CD000261.

67.   Furlan AD, Imamura M, Dryden T, et al. Massage for low-back pain. *The Cochrane database of systematic reviews.*2008 (4):CD001929.

68.   Clarke JA, van Tulder MW, Blomberg SE, et al. Traction for low-back pain with or without sciatica. *The Cochrane database of systematic reviews.*2007 (2):CD003010.

69.   Khadilkar A, Odebiyi DO, Brosseau L, et al. Transcutaneous electrical nerve stimulation (TENS) versus placebo for chronic low-back pain. *The Cochrane database of systematic reviews.*2008 (4):CD003008.

70.   Urrutia G, Burton AK, Morral A, et al. Neuroreflexotherapy for non-specific low-back pain. *The Cochrane database of systematic reviews.*2004 (2):CD003009.

71.   Rubinstein SM, van Middelkoop M, Assendelft WJ, et al. Spinal manipulative therapy for chronic low-back pain. *The Cochrane database of systematic reviews.*2011 (2):CD008112.

72.   Henschke N, Ostelo RW, van Tulder MW, et al. Behavioural treatment for chronic low-back pain. *The Cochrane database of systematic reviews.*2010 (7):CD002014.

73.   Ebadi S, Henschke N, Nakhostin Ansari N, et al. Therapeutic ultrasound for chronic low-back pain. *The Cochrane database of systematic reviews.*2014;3:CD009169.

74.   Kamper SJ, Apeldoorn AT, Chiarotto A, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain. *The Cochrane database of systematic reviews.*2014;9:CD000963.

75.   Wegner I, Widyahening IS, van Tulder MW, et al. Traction for low-back pain with or without sciatica. *The Cochrane database of systematic reviews.*2013;8:CD003010.

76.   Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Annals of internal medicine.*2013 Feb 5;158(3):200-207.

77.   COMET Database CI, accessed on January 2015 http://www.comet-initiative.org/studies/searchresults/. COMET Initiative, accessed on January 2015 http://www.comet-initiative.org/studies/searchresults/

78.   The COMET Initiative. http://www.comet-initiative.org/. Accessed access on 16 February 2015

79.   McDowell I. Measuring Health: a guide to rating scales and questionnaires, third Edition. *Oxford University Press.*2006.

80.   Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet.*2002 Feb 23;359(9307):696-700.

81.   Tugwell P, Boers M. Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. OMERACT Committee. *The Journal of rheumatology.*1993 Mar;20(3):555-556.

82.   Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of clinical epidemiology.*2010 Jul;63(7):737-745.

83.   Higgins J, Altman D, Sterne Je. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. . *Version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from www.cochrane-handbook.org.*2011.

84.   R Core Team. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2014.

85.   Rubinstein SM, Terwee CB, Assendelft WJ, et al. Spinal manipulative therapy for acute low-back pain. *The Cochrane database of systematic reviews.*2012;9:CD008880.

86. Yousefi-Nooraie R, Schonstein E, Heidari K, et al. Low level laser therapy for nonspecific low-back pain. *The Cochrane database of systematic reviews.*2011 (2):CD005107.

87. Hayden JA, van Tulder MW, Malmivaara A, et al. Exercise therapy for treatment of non-specific low back pain. *The Cochrane database of systematic reviews.*2011 (3):CD000335.

88. Heymans MW, van Tulder MW, Esmail R, et al. Back schools for non-specific low-back pain. *The Cochrane database of systematic reviews.*2011 (4):CD000261.

89. Furlan AD, Imamura M, Dryden T, et al. Massage for low-back pain. *The Cochrane database of systematic reviews.*2011 (4):CD001929.

90. Clarke JA, van Tulder MW, Blomberg SE, et al. Traction for low-back pain with or without sciatica. *The Cochrane database of systematic reviews.*2010 (2):CD003010.

91. Khadilkar A, Odebiyi DO, Brosseau L, et al. Transcutaneous electrical nerve stimulation (TENS) versus placebo for chronic low-back pain. *The Cochrane database of systematic reviews.*2010 (4):CD003008.

92. Urrutia G, Burton AK, Morral A, et al. Neuroreflexotherapy for non-specific low-back pain. *The Cochrane database of systematic reviews.*2011 (2):CD003009.

93. Henschke N, Ostelo RW, van Tulder MW, et al. Behavioural treatment for chronic low-back pain. *The Cochrane database of systematic reviews.*2011 (7):CD002014.

94. Pino C, Boutron I, Ravaud P. Inadequate description of educational interventions in ongoing randomized controlled trials. *Trials.*2012;13:63.

95. Glasziou P, Chalmers I, Altman DG, et al. Taking healthcare interventions from trial to practice. *BMJ.*2010;341:c3852.

96. Clarke M. Standardising outcomes for clinical trials and systematic reviews. *Trials.*2007;8:39.

97. Kirkham JJ, Gargon E, Clarke M, et al. Can a core outcome set improve the quality of systematic reviews?--a survey of the Co-ordinating Editors of Cochrane Review Groups. *Trials.*2013;14:21.

98. Mannion AF, Boneschi M, Teli M, et al. Reliability and validity of the cross-culturally adapted Italian version of the Core Outcome Measures Index. *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.*2012 Aug;21 Suppl 6:S737-749.

99. Costa LO, Maher CG, Lopes AD, et al. Transparent reporting of studies relevant to physical therapy practice. *Rev Bras Fisioter.*2011 Aug-Sep;15(4):267-271.

100. Cook JA, Hislop J, Altman DG, et al. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials.*2015;16(1):12.

101. Chiarotto A, Terwee CB, Deyo RA, et al. A core outcome set for clinical trials on non-specific low back pain: study protocol for the development of a core domain set. *Trials.*2014;15:511.

102. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain.*2018 Mar;159(3):481-495.

103. Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. *The spine journal : official journal of the North American Spine Society.*2007 Sep-Oct;7(5):541-546.

104. Rodriguez-Mansilla J, Espejo-Antunez L, Bustamante-Lopez AI. [Effectiveness of acupuncture in spasticity of the post-stroke patient. Systematic review]. *Atencion primaria / Sociedad Espanola de Medicina de Familia y Comunitaria.*2015 Jul 10.

105. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of consulting and clinical psychology.*1991 Feb;59(1):12-19.

106. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976).*2008 Jan 1;33(1):90-94.

107. Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine (Phila Pa 1976).*2000 Dec 15;25(24):3100-3103.

108. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of clinical epidemiology.*2008 Feb;61(2):102-109.

109. Guyatt GH, Deyo RA, Charlson M, et al. Responsiveness and validity in health status measurement: a clarification. *Journal of clinical epidemiology.*1989;42(5):403-408.

110. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled clinical trials.*1989 Dec;10(4):407-415.

111. Copay AG, Glassman SD, Subach BR, et al. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *The spine journal : official journal of the North American Spine Society.*2008 Nov-Dec;8(6):968-974.

112. Stratford PW, Binkley JM, Riddle DL, et al. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Physical therapy.*1998 Nov;78(11):1186-1196.

113. Hajiro T, Nishimura K. Minimal clinically significant difference in health status: the thorny path of health status measures? *The European respiratory journal.*2002 Mar;19(3):390-391.

114. Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. *The Journal of rheumatology.*2001 Feb;28(2):406-412.

115. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current opinion in rheumatology.*2002 Mar;14(2):109-114.

116. Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. *Best practice & research. Clinical rheumatology.*2005 Aug;19(4):593-607.

117. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clinic proceedings.*2002 Apr;77(4):371-383.

118. Wyrwich KW, Bullinger M, Aaronson N, et al. Estimating clinically significant differences in quality of life outcomes. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation.*2005 Mar;14(2):285-295.

119. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical care.*2003 May;41(5):582-592.

120. Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *The Journal of rheumatology.*2001 Feb;28(2):400-405.

121. Bolton JE. Sensitivity and specificity of outcome measures in patients with neck pain: detecting clinically significant improvement. *Spine (Phila Pa 1976).*2004 Nov 1;29(21):2410-2417; discussion 2418.

122. Cohen J. *Statistical power: analysis for the behavioural sciences*. New York: Academic Press1977.

123. Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Medical care.*2003 May;41(5):593-596.

124. Sloan JA, Dueck A. Issues for statisticians in conducting analyses and translating results for quality of life end points in clinical trials. *Journal of biopharmaceutical statistics.*2004 Feb;14(1):73-96.

125. Lassere MN, van der Heijde D, Johnson KR. Foundations of the minimal clinically important difference for imaging. *The Journal of rheumatology.*2001 Apr;28(4):890-891.

126. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials.*1986 Sep;7(3):177-188.

127. Johnston BC, Thorlund K, Schunemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. *Health Qual Life Outcomes.*2010;8:116.

128. Johnston BC, Thorlund K, da Costa BR, et al. New methods can extend the use of minimal important difference units in meta-analyses of continuous outcome measures. *Journal of clinical epidemiology.*2012 Aug;65(8):817-826.

129. Good M, Stiller C, Zauszniewski JA, et al. Sensation and Distress of Pain Scales: reliability, validity, and sensitivity. *Journal of nursing measurement.*2001 Winter;9(3):219-238.

130. Ware JE, Jr., Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *Journal of clinical epidemiology.*1998 Nov;51(11):903-912.

131. Tavafian SS, Jamshidi AR, Montazeri A. A randomized study of back school in women with chronic low back pain: quality of life at three, six, and twelve months follow-up. *Spine (Phila Pa 1976).*2008 Jul 1;33(15):1617-1621.

132. Tavafian SS, Jamshidi AR, Mohammad K. Treatment of chronic low back pain: a randomized clinical trial comparing multidisciplinary group-based rehabilitation program and oral drug treatment with oral drug treatment alone. *The Clinical journal of pain.*2011 Nov-Dec;27(9):811-818.

133. Moix J, Canellas M, Osorio C, et al. Efficacy of an interdisciplinary educational program in patients with chronic back pain. *Dolor.*2003;18(3):149–157.

134. Ferreira ML, Herbert RD, Ferreira PH, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *Journal of clinical epidemiology.*2012 Mar;65(3):253-261.

135. Gianola S, Frigerio P, Agostini M, et al. Completeness of outcomes description reported in low-back pain rehabilitation interventions: a survey of 185 randomized trials. *accepted on Physiotherapy Canada.*2016 March.

136. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS medicine.*2009 Jul 21;6(7):e1000100.

137. Young C, Horton R. Putting clinical trials into context - Reply. *Lancet.*2005 Sep;366(9490):980-980.

138. Moseley AM, Herbert RD, Sherrington C, et al. Evidence for physiotherapy practice: A survey of the Physiotherapy Evidence Database (PEDro). *Australian Journal of Physiotherapy.*2002;48(1):43-49.

139. Jansen JP, Crawford B, Bergman G, et al. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.*2008 Sep-Oct;11(5):956-964.

140. Jansen JP, Trikalinos T, Cappelleri JC, et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.*2014 Mar;17(2):157-173.

141. Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.*2011 Jun;14(4):429-437.

142. Chaimani A, Caldwell D, Lic T, et al. Undertaking network meta-analyses. *Unpublished draft chapter prepared for the Cochrane Handbook for Systematic Reviews of Interventions.*2016.

143. Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *Journal of clinical epidemiology.*2009 Aug;62(8):857-864.

144. Donegan S, Williamson P, Gamble C, et al. Indirect comparisons: a review of reporting and methodological quality. *PloS one.*2010 Nov 10;5(11):e11054.
145. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in medicine.*2004 Oct 30;23(20):3105-3124.
146. Lu. G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. *Journal of the American Statistical Association.*2006;101(474).
147. Song F, Loke YK, Walsh T, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ.*2009 Apr 3;338:b1147.
148. Wong JJ, Cote P, Sutton DA, et al. Clinical practice guidelines for the noninvasive management of low back pain: A systematic review by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. *Eur J Pain.*2017 Feb;21(2):201-216.
149. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Annals of internal medicine.*2015 Jun 02;162(11):777-784.
150. Chaimani A, Caldwell DM, Li T, et al. Additional considerations are required when preparing a protocol for a systematic review with multiple interventions. *Journal of clinical epidemiology.*2017 Jan 11.
151. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ.*2011 Oct 18;343:d5928.
152. Furlan AD, van Tulder M, Cherkin D, et al. Acupuncture and dry-needling for low back pain: an updated systematic review within the framework of the cochrane collaboration. *Spine (Phila Pa 1976).*2005 Apr 15;30(8):944-963.
153. COEVIDENCE. https://www.covidence.org/reviews. *accessed on January 2018.*
154. Higgins J, Deeks J, Altman D. Chapter 16: Special topics in statistics. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. *Version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from www.cochrane-handbook.org.*2011.
155. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ.*2003 Sep 6;327(7414):557-560.
156. White IR, Barrett JK, Jackson D, et al. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research synthesis methods.*2012 Jun;3(2):111-125.
157. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research synthesis methods.*2012 Jun;3(2):80-97.
158. Miladinovic B, Hozo I, Chaimani A, et al. Indirect treatment comparison. *Stata Journal.*2014;14(1):76-86.
159. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of clinical epidemiology.*2011 Feb;64(2):163-171.
160. Jackson D, Barrett JK, Rice S, et al. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Statistics in medicine.*2014 Sep 20;33(21):3639-3654.
161. Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in medicine.*2010 Mar 30;29(7-8):932-944.
162. Higgins JP, Jackson D, Barrett JK, et al. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research synthesis methods.*2012 Jun;3(2):98-110.
163. Veroniki AA, Vasiliadis HS, Higgins JP, et al. Evaluation of inconsistency in networks of interventions. *International journal of epidemiology.*2013 Feb;42(1):332-345.
164. Stata Statistical Software, v 15. College Station, TX: StataCorp LLC [computer program]2017.

165. Chaimani A, Higgins JP, Mavridis D, et al. Graphical tools for network meta-analysis in STATA. *PloS one.*2013;8(10):e76654.
166. (MTM). M-TM-a. A framework for evaluating and ranking multiple healthcare technologies. Accessed http://www.mtm.uoi.gr/ (accessed 8 February 2017).
167. White I. Multivariate random-effects meta-regression: updates to mvmeta. *The STATA Journal.*2011;11:255–270.
168. Casale R. Sintomatic treatment with a muscle relaxant drug. *The Clinical journal of pain.*1988 (4):81-88.
169. 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.
170. 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.
171. Hagen EM, Odelien 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 Healt.*2010 Nov;38(7):731-738.
172. Hindle TH. Comparison of Carisoprodol, Butabarbital, and Placebo in Treatment of Low Back Syndrome. *Calif Med.*1972;117(2):7-&.
173. 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.
174. 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.
175. 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.
176. 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.
177. 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.
178. 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.
179. 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.
180. 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.
181. 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.
182. 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.
183. 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.
184. 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.

185. Postacchini F, Facchini M, Palmieri P. Efficacy of various forms of conservative treatments in low back pain. *Neur Orthop.*1988 Dec;6:28-35.

186. 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.

187. 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.

188. 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.

189. Roberts L, Little P, Chapman J, et al. The Back Home Trial: General practitioner-supported leaflets may change back pain behavior. *Spine.*2002 Sep 1;27(17):1821-1828.

190. 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.

191. 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.

192. 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.

193. 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.

194. 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.

195. 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.

196. 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.

197. 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.

198. 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.

199. 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.

200. 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.

201. 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.

202.   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.

203.   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.

204.   Abdel Shaheed C, Maher CG, Williams KA, et al. Efficacy and tolerability of muscle relaxants for low back pain: Systematic review and meta-analysis. *Eur J Pain.*2017 Feb;21(2):228-237.

205.   Paige NM, Miake-Lye IM, Booth MS, et al. Association of Spinal Manipulative Therapy With Clinical Benefit and Harm for Acute Low Back Pain: Systematic Review and Meta-analysis. *Jama.*2017 Apr 11;317(14):1451-1460.

206.   Lin I, Wiles LK, Waller R, et al. Poor overall quality of clinical practice guidelines for musculoskeletal pain: a systematic review. *British journal of sports medicine.*2018 Mar;52(5):337-343.

207.   O'Connell NE, Cook CE, Wand BM, et al. Clinical guidelines for low back pain: A critical review of consensus and inconsistencies across three major guidelines. *Best practice & research. Clinical rheumatology.*2016 Dec;30(6):968-980.

208.   Qaseem A, Wilt TJ, McLean RM, et al. Noninvasive Treatments for Acute, Subacute, and Chronic Low Back Pain: A Clinical Practice Guideline From the American College of Physicians. *Annals of internal medicine.*2017 Apr 4;166(7):514-530.

209.   Hodder RK, Wolfenden L, Kamper SJ, et al. Developing implementation science to improve the translation of research to address low back pain: A critical review. *Best practice & research. Clinical rheumatology.*2016 Dec;30(6):1050-1073.