Blinding and sham control methods in trials of physical, psychological, and self-management interventions for pain (article II):
a meta-analysis relating methods to trial results

Authors

David Hohenschurz-Schmidt[1] *, Dr Jerry Draper-Rodi[2], Prof Lene Vase[3], Dr Whitney Scott[4,5], Prof Alison McGregor[6], Dr Nadia Soliman[1], Andrew MacMillan[2], Axel Olivier[7], Cybill Ann Cherian[8], Daniel Corcoran[9], Dr Hilary Abbey[2], Sascha Freigang[10], Jessica Chan[7], Jules Phalip[11], Lea Nørgaard Sørensen[12], Maite Delafin[13], Margarida Baptista[14], Naomi R Medforth[15], Nuria Ruffini[16], Stephanie Skøtt Andresen[17], Sylvain Ytier[18], Dorota Ali[19], Harriet Hobday[20], Anak Agung Ngurah Agung Adhiyoga Santosa[21], Dr Jan Vollert[1,22,23,24], Prof Andrew SC Rice[1]

Affiliations

[1] Pain Research, Dept. Surgery & Cancer, Faculty of Medicine, Imperial College, Chelsea & Westminster Hospital campus, 369 Fulham Road London, London SW10 9NH, UK

[2] Research Centre, University College of Osteopathy, 275 Borough High Street, London SE1 1JE, UK

[3] Aarhus University, Department of Psychology and Behavioural Sciences, Section for Psychology and Neuroscience, Bartholins Allé 11, Building 1350, 316; 8000 Aarhus C, DK

[4] Health Psychology Section, Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[5] INPUT Pain Management Unit, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

[6] Human Performance Group, Dept. Surgery & Cancer, Faculty of Medicine, Imperial College, White City Campus, 80–92 Wood Lane, London W12 0BZ, UK

[7] Faculty of Medicine, Imperial College London, London, England

[8] Khalifa University, Abu Dhabi, United Arab Emirates

[9] no institutional affiliations, West Ballina, Australia

[10] Department of Neurosurgery, Medical University Graz, Auenbruggerplatz 29, Graz 8036 Austria

[11] Résidence les Estrangers, 60 bis Place de la Victoire, 63150 La Bourboule, France

[12] Department of Occupational Medicine, Danish Ramazzini Centre, Aarhus University Hospital, Denmark

[13] The Penn Clinic, Hertfordshire, Hatfield, United Kingdom

[14] Wolfson Centre for Age Related Diseases, Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[15] no institutional affiliations, London, United Kingdom

[16] National Centre Germany, Foundation C.O.M.E. Collaboration, Berlin, 10825, Germany

[17] Department of Psychology and Behavioural Sciences, Aarhus University, Bartolins Allé 11, Building 1350, 8000 Aarhus C, DK.

[18] no institutional affiliations, Vevey, Switzerland

[19] Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[20] Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[21] no institutional affiliations, Denpasar, Indonesia.

[22] Division of Neurological Pain Research and Therapy, Department of Neurology, University Hospital of Schleswig-Holstein, Campus Kiel, Germany

[23] Neurophysiology, Mannheim Center of Translational Neuroscience (MCTN), Medical Faculty Mannheim, Heidelberg University, Germany

[24] Department of Anaesthesiology, Intensive Care and Pain Medicine, University Hospital Muenster, Germany

* Corresponding author details:
Email: d.hohenschurz-schmidt19@imperial.ac.uk
URL: https://www.imperial.ac.uk/people/d.hohenschurz-schmidt19
phone: +44 (0)20 3315 8816
Postal: as above

Conflict of interest and further declarations

We have no conflict of interest to declare in relation to this work.
This work has not been presented publicly or published elsewhere. Data have been used to inform a guideline development process, results of which have not been publicised yet.
All listed authors have contributed substantially to the project and fulfil ICMJE criteria.
No approval from an ethics committee was required.

Introduction

Clinical trials are thought to show larger effects when they are not blinded [45,57]. Blinding or masking refers to preventing trial participants from knowing which of the available treatments they receive, so that such knowledge does not affect their expectation of treatment benefit and thus bias trial results. Expectancies are widely assumed to mediate health benefits due to contextual factors - the placebo effect. Further placebo mechanisms include learning and conditioning, where various symptom-modifying neurobiological pathways are activated and which is dependent on contextual factors of a treatment. Related placebo contributors are the therapeutic interaction and the treatment's meaning to patients [48,7,26].

In clinical trials, 'placebo' or 'sham' control interventions are employed for blinding, and to control for the psychosocial context of the treatment, spontaneous remission, and other confounding factors [48,76]. In principle, this can be achieved through a therapeutic ritual that does not include features of the experimental treatment thought to produce the treatment effects and subject to study in a given trial. The concept of placebo controls is best illustrated by drug studies, where identical pills can be employed, not containing the pharmacologically active agent but otherwise allowing for the same doctor interactions and rituals of pill-taking [76]. Non-pharmacological therapies, however, are often more complex in their procedures, and more reliant on patient-provider interactions than drug therapies. Further, the distinction between specific and contextual treatment components can be highly challenging and controversial [61,77,23]. Consequently, blinding is more difficult in trials of non-pharmacological treatments [16]. In non-pharmacological studies, such specifically-designed control interventions are usually called 'sham' or 'attention controls', despite slight differences in meaning.

For complex intervention studies, inert pills are not considered adequate controls, mainly because less elaborate placebos produce less pronounced placebo effects [46,89,54]. Efficacy trials of complex interventions require complex control interventions, matching some or most features of the intervention. For example, sham ultrasound is often used in manual therapy trials, which is a largely dissimilar complex control treatment in this context. Whilst low-similarity shams are easy to design and have, on occasion, been shown to effectively blind participants to group allocation [8], they may have a higher risk of unblinding and are unable control for influential but unknown contextual factors. Importantly, trials that employ dissimilar controls are thought to over-estimate treatment efficacy [5]. Further, the placebo response is highly variable [83,78] and it is often unclear which aspects of the psychosocial treatment context influence the individual trial and to what extent, making it desirable to fully replicate the experimental treatment's context.

Therefore, a high degree of similarity between control and experimental interventions is commonly recommended for efficacy trials [75,5,52,18,34,39,64,3], but quantitative data to substantiate such recommendations are sparse. Some authors have used the concepts of 'indistinguishability', 'sham fidelity', and 'structural equivalence' to denote this similarity [5,52,64]. Despite 'indistinguishability' being commonly recommended, it has not been systematically assessed which aspects in the resemblance between experimental and control interventions predict differences in trial outcomes. Such knowledge would enable researchers to prioritise and target efforts during the development of novel control interventions for efficacy trials, acknowledging the practical constraints of clinical trials. The present meta-analysis draws on previous work to define numerous features across which control and

treatment interventions may be compared [3,5,14,17,18,21,24,34,39,45,64,68,73,75], but refrains from a dichotomous distinction into 'similar' and 'dissimilar'.

There are currently no guidelines for dedicated control interventions in efficacy or mechanistic trials of physical, psychological, and self-management therapies (PPS) for pain. Such guidelines can only be sensibly developed based on improved insights regarding the effect of design decisions on trial feasibility, participant blinding, and outcomes. As such, the question of whether similarity between test and control interventions matters is of primary importance. Also, with blinding effectiveness rarely reported in trial publications [41], the retrospective assessment of control intervention quality would be facilitated by establishing quality standards. To advance this field, a systematic review of sham-controlled RCTs of PPS interventions was conducted. In a linked publication, we have reported the results of similarity assessments between control and experimental treatments, and information on additional blinding methods, control intervention development processes, blinding assessments, and reporting guideline compliance [41]. Here, we compare the control and tested interventions across multiple features and test whether different levels of similarity between interventions predict differences in trial results, including pain-related outcomes, blinding effectiveness, and differential attrition.

Methods

Reporting of this systematic review follows the PRISMA 2020 statement [62]. Further methodological detail is reported in a parallel article [41] where the included trial methods were described in detail, including discussion of similarity features between interventions and differences between therapy groups. In short, the following methods were employed for the systematic search, record screening, and data extraction.

Protocol and registration

A protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) (registration ID: CRD42020206590). This publication reports the second part of the review, including outcome data and the meta-analysis. Protocol deviations occurred in relation to the employed meta-analysis methods as specified below.

Eligibility criteria

We included randomized controlled trials of PPS interventions for adults living with pain. Studies had to report at least one pain-related outcome measure. PPS interventions included all forms of manual and physical therapy, exercise and rehabilitation therapy, conversation-based and psychological therapies, body-mind, spiritual, religious, other non-material healing practices, web-based treatments, relaxation, and educational interventions (the latter two were classified as 'self-management' here). We excluded drug studies, surgery, nutrition and infusions, device-based treatments, acupuncture and dry needling, and any other treatment based on meridian or reflex point considerations, because these provide different challenges and opportunities for control intervention design than the group of non-pharmacological interventions studied here [18,6]. To be eligible for this review, trials had to employ a sham control intervention (or 'attention' or 'placebo control'). Pilot or feasibility studies were only included if they specifically assessed features of the control intervention in a pain population, irrespective of employed outcome measures (henceforth called validation studies). This review systematically assessed studies published from 2008 onwards [15].

Data sources

The following databases were searched from January 2008 to 24 June 2020. MEDLINE, EMBASE, PsychInfo, the Cochrane Database of Systematic Reviews, the Cochrane Central Register of Controlled Trials (CENTRAL), NIH Clinicaltrials.gov, AMED (Allied and Complementary Medicine), CINAHL (nursing and allied health), the Physiotherapy evidence database (pedro.org.au), ostmed.dr (ostmed-dr.oclc.org), osteopathic research web (osteopathic-research.com), and the index to chiropractic literature (chiroindex.org). The year 2008 was used as a cut-off as the first reporting guideline for non-pharmacological trials was published in that year [15].

Search strategy

The search strategy was built around the following keywords and is provided in full for each database as supplement (spreadsheet, Supplemental Digital Content 1). Aligning with the PICOS framework, this review's target population, interventions, control conditions, outcomes, and study designs are encompassed by the following:

Pain OR painful conditions AND Physical, Psychological, Self-management therapies (specific therapy and technique names) AND placebo control OR sham control OR attention control AND controlled clinical trials. Limit: 2008-present. Pain-related outcome measures were not searched for specifically.

Study selection

Eligibility screening was performed in duplicate by two independent reviewers drawn from a pool of specifically trained research contributors. Disagreements were resolved by a third independent reviewer. Screening was first performed based on study title and abstract, and full-text second.

Instead of excluding smaller trials over risk of bias concerns [37,30,31,85,84], this review provides a descriptive overview of existing, otherwise eligible trials of less than 20 participants per arm at randomization, allowing for the potential identification of novel or promising alternative methods of placebo controlling. Further, we included these smaller studies in sensitivity analyses of the meta-regression modelling. Thus, and in extension of our initial protocol, outcome data were extracted and risk-of-bias assessments performed for these trials.

Data extraction

Data extraction also required two independent reviewers, with discrepancies resolved through discussion or by a third independent reviewer. In trials with more than one sham control group, data for all sham control groups were extracted and treated independently in the analysis. Publications reporting multiple sham controls were assessed independently for each pair of intervention and control intervention, with data from an active intervention arm used twice for comparisons with control interventions if required (sample size of the respective group was halved to account for duplicate inclusion). Where a single sham group acted as comparator for multiple active interventions, data were extracted from the active intervention that most resembled the control intervention.

The domains of data extraction were bibliographic information, general study design, trial reporting, sham control and blinding methods, trial result (including attrition and blinding effectiveness), and risk of bias using the Cochrane Risk of Bias tool 2 (RoB 2) (Sterne et al., 2019). Resemblance between sham and experimental treatments was rated for 25 features.

Pain-related outcome measures were extracted for sham control and active intervention groups at baseline and earliest follow-up after treatment. Pain-related outcomes can be unidimensional (usually pain intensity rating scales) or multidimensional (e.g., questionnaires assessing disease-related function, pain interference or quality of life) [74]. In the absence of research on whether these are differentially susceptible to placebo effects, we strove to extract both a unidimensional and a multidimensional outcome measure per study. Specifically, data were sampled for the declared primary pain-related outcome, irrespective of its nature. Where available, another pain-related outcome of the opposite dimensionality was extracted. Where authors did not declare a primary outcome measure, data for a unidimensional and a multidimensional measure were extracted if available. Direction of effect was considered. Where necessary, data were extracted from figures using the Adobe Reader measurement tool. Authors were contacted via email if data were missing that were required for the calculation of effect sizes or if data appeared erroneous. For cross-over designs, only results from the first phase were extracted.

Data analysis

Descriptives and subgroups

A descriptive overview of blinding methods used in the field of PPS interventions for pain is provided in a parallel publication [41] and the Supplemental Digital Content 2, including a basic description of placebo control interventions, their development and reported rationale, the similarity between control and active interventions, compliance with relevant reporting guidelines, and reports of blinding effectiveness.

Studies were sub-grouped into large and small trials or by therapy type where appropriate. Trials of less than 20 participants per arm at randomization and placebo control validation studies without pain-related outcome measures were not used for primary meta-analyses but included in sensitivity analyses. Trials that could not be pooled were only analysed descriptively [41]. Criteria for non-pooling were pain or disability that was not expected to improve in a comparable manner over the course of the study (cancer- and pregnancy-related pain) and studies with patients who had no pain or pharmaceutically reduced pain at outset (pain interventions during or immediately after surgery). Outliers were checked for errors in data reporting or entry and removed if errors in the primary data were suspected or if between-group standardised mean changes deviated from the group mean by more than 2.5 standard deviations.

Meta-analysis: Placebo responses and treatment effects

For all studies with more than 20 participants per arm, we synthesised outcome data and present a risk of bias (RoB) assessment (Sterne et al., 2019). For each included control group, placebo responses were calculated as standardised mean changes (SMC) from baseline to first follow-up after the end of the treatment period, for both a uni- and a multidimensional pain-related outcome measure, where available. To calculate SMCs, a change score was divided by the pooled standard deviations.

Between-group differences were calculated as the difference in SMCs between active and control groups for the same measures and timepoints. Meta-analyses of between-group differences were performed per therapy subgroup and for uni- and multidimensional measures, separately. For each group of therapies, summary effects were calculated using random effects models weighted by the inverse of the variance and plotted using RevMan 5 software [72]. The heterogeneity of overall effects was estimated using Tau^2 (T^2) and I^2 statistics and tested for significance using Z statistics [13]. Data are presented in forest plots, also illustrating study-specific risk-of-bias.

Meta-analysis: Similarity ratings

'Similarity ratings' were calculated by converting the evaluation of how similar individual features were between active and sham control interventions into numerical scores. Specifically, clear 'Yes' (similar) and 'No' (dissimilar) evaluations were rated as 2 and -2, respectively. 'Probably Yes' and 'Probably No' were awarded 1 and -1 points, and 0 points were given for items that could not be rated due to insufficient information. Non-applicable items were not rated. Results of these ratings are presented in the parallel publication [41] and as supplement to this article (Supplemental Digital Content 3, table). These data were used as covariates for meta-regression in the present paper.

Meta-analysis: Blinding effectiveness, blinding indices, and treatment expectations

During data extraction, we identified all studies that provided an indication as to the effectiveness of the employed blinding methods. Where group guesses were reported in a manner that allowed for the calculation of Bang's blinding index (BI) the index was calculated for active and control groups individually [4] and a ratio calculated as Hedge's g for each comparison between active and sham control group [25]. Descriptive results are again provided as a supplemental table (Supplemental Digital Content 4). The effect size of Hedge's g was estimated irrespectively of the interventions studied, using meta-analysis methods as above.

In addition, we identified trials that reported measuring participant expectancy or related concepts (treatment credibility and satisfaction). On the suggestion of a reviewer and not specified in the original protocol, we examined these reports for the possibility of data pooling and meta-analysis of expectancy measures. For meta-analysis and regression with similarity ratings, reported expectancy data had to be i) clearly attributable to expectation of treatment benefit only (i.e., not presented as a compound measure with questions on treatment credibility, or evaluated as treatment satisfaction), ii) sampled after at least one exposure to the test or control intervention but not after the final of multiple treatment sessions (to avoid confounding with treatment satisfaction), and iii) reported in full and per trial group.

Multiple meta-regression analysis: the role of similarity between intervention and sham controls in predicting trial outcomes or blinding effectiveness

Based on interim feedback from subject experts, we deviated from the pre-registered protocol to perform a more rigorous meta-regression analysis instead of simple correlation testing. Specifically, to assess the potential relationship between the trial results and how (dis)similar sham and active interventions were, for each individual therapy subgroup, meta-regression analyses were performed using methods-of-moments random effects models [13,86]. The SMC between sham control and active intervention groups was used as the dependent variable and models were weighted by the inverse of their variances. Models were built for each subgroup individually, identifying potential predictive variables from the pool of similarity ratings. This was done based on non-parametric correlation analyses between all ratings for each subgroup to identify a selection of variables with little interdependence between each other, and then further-refined by iterative adjustment of the model until a model of supposed best-fit was found. Put simply, we tested if similarity ratings could account for some of the differences in pain-related outcomes between trials. In particular, this method examined differences that were not likely due to the 'true' difference in treatment efficacy but due to other factors, for example control methods.

Meta-regression modelling was also performed for a subgroup of studies for which Bang's blinding index could be calculated, testing if similarity features could predict variance in the studies' blinding success as well as testing if the blinding index could predict the variance in studies' effect sizes. These analyses were irrespective of the type of therapy tested in the trials. Similarly, two meta-regression models were computed to examine if differences in attrition between studies predicted trial outcomes and, further, whether the degree of similarity between active and sham interventions predicted the degree of differential attrition. The results of meta-regression models are presented per therapy group below. Primary analyses were performed with large studies only (20 or more participants per group) and sensitivity-tested using the complete dataset and/or excluding studies that did not formally qualify as outliers (i.e., were included in the meta-analysis and primary modelling), but whose confidence intervals did not overlap with those of the aggregate effect.

We initially planned to use a trial-level average of similarity ratings for meta-regression analyses but decided to use ratings for individual items only, given concerns about the validity of a compound score.

Notably, many of the individual items' ratings were inter-correlated so that an overall score would have been biased. Also, an equal weighting of all items as part of an average was seen as an undue assumption. We do, however, explore the average of all similarity ratings as part of our descriptive analysis (reported separately [41]).

Results

Sample description

Included trials' characteristics are illustrated in table 1. Figure 1 shows the study selection process. We reviewed 194 publications (plus protocols where available), extracting data for 197 unique sham interventions and 198 comparisons between sham and experimental interventions. Manual therapy trials were most common, and there were multiple psychological and rehabilitation trials. Mostly, patients with musculoskeletal pain were recruited. Whilst sham control interventions were not always well-described, we were able to classify a range of employed methods, including control interventions that were clearly modelled based on the active treatment under investigation and such that were very dissimilar. Further-describing levels of similarity between control and experimental interventions, we identified features for which similarity was frequently given, such as the amount and frequency of treatments. For other aspects similarity was more variable, often also depending on the category of intervention studied. The first part of this results section will describe the entire sample, highlighting the subset of large trials eligible for primary meta-analyses, whilst the second part presents the meta-analysis of eligible studies only (Table 1). A table with each study's details is provided as supplement (Supplemental Digital Content 2).

Figure 1: PRISMA flow diagram of the systematic search and selection process. Complete search strings per database and a list of all studies excluded at the full-text screening stage are provided in the Supplementary Digital Content file 1.

Identification

Records identified through database searching
(n = 21,259)

Additional records identified through other sources
(n = 227)

Records after duplicates removed
(n = 12,529)

Screening

Records screened
(n = 12,529)

Records excluded
(n = 11,854)

Eligibility

Full-text articles assessed for eligibility
(n = 722)

Full-text articles excluded
(n = 528)

246 Literature review
68 No full text published or found
62 Wrong comparator
57 Pain not the main outcome
49 Duplicate paper or study
43 Wrong intervention
41 Wrong patient population
28 Irrelevant pilot /
feasibility trial
11 Publication before 2008
4 Not an RCT
3 Commentary on study
2 Full text not translatable

(multiple reasons for exclusion possible)

Included

Articles included in **Qualitative synthesis**
(n = 194)
(198 active/control intervention pairs)

Not included in meta-analysis
(n = 28)

5 Outliers / data errors
1 Pregnancy-related pain
6 Cancer-related pain
1 Intraoperative pain
Therapy subgroups too small for pooling:
8 Spiritual / Energy healing
5 Self-management
2 Other

Comparisons included in **Meta-analysis**
(n = 121)

Additional pairs included for **Sensitivity analyses**
(n = 49)

11

Table 1: Sample overview. *The types of therapies, intervention complexity, and pain population are provided for the entire sample and per group. "Large" trials had 20 or more participants per arm and their pain-related outcome data were used for meta-analyses. Special cases (large trials): In one trial, data from the active intervention group was used twice to compare it with two different sham controls: Bialosky et al. (2014) used a 'standard' and an 'enhanced' sham control. Three publications reported more than one trial: D'Souza et al. (2008) studied two groups with different types of headaches, and Assefi et al.'s (2008) publication included two active interventions and a matching sham control each. Finally, Sharpe et al. (2012) reported two trials in a single publication, which were treated entirely independently here. In general, only patients that informed the present analyses are counted in this table, patients were not counted twice, and analyses of reporting refer to individual trials.* [*] Intervention complexity: Single-step or single-technique interventions were judged as 'simple', irrespective of how often these were applied, and others as complex. N = 194 unique trials.

	Total sample		Large trials		Small trials	
Therapy types	n	%	n	%	n	%
Manual therapy with spinal manipulation	48	24.2	36	25.0	12	22.2
Craniosacral therapy and gentle myofascial release	22	11.1	16	11.1	6	11.1
Other Manual Therapy	64	32.3	40	27.8	24	44.4
Rehabilitation / Physiotherapy	22	11.1	16	11.1	6	11.1
Self-management	5	2.5	4	2.7	1	1.9
Cognitive behavioural & other psychotherapy	27	13.6	26	18.1	1	1.9
Spiritual / energetic / esoteric healing	8	4.0	5	3.5	3	5.6
Other	2	1.0	1	0.7	1	1.9
Intervention complexity *	**n**	**%**	**n**	**%**	**n**	**%**
Simple	112	56.6	73	50.7	39	72.2
Complex	82	41.4	71	49.3	11	20.4
Pain descriptor	**n**	**%**	**n**	**%**	**n**	**%**
Musculoskeletal pain	121	61.1	88	61.1	33	61.1
Headaches and orofacial pain	22	11.1	15	10.4	7	13.0
Diffuse chronic pain	18	9.1	16	11.1	2	3.7
Injury- and medical intervention-related pain	19	9.6	8	5.6	11	20.4
Cancer-related pain	6	3.0	6	4.2	0	0.0
Visceral pain	5	2.5	4	2.8	1	1.9
Neuropathic pain	5	2.5	5	3.5	0	0.0
Pregnancy-related pain	1	0.5	1	0.7	0	0.0
Pre-registered trial protocol available	**n**	**%**	**n**	**%**	**n**	**%**
Pre-registered	114	57.9	90	62.9	24	44.4
Total included	**Total sample**		**Large trials**		**Small trials**	
	198		144		54	

Validation studies

Two [80,35] of eight validation studies [38,47,29,19,56,24] were eligible for primary meta-analysis of pain-related outcomes. Three validation studies with more than 20 participants per group [24,29,56] did not provide pain-related outcome data and were thus not included into respective meta-analyses. They were, however, included into the meta-analysis of trials reporting on blinding success.

Placebo and sham control intervention designs

Employed sham control interventions are listed and classified in the companion paper, as are the results of the similarity assessment between sham and tested interventions, and differences between therapies [41]. Importantly, average similarity ratings were not significantly different between large and small trials (F(1,197) = 3.56, $p = 0.061$). However, physiotherapy / rehabilitation trials employed overall more dissimilar sham interventions than spinal manipulation trials, other manual therapies (excluding craniosacral therapy), and trials of spiritual or energetic therapies. Apart from the difference to other manual therapy trials, these differences were still significant when only large trials were tested and when pregnancy and cancer-related trials were excluded as in our meta-regression analyses below.

Reports of blinding effectiveness

In 19 reports, blinding indices were provided or data were reported in a manner that allowed for calculating Bang's index [4]. These studies were included for meta-analysis.

Reports of expectation of benefit from interventions

Only six trials (with seven control interventions) reported expectancy data in a manner that allowed for data pooling [10,29,60,11,56,28]. Apart from being an excessively small sample for meta-analysis [13], none of those studies reported significant differences in expectations between groups after an initial exposure to trial interventions. Meta-analysis and regression testing were thus not deemed promising and were not performed. In the remaining 23 trials with some mention of expectation-related assessments, the reasons for non-comparability were: 1) Inappropriate timepoint of assessments (either unexposed at baseline, or retrospectively after completing multi-session treatment programme and thus likely confounded by satisfaction), 2) Compound assessment with treatment credibility without individually reporting data of expectancy-only questions [27], 3) Confounding expectancy and satisfaction, and 4) Insufficient outcome data reporting. More details are presented in Supplementary Digital Content 4.

Study outcomes

Employed outcome measures

The most common unidimensional outcome measures employed were pain intensity rating scales (92%). Pressure pain thresholds were used in 8% of the 168 pairs of active and control interventions with extractable unidimensional outcome measures. Multidimensional outcome measures were available for 130 comparisons. These measures were mainly disability questionnaires (68%), followed by multidimensional pain questionnaires (19%), functional tests (8%), disease activity or symptom scores (5%), and general health questionnaires (2%). Relatively more objective and supposedly pain-related outcome measures were employed in 37% of all trials, including functional tests, disease

markers, nerve conduction, autonomic nervous system parameters, brain imaging, work absenteeism or medication use.

Attrition

The difference in percentual attrition (i.e., the differential attrition) between active and sham control groups was a mean of 0.4 % more in active treatment groups (-.74 to 1.5 95% CI, Q(136) = 6054, p = .51, T^2 = 43.52, I^2 = 97.8 %), as estimated with a random-effects model with all studies weighted equally and applied to all large studies irrespective of therapy type (participants = 13,150, studies = 137, not reported in 6 studies). Longer studies reported more percentual attrition (r_s's(135) = 0.43, $p < 0.001$; 0.42, $p < 0.001$, respectively for active and control groups). Drop-out reasons related to the nature of the control intervention were reported in 9 studies (6.3%).

Risk-of-bias

In studies used for the primary meta-analysis, the overall risk-of-bias was judged low in 17% of studies, high in 44% and some concerns existed for 38% (figure 2). In the smaller studies, 15% were rated as low risk, 82% high risk, and there were some concerns for 4% (not illustrated).

Figure 2: Risk of Bias assessment for all comparisons with 20 or more participants per group (n = 128). Overall risk-of-bias was judged "High" if at least one domain had a rating of high risk-of-bias or if there were concerns in more than one domain. The overall rating of "Some concerns" was given if there were concerns in any one domain, and "Low" if all domains were rated as of low risk-of-bias, as per the Cochrane Risk of Bias tool 2 (RoB 2) [71].
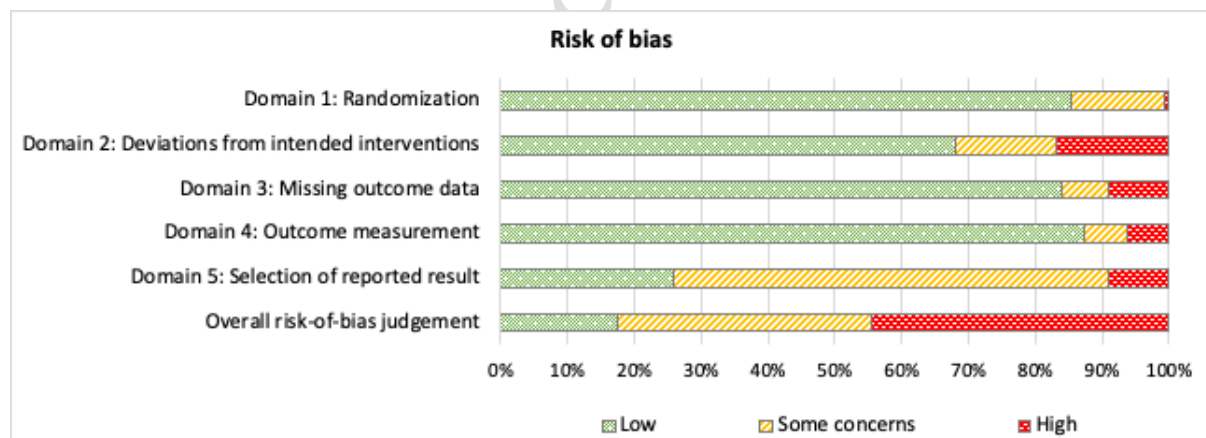
Pain-related outcome measures

Effect sizes (as standardised mean changes, SMC) could be calculated for 166 intervention-vs-control comparisons for unidimensional outcomes and 125 comparisons using multidimensional outcome measures. Between-groups SMCs are illustrated below per subgroup. For the entire sample, between-group differences were not significantly different when studies were grouped by high, some or low overall risk-of-bias (unidimensional outcomes: F(2,164) = 1.217, p = 0.3.; multidimensional: F(2,123) = 0.231, p = 0.79).

Meta-analyses and meta-regression

Out of 198 included comparisons, 120 were included in the primary meta-analysis, with a further 49 smaller studies used for secondary sensitivity analyses. Reasons for non-pooling included classification as outlier or apparent data errors (n = 5), patient populations in which no comparable improvements in pain report were expected (n = 8), and therapy subgroups too small for pooling (n = 16) (Figure 1 – PRISMA flowchart). Since not all studies provided data for each outcome, study numbers in the following analyses vary.

Sham control within-group effects (Placebo responses)

Across all studies included in the primary meta-analysis, changes from baseline to earliest follow-up within sham control groups had an average SMC effect size of -0.46 (-0.53 to -0.38 95% CI, $p < 0.0001$, $T^2 = 0.12$, $I^2 = 69.9$ %, participants = 10.557, studies = 112) for unidimensional outcomes and -0.32 (-0.39 to -0.24 95% CI, $p < 0.0001$, $T^2 = 0.08$, $I^2 = 62.06$ %, participants = 9.447, studies = 95), indicating small to moderate placebo responses.

In the following, between-group differences will be presented per therapy subgroup and meta-regression analyses examining the role of different variables in predicting heterogeneity in trial outcomes.

Spinal manipulation subgroup

With thirty-five comparisons between an active treatment and a sham control, the overall effect in spinal manipulation trials was -0.36 (SMC) in favour of treatment groups for unidimensional pain measures (-0.51 to -0.21 95% CI, 3.084 participants, studies = 35, $T^2 = 0.14$, $I^2 = 71.1\%$) (Figure 3) and -0.26 for multidimensional measures (-0.37 to -0.15 95% CI, participants = 2.384, reported in 24 studies, $T^2 = 0.02$, $I^2 = 26.7\%$) (Figure 4).

15

Figure 3: Spinal manipulation trials, unidimensional outcome measures as standardised mean changes with risk-of-bias assessment per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Arens 2017	−0.679	0.29	2.7%	−0.68 [−1.25, −0.11]
Aspinall 2019	0.04	0.22	3.2%	0.04 [−0.39, 0.47]
Bautista–Auguirre 2017	−0.184	0.26	2.9%	−0.18 [−0.69, 0.33]
Bernal-Utrera 2020	−2.416	0.41	1.9%	−2.42 [−3.22, −1.61]
Bialosky 2014 a	0.114	0.33	2.4%	0.11 [−0.53, 0.76]
Bialosky 2014 b	−0.075	0.33	2.4%	−0.07 [−0.72, 0.57]
Chaibi 2017	0.042	0.26	2.9%	0.04 [−0.47, 0.55]
Corum 2018	−1.193	0.36	2.2%	−1.19 [−1.90, −0.49]
Coste 2021	−0.32	0.23	3.1%	−0.32 [−0.77, 0.13]
da Silva 2019	−0.099	0.26	2.9%	−0.10 [−0.61, 0.41]
Dougherty 2014	−0.31	0.18	3.5%	−0.31 [−0.66, 0.04]
Elleuch 2009	−1.689	0.26	2.9%	−1.69 [−2.20, −1.18]
Espí–López 2014	0.123	0.32	2.4%	0.12 [−0.50, 0.75]
Fisher 2020	−0.043	0.23	3.1%	−0.04 [−0.49, 0.41]
Grimes 2019	−0.296	0.32	2.4%	−0.30 [−0.92, 0.33]
Haik 2014	−0.24	0.28	2.7%	−0.24 [−0.79, 0.31]
Haik 2017	−0.235	0.27	2.8%	−0.23 [−0.76, 0.29]
Kardouni 2015	0.609	0.31	2.5%	0.61 [0.00, 1.22]
Lehtola 2010	−0.468	0.24	3.0%	−0.47 [−0.94, 0.00]
Licciardone 2013	−0.342	0.1	4.1%	−0.34 [−0.54, −0.15]
Martì–Salvador 2018	−1.053	0.27	2.8%	−1.05 [−1.58, −0.52]
Mollins–Cubero 2014	−0.417	0.32	2.4%	−0.42 [−1.04, 0.21]
Muñoz–Gómez 2021	−1.29	0.31	2.5%	−1.29 [−1.90, −0.68]
Nguyen 2021	−0.046	0.11	4.0%	−0.05 [−0.26, 0.17]
Peña–Salinas 2017	−0.272	0.28	2.7%	−0.27 [−0.82, 0.28]
Reynolds 2020	−0.586	0.29	2.7%	−0.59 [−1.15, −0.02]
Schwerla 2008	−0.621	0.34	2.3%	−0.62 [−1.29, 0.05]
Senna 2011	−0.345	0.26	2.9%	−0.34 [−0.85, 0.16]
Sillevis 2010	−0.048	0.2	3.3%	−0.05 [−0.44, 0.34]
Silva 2019	−0.099	0.26	2.9%	−0.10 [−0.61, 0.41]
Thomas 2020	0.03	0.2	3.3%	0.03 [−0.36, 0.42]
Vernon 2012	−0.137	0.25	3.0%	−0.14 [−0.63, 0.35]
Walker 2013	−0.209	0.15	3.7%	−0.21 [−0.50, 0.08]
Young 2009	−0.191	0.24	3.0%	−0.19 [−0.66, 0.28]
Young 2019	−0.893	0.32	2.4%	−0.89 [−1.52, −0.27]
Total (95% CI)			**100.0%**	**−0.36 [−0.51, −0.21]**

Heterogeneity: Tau² = 0.13; Chi² = 117.04, df = 34 (P < 0.00001); I² = 71%
Test for overall effect: Z = 4.79 (P < 0.00001)

Favours [experimental]   Favours [control]

Risk of Bias: A B C D E

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

Figure 4: Spinal manipulation trials, multidimensional outcome measures as standardised mean changes with risk-of-bias assessment per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Arens 2017	0	0		Not estimable
Aspinall 2019	0	0		Not estimable
Bautista–Auguirre 2017	0	0		Not estimable
Bernal–Utrera 2020	−0.75	0.32	2.5%	−0.75 [−1.38, −0.12]
Bialosky 2014 a	−0.29	0.33	2.4%	−0.29 [−0.94, 0.36]
Bialosky 2014 b	−0.18	0.33	2.4%	−0.18 [−0.83, 0.47]
Chaibi 2017	0.32	0.26	3.5%	0.32 [−0.19, 0.83]
Corum 2018	−0.07	0.33	2.4%	−0.07 [−0.72, 0.58]
Coste 2021	−0.1	0.23	4.2%	−0.10 [−0.55, 0.35]
da Silva 2019	0	0		Not estimable
Dougherty 2014	−0.34	0.18	6.0%	−0.34 [−0.69, 0.01]
Elleuch 2009	−0.45	0.22	4.5%	−0.45 [−0.88, −0.02]
Espí–López 2014	0.01	0.32	2.5%	0.01 [−0.62, 0.64]
Fisher 2020	−0.43	0.23	4.2%	−0.43 [−0.88, 0.02]
Grimes 2019	0.05	0.32	2.5%	0.05 [−0.58, 0.68]
Haik 2014	0	0		Not estimable
Haik 2017	−0.14	0.27	3.3%	−0.14 [−0.67, 0.39]
Kardouni 2015	0.16	0.3	2.8%	0.16 [−0.43, 0.75]
Lehtola 2010	0	0		Not estimable
Licciardone 2013	−0.08	0.09	11.6%	−0.08 [−0.26, 0.10]
Marti–Salvador 2018	−0.69	0.25	3.7%	−0.69 [−1.18, −0.20]
Mollins–Cubero 2014	0	0		Not estimable
Muñoz–Gómez 2021	−0.69	0.29	2.9%	−0.69 [−1.26, −0.12]
Nguyen 2021	−0.34	0.11	10.0%	−0.34 [−0.56, −0.12]
Peña–Salinas 2017	0	0		Not estimable
Reynolds 2020	−0.19	0.29	2.9%	−0.19 [−0.76, 0.38]
Schwerla 2008	−0.33	0.33	2.4%	−0.33 [−0.98, 0.32]
Senna 2011	−0.68	0.26	3.5%	−0.68 [−1.19, −0.17]
Sillevis 2010	0	0		Not estimable
Silva 2019	0	0		Not estimable
Thomas 2020	−0.01	0.2	5.2%	−0.01 [−0.40, 0.38]
Vernon 2012	0	0		Not estimable
Walker 2013	−0.36	0.15	7.4%	−0.36 [−0.65, −0.07]
Young 2009	−0.1	0.22	4.5%	−0.10 [−0.53, 0.33]
Young 2019	−0.87	0.32	2.5%	−0.87 [−1.50, −0.24]
Total (95% CI)			**100.0%**	**−0.26 [−0.37, −0.16]**

Heterogeneity: Tau² = 0.02; Chi² = 31.95, df = 23 (P = 0.10); I² = 28%
Test for overall effect: Z = 4.79 (P < 0.00001)

Favours Experimental Favours Sham (axis −2, −1, 0, 1, 2)

Risk of Bias: A B C D E

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

A meta-regression model with the ratings for the following three features was able to predict 59% of the unaccounted between-study variance in unidimensional outcomes: Similarity between active and control groups for the 1) number of treatment sessions, 2) the information about intervention efficacy provided to trial participants, and 3) treatment environment (Q(3,31) = 50.33, $p < 0.0001$, $R^2 = 0.59$, participants = 3.084, studies = 35, residual $T^2 = 0.006$, residual I² = 10.1%) (Table 2). Differences in the number of sessions was the best predictor of differential outcomes. Exclusion of the two studies with the largest effect sizes [32,9] rendered the model non-functional, making necessary the removal of the variable with the least variability (similarity in session number) which then resulted in a model with a reduced but significant predictive value (details not reported in table; Q(2,30) = 7.67, $p = 0.02$, $R^2 = 0.19$, participants = 2.953, studies = 33, residual $T^2 = 0.006$, residual I² = 11.0%). A sensitivity analysis with all spinal manipulation therapy trials irrespective of sample size confirmed the validity of the complete model (details not reported in table; Q(3,38) = 43.86, $p < 0.0001$, $R^2 = 0.48$, participants = 3.240, studies = 42, residual $T^2 = 0.02$, residual I² = 20.3%).

Table 2: Multiple meta-regression analysis, predicting between-study variance in unidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in spinal manipulation trials. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (n of sessions)	Mean (SD) similarity rating (efficacy information)	Mean (SD) similarity rating (environment)	R^2	k		
-.33	1.91 (0.37)	0.6 (1.03)	1.43 (0.74)	.5934	35		
Homogeneity Analysis							
	Q	df	p	T^2	I^2		
Model	50.33	3	.0000	0.077	94.0%		
Residual	34.49	31	.30	0.006	10.1%		
Total	82.15	34	.0000				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	p	Beta
Constant	-2.64	.36	-3.35	-1.93	-7.27	.0000	.0000
Similarity: N of sessions	1.09	.19	. 72	1.47	5.75	.0000	. 65
Similarity: Efficacy information	.13	.05	.02	.23	2.3	.02	.27
Similarity: Environment	.07	.08	-.08	.22	.89	.37	.11
Method of Moments Random Effects Variance Component							
v = .03							

This model was not able to predict a significant proportion of the variance when applied to between-group differences in multidimensional outcome measures, such as pain or disability questionnaires ($Q(3,20) = 3.59$, $p = .31$, $R^2 = .15$, participants = 2.384, studies = 24, residual $T^2 = 0.00$, residual $I^2 = 0\%$, results not shown in table), also not with small studies included as a sensitivity analysis ($Q(3,24) = 3.6$, $p = .31$, $R^2 = .13$, participants = 2.488, studies = 28, residual $T^2 = 0.00$, residual $I^2 = 0\%$). In this model, however, the (dis)similarity in treatment environment was clearly best able to predict outcomes. Simplification of the model to only include this variable improved its ability to account for between-study variance ($Q(1,22) = 2.82$, $p = .09$, $R^2 = .11$, residual $T^2 = 0.00$, $I^2 = 0\%$). It needs to be noted, however, that the unexplained variance in effect sizes was low for multidimensional outcomes ($I^2 = 28\%$, Figure 4), leaving little scope for meta-regression analyses.

Craniosacral therapy trials

The meta-analysis included thirteen studies of craniosacral or gentle myofascial interventions. Effect sizes compared to control interventions were -0.52 (-0.84 to -0.2 95% CI, $p = 0.001$, participants = 1.022, studies = 11, $I^2 = 78.1\%$) for unidimensional measures (Figure 5) and -0.47 (-0.81 to -0.12 95% CI, $p < 0.0001$, participants = 1.162, studies = 13, $I^2 = 83.8\%$) for multidimensional ones (Figure 6).

Figure 5: Trials of craniosacral therapies, unidimensional outcome measures as standardised mean changes with risk-of-bias assessment per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Ajimsha 2014	0	0		Not estimable
Castro–Sanchez 2011 a	−1.067	0.23	9.6%	−1.07 [−1.52, −0.62]
Castro–Sanchez 2011 b	−0.146	0.26	9.1%	−0.15 [−0.66, 0.36]
Cerritelli 2015	−0.819	0.25	9.3%	−0.82 [−1.31, −0.33]
De Groef 2017	−0.773	0.3	8.4%	−0.77 [−1.36, −0.19]
De Groef 2018	0.061	0.17	10.5%	0.06 [−0.27, 0.39]
Ferragut–Garcías 2019	−1.27	0.32	8.1%	−1.27 [−1.90, −0.64]
Haller 2016	−1.251	0.31	8.3%	−1.25 [−1.86, −0.64]
Hensel 2015	−0.028	0.26	9.1%	−0.03 [−0.54, 0.48]
Matarán–Peñarrocha 2009	−0.469	0.22	9.8%	−0.47 [−0.90, −0.04]
Mazreati 2021	0	0		Not estimable
Paulo 2021	0.192	0.22	9.8%	0.19 [−0.24, 0.62]
Simoni 2021	−0.412	0.33	8.0%	−0.41 [−1.06, 0.23]
Total (95% CI)			100.0%	−0.52 [−0.84, −0.20]

Heterogeneity: Tau^2 = 0.22; Chi^2 = 45.26, df = 10 (P < 0.00001); I^2 = 78%
Test for overall effect: Z = 3.22 (P = 0.001)

Favours Experimental Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
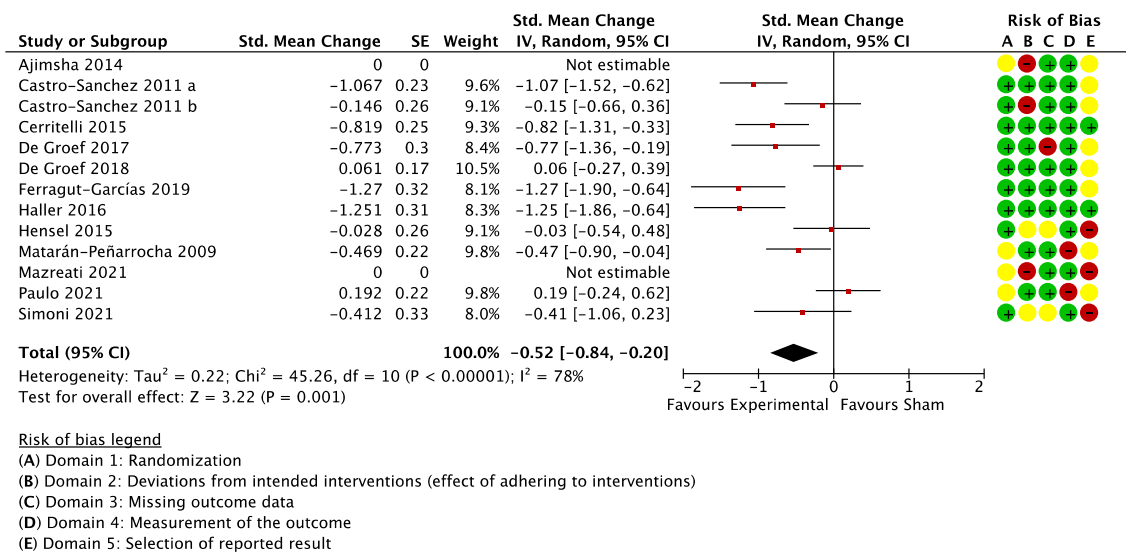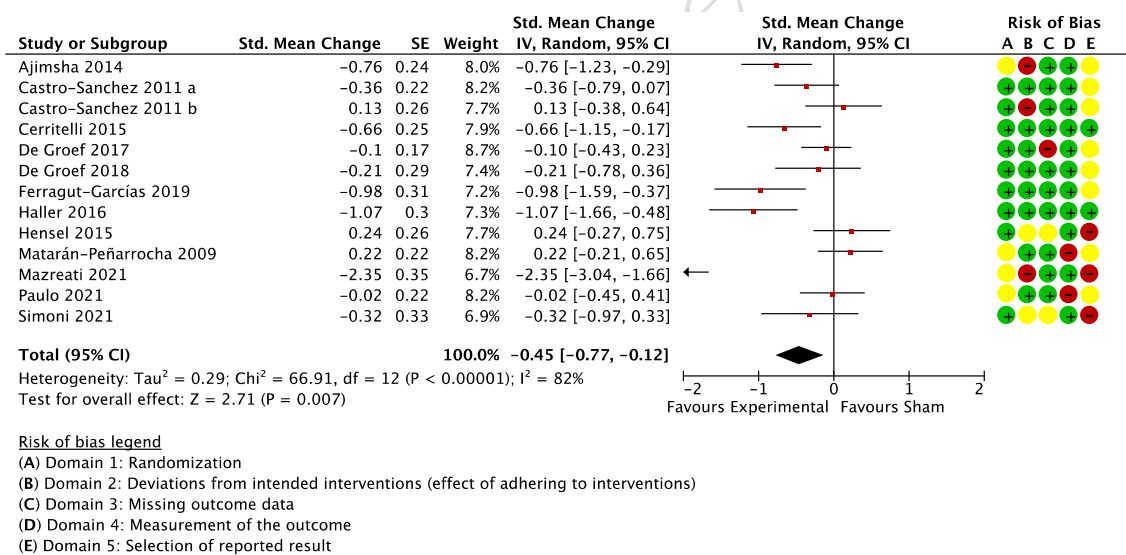(**E**) Domain 5: Selection of reported result

Figure 6: Trials of craniosacral therapies, multidimensional outcome measures as standardised mean changes with risk-of-bias assessment per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Ajimsha 2014	−0.76	0.24	8.0%	−0.76 [−1.23, −0.29]
Castro–Sanchez 2011 a	−0.36	0.22	8.2%	−0.36 [−0.79, 0.07]
Castro–Sanchez 2011 b	0.13	0.26	7.7%	0.13 [−0.38, 0.64]
Cerritelli 2015	−0.66	0.25	7.9%	−0.66 [−1.15, −0.17]
De Groef 2017	−0.1	0.17	8.7%	−0.10 [−0.43, 0.23]
De Groef 2018	−0.21	0.29	7.4%	−0.21 [−0.78, 0.36]
Ferragut–Garcías 2019	−0.98	0.31	7.2%	−0.98 [−1.59, −0.37]
Haller 2016	−1.07	0.3	7.3%	−1.07 [−1.66, −0.48]
Hensel 2015	0.24	0.26	7.7%	0.24 [−0.27, 0.75]
Matarán–Peñarrocha 2009	0.22	0.22	8.2%	0.22 [−0.21, 0.65]
Mazreati 2021	−2.35	0.35	6.7%	−2.35 [−3.04, −1.66]
Paulo 2021	−0.02	0.22	8.2%	−0.02 [−0.45, 0.41]
Simoni 2021	−0.32	0.33	6.9%	−0.32 [−0.97, 0.33]
Total (95% CI)			100.0%	−0.45 [−0.77, −0.12]

Heterogeneity: Tau^2 = 0.29; Chi^2 = 66.91, df = 12 (P < 0.00001); I^2 = 82%
Test for overall effect: Z = 2.71 (P = 0.007)

Favours Experimental Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

During meta-regression modelling for the differences in effects sizes in unidimensional outcomes between craniosacral trials, a model with the similarity ratings 'body areas', 'application mode' and 'efficacy information' was found to predict 74% of the heterogeneity between studies (Table 3). Notably, directions of effects were such that trials showed smaller effects if control interventions consisted of devices (different application mode) and if participants were given different information about effectiveness of the study treatments. When similar body areas were treated in both groups, effect sizes were smaller, too. In this subgroup similarity in the number of sessions could not be used as predictor variable because all studies had the same number of treatment sessions between experimental and control groups. Adding five smaller studies reduced the usefulness of the model (details not reported in table; Q(3,12) = 3.35, p = 0.34, R^2 = 0.15, residual T^2 = 0.04, residual I^2 = 37.9%, participants = 1.146, studies = 16).

Table 3: Multiple meta-regression analysis, predicting between-study variance in unidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in trials of craniosacral and gentle myofascial treatments. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (body areas)	Mean (SD) similarity rating (efficacy information)	Mean (SD) similarity rating (application mode)	R^2	k		
-.48	0.18 (1.9)	0.55 (0.93)	.18 (2.1)	.7299	11		
Homogeneity Analysis							
	Q	df	p	T^2	I^2		
Model	19.25	3	.0002	0.1	84.4%		
Residual	6.77	7	.45	0.00	0%		
Total	26.02	10	.0037				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	p	Beta
Constant	-.37	.11	-.6	-.14	-3.18	.0015	.0
Similarity: Body areas addressed	.17	.06	.06	.28	2.98	.0028	.62
Similarity: Information about treatment effectiveness provided to participants	-.33	.14	-.61	-.05	-2.34	.0195	-.54
Similarity: Application mode	-.06	.06	-.18	.06	-1.04	.2984	-.25
Method of Moments Random Effects Variance Component							
v = .04							

For multidimensional outcome measures, the model predicted 38% of the between-studies variance ($Q(3,9) = 7.0$, $p = .07$, $R^2 = .38$, residual $T^2 = 0.01$, residual $I^2 = 21.6\%$, participants = 1.162, studies = 13, Table 4) and was confirmed in a sensitivity analysis with two additional, smaller studies (details not reported in table; $Q(3,11) = 7.54$, $p = .06$, $R^2 = .37$, residual $T^2 = 0.01$, $I^2 = 15.2\%$, participants = 1209, studies = 15).
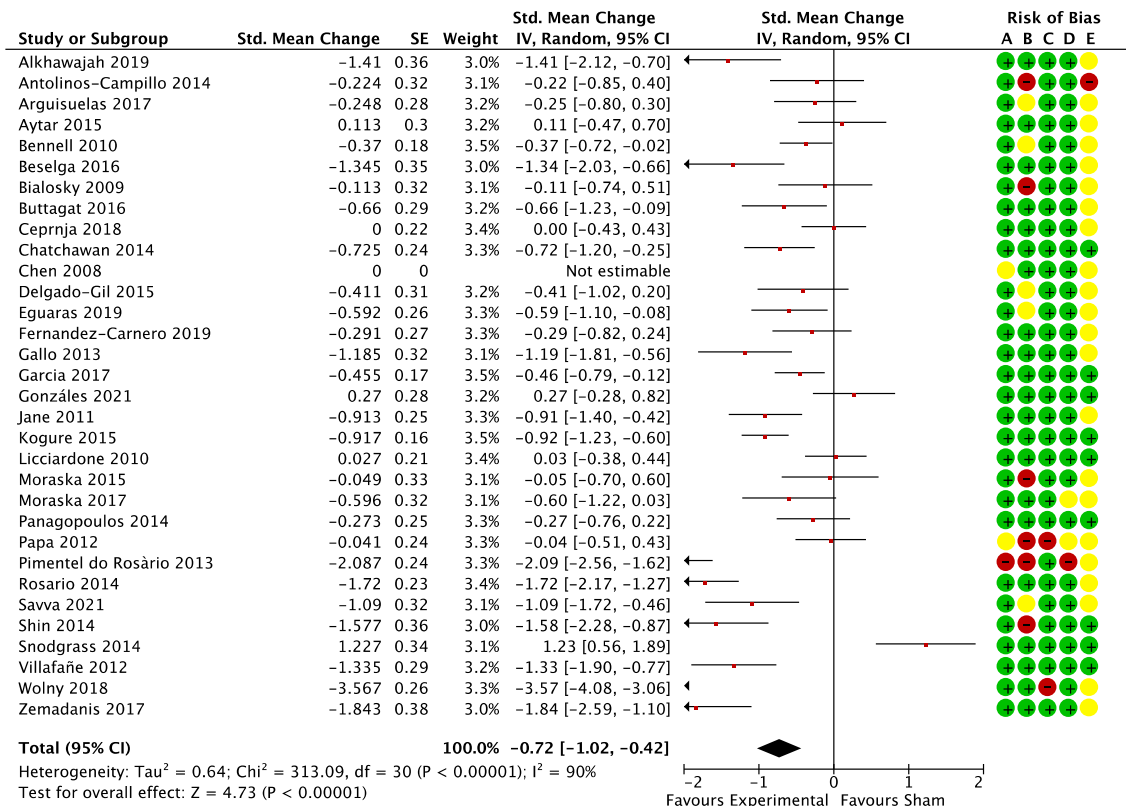
Table 4: Multiple meta-regression analysis, predicting between-study variance in multidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in trials of craniosacral and gentle myofascial treatments. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (body areas)	Mean (SD) similarity rating (efficacy information)	Mean (SD) similarity rating (application mode)	R^2	k		
-.46	0.38 (1.9)	0.46 (0.88)	0.46 (2.03)	.3789	13		
Homogeneity Analysis							
	Q	df	p	T^2	I^2		
Model	7.0	3	.07	0.021	57.1%		
Residual	11.48	9	.24	0.013	21.6%		
Total	18.48	12	.1				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	p	Beta
Constant	-.42	.19	-.79	-.06	-2.25	.02	.0
Similarity: Body areas addressed	.08	.1	-.12	.27	.79	.43	.2
Similarity: Information about treatment effectiveness provided to participants	.07	.22	-.37	.5	.32	.75	.08
Similarity: Application mode	-.25	.1	-.44	-.05	-2.47	.02	-.69
Method of Moments Random Effects Variance Component							
v = .27							

Other manual therapy trials

Meta-analysed trials of other manual therapy encompassed 35 trials of massage, articulation, and manual therapies other than spinal manipulation and craniosacral or gentle myofascial techniques. The combined effect size for unidimensional outcome measures was -0.72 (-1.02 to -0.42 95% CI, $p <$ 0.0001, participants = 2.170, studies = 31, $I^2 = 90\%$) (Figure 7) and -0.45 (-0.68 to -0.22 95% CI, $p <$ 0.0001, participants = 1.647, studies = 22, $I^2 = 79\%$) for multidimensional measures (Figure 8).
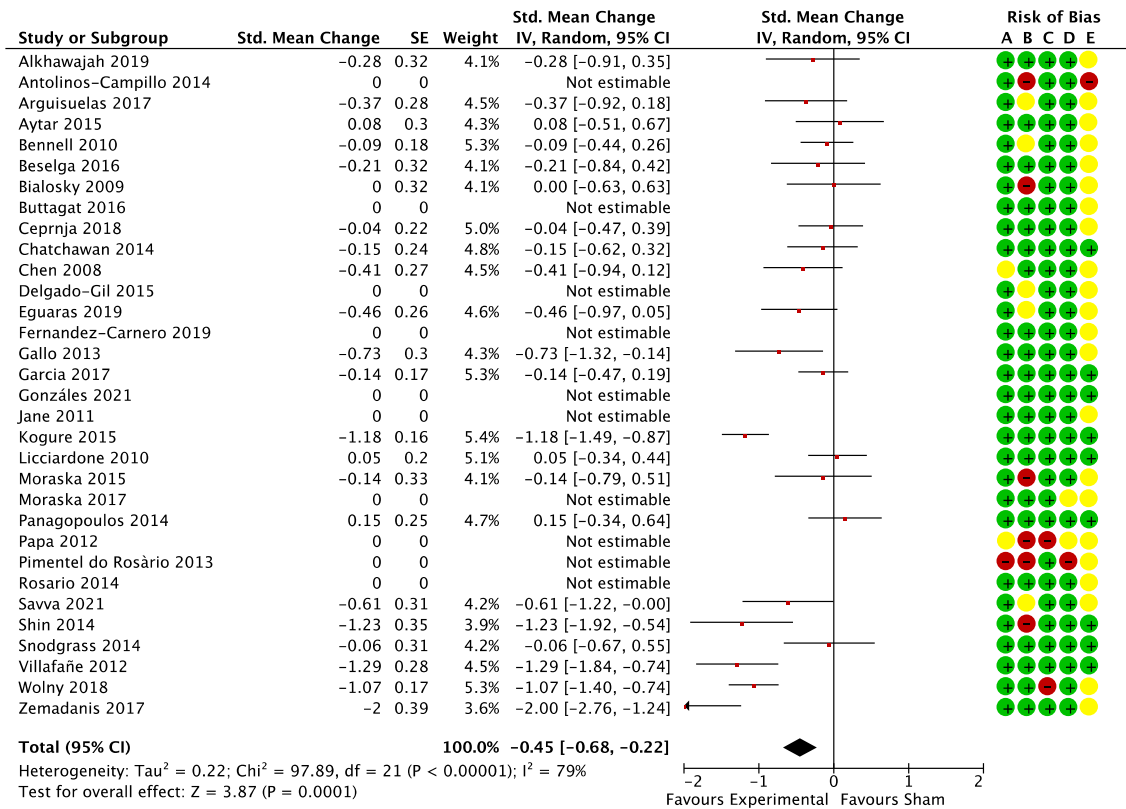
Figure 7: Other manual therapy trials (i.e., excluding spinal manipulation and craniosacral techniques and including massage, articulation etc.), **unidimensional outcome measures** as standardised mean changes with risk-of-bias assessment per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Alkhawajah 2019	-1.41	0.36	3.0%	-1.41 [-2.12, -0.70]
Antolinos-Campillo 2014	-0.224	0.32	3.1%	-0.22 [-0.85, 0.40]
Arguisuelas 2017	-0.248	0.28	3.2%	-0.25 [-0.80, 0.30]
Aytar 2015	0.113	0.3	3.2%	0.11 [-0.47, 0.70]
Bennell 2010	-0.37	0.18	3.5%	-0.37 [-0.72, -0.02]
Beselga 2016	-1.345	0.35	3.0%	-1.34 [-2.03, -0.66]
Bialosky 2009	-0.113	0.32	3.1%	-0.11 [-0.74, 0.51]
Buttagat 2016	-0.66	0.29	3.2%	-0.66 [-1.23, -0.09]
Ceprnja 2018	0	0.22	3.4%	0.00 [-0.43, 0.43]
Chatchawan 2014	-0.725	0.24	3.3%	-0.72 [-1.20, -0.25]
Chen 2008	0	0		Not estimable
Delgado-Gil 2015	-0.411	0.31	3.2%	-0.41 [-1.02, 0.20]
Eguaras 2019	-0.592	0.26	3.3%	-0.59 [-1.10, -0.08]
Fernandez-Carnero 2019	-0.291	0.27	3.3%	-0.29 [-0.82, 0.24]
Gallo 2013	-1.185	0.32	3.1%	-1.19 [-1.81, -0.56]
Garcia 2017	-0.455	0.17	3.5%	-0.46 [-0.79, -0.12]
Gonzáles 2021	0.27	0.28	3.2%	0.27 [-0.28, 0.82]
Jane 2011	-0.913	0.25	3.3%	-0.91 [-1.40, -0.42]
Kogure 2015	-0.917	0.16	3.5%	-0.92 [-1.23, -0.60]
Licciardone 2010	0.027	0.21	3.4%	0.03 [-0.38, 0.44]
Moraska 2015	-0.049	0.33	3.1%	-0.05 [-0.70, 0.60]
Moraska 2017	-0.596	0.32	3.1%	-0.60 [-1.22, 0.03]
Panagopoulos 2014	-0.273	0.25	3.3%	-0.27 [-0.76, 0.22]
Papa 2012	-0.041	0.24	3.3%	-0.04 [-0.51, 0.43]
Pimentel do Rosàrio 2013	-2.087	0.24	3.3%	-2.09 [-2.56, -1.62]
Rosario 2014	-1.72	0.23	3.4%	-1.72 [-2.17, -1.27]
Savva 2021	-1.09	0.32	3.1%	-1.09 [-1.72, -0.46]
Shin 2014	-1.577	0.36	3.1%	-1.58 [-2.28, -0.87]
Snodgrass 2014	1.227	0.34	3.1%	1.23 [0.56, 1.89]
Villafañe 2012	-1.335	0.29	3.2%	-1.33 [-1.90, -0.77]
Wolny 2018	-3.567	0.26	3.3%	-3.57 [-4.08, -3.06]
Zemadanis 2017	-1.843	0.38	3.0%	-1.84 [-2.59, -1.10]
Total (95% CI)			**100.0%**	**-0.72 [-1.02, -0.42]**

Heterogeneity: Tau^2 = 0.64; Chi^2 = 313.09, df = 30 (P < 0.00001); I^2 = 90%
Test for overall effect: Z = 4.73 (P < 0.00001)

Favours Experimental Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

Figure 8: Other manual therapy trials, multidimensional outcome measures as standardised mean changes with risk-of-bias assessment per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI	Risk of Bias A B C D E
Alkhawajah 2019	-0.28	0.32	4.1%	-0.28 [-0.91, 0.35]	
Antolinos-Campillo 2014	0	0		Not estimable	
Arguisuelas 2017	-0.37	0.28	4.5%	-0.37 [-0.92, 0.18]	
Aytar 2015	0.08	0.3	4.3%	0.08 [-0.51, 0.67]	
Bennell 2010	-0.09	0.18	5.3%	-0.09 [-0.44, 0.26]	
Beselga 2016	-0.21	0.32	4.1%	-0.21 [-0.84, 0.42]	
Bialosky 2009	0	0.32	4.1%	0.00 [-0.63, 0.63]	
Buttagat 2016	0	0		Not estimable	
Ceprnja 2018	-0.04	0.22	5.0%	-0.04 [-0.47, 0.39]	
Chatchawan 2014	-0.15	0.24	4.8%	-0.15 [-0.62, 0.32]	
Chen 2008	-0.41	0.27	4.5%	-0.41 [-0.94, 0.12]	
Delgado-Gil 2015	0	0		Not estimable	
Eguaras 2019	-0.46	0.26	4.6%	-0.46 [-0.97, 0.05]	
Fernandez-Carnero 2019	0	0		Not estimable	
Gallo 2013	-0.73	0.3	4.3%	-0.73 [-1.32, -0.14]	
Garcia 2017	-0.14	0.17	5.3%	-0.14 [-0.47, 0.19]	
Gonzáles 2021	0	0		Not estimable	
Jane 2011	0	0		Not estimable	
Kogure 2015	-1.18	0.16	5.4%	-1.18 [-1.49, -0.87]	
Licciardone 2010	0.05	0.2	5.1%	0.05 [-0.34, 0.44]	
Moraska 2015	-0.14	0.33	4.1%	-0.14 [-0.79, 0.51]	
Moraska 2017	0	0		Not estimable	
Panagopoulos 2014	0.15	0.25	4.7%	0.15 [-0.34, 0.64]	
Papa 2012	0	0		Not estimable	
Pimentel do Rosàrio 2013	0	0		Not estimable	
Rosario 2014	0	0		Not estimable	
Savva 2021	-0.61	0.31	4.2%	-0.61 [-1.22, -0.00]	
Shin 2014	-1.23	0.35	3.9%	-1.23 [-1.92, -0.54]	
Snodgrass 2014	-0.06	0.31	4.2%	-0.06 [-0.67, 0.55]	
Villafañe 2012	-1.29	0.28	4.5%	-1.29 [-1.84, -0.74]	
Wolny 2018	-1.07	0.17	5.3%	-1.07 [-1.40, -0.74]	
Zemadanis 2017	-2	0.39	3.6%	-2.00 [-2.76, -1.24]	
Total (95% CI)			100.0%	**-0.45 [-0.68, -0.22]**	

Heterogeneity: Tau^2 = 0.22; Chi^2 = 97.89, df = 21 (P < 0.00001); I^2 = 79%
Test for overall effect: Z = 3.87 (P = 0.0001)

Favours Experimental   Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

When considering the regression model for this subgroup, it is worthwhile noting that certain similarity parameters had little variability, i.e., ratings were consistently high across studies, and were thus unlikely to be of great predictive value in the model. These are 'frequency of treatment' (1.97, 0.18 M, SD), 'assessments' (1.97, 0.18), 'delivery format' (1.97, 0.18), and 'follow-up' (2.0, 0.0).

The model with the best fit included three covariates: Similarity ratings between active and control for 'number of sessions', 'efficacy information', and 'environment'. Nonetheless, this model only predicted 16.3% of the unexplained variance in between-group effect size differences measured in unidimensional outcomes (Table 5). To evaluate how dependent those findings were on the large between-group differences seen in some studies, all studies with confidence intervals that were not overlapping with those of the pooled effect [49,66,67,70,87,35] were removed. This did not enhanced the model's predictive power (details not provided in table; Q(3,16) = 1.07, p = .78, R^2 = .06, residual T^2 = 0.01, residual I^2 = 7.8 %, participants = 1.587, studies = 25), highlighting the (dis)similarity in the treatment environment as the most important contributor to the model. A sensitivity analysis including the smaller studies provided similar results (Q(3,48) = 3.58, p = 0.31, R^2 = .07, residual T^2 = 0.00, residual I^2 = 0%, participants = 2.752, studies = 52).

Table 5: Multiple meta-regression analysis, predicting between-study variance in unidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in other manual therapy trials. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

TABLE 5

Descriptives							
Mean ES	Mean (SD) similarity rating (number of sessions)	Mean (SD) similarity rating (efficacy information)	Mean (SD) similarity rating (environment)	R^2	k		
-.72	1.81 (0.75)	0.58 (0.85)	1.61 (0.62)	.1631	31		
Homogeneity Analysis							
	Q	df	*p*	T^2	I^2		
Model	5.35	3	.1482	0.00	43.9%		
Residual	27.43	27	.4408	0.00	1.6%		
Total	32.78	30	.3324				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	*p*	Beta
Constant	-1.83	.63	-3.07	-.59	-2.9	.0038	.00
Similarity: Environment	.52	.26	.01	1.03	2.0	.0457	.36
Similarity: Efficacy information	.17	.19	-.19	.54	.92	.3578	.16
Similarity: Number of sessions	.1	.22	-.33	. 52	.44	.65	.08
Method of Moments Random Effects Variance Component							
v = .64							

For multidimensional outcome measures, the above model was better-able to predict between-study variance (Q(3,18) = 9.97, *p* = .019, R^2 = .37, residual T^2 = 0.0 , residual I^2 = 0%, participants = 1.647, studies = 22, Table 6), also in a sensitivity analysis with smaller studies that had multidimensional outcome measures available (Q(3,29) = 10.1, *p* = .017, R^2 = .2, residual T^2 = 0.03 , residual I^2 = 27.5%, participants = 1.957, studies = 33) or when an outlier was removed from the pool of large studies [88] (Q (3,17) = 7.66, *p* = .05, R^2 = .34, residual T^2 = 0.00 , I^2 = 0%, participants = 1.607, studies = 21). In either case, the (dis)similarity in the number of sessions was the best predictor of between-study variance.

Table 6: Multiple meta-regression analysis, predicting between-study variance in multidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in other manual therapy trials. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (number of sessions)	Mean (SD) similarity rating (efficacy information)	Mean (SD) similarity rating (environment)	R^2	k		
-.37	1.7 (0.88)	0.5 (0.8)	1.6 (0.7)	.37	22		
Homogeneity Analysis							
	Q	df	*p*	T^2	I^2		
Model	9.97	3	.0188	0.02	69.9 %		
Residual	16.85	18	.5338	0.00	0 %		
Total	26.82	21	.18				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	*p*	Beta
Constant	-.27	.44	-1.13	-.6	-.61	.54	.0
Similarity: Number of sessions	-.34	.14	-.62	-.05	-2.32	.0203	-.47
Similarity: Environment	.29	.19	-.08	.66	1.56	.12	.31
Similarity: Efficacy information	.02	.15	-.27	.32	.17	.87	.03
Method of Moments Random Effects Variance Component							
v = .22							

All manual therapy trials combined

Albeit a more heterogeneous group, we explored if the influential variables from the manual therapy subgroups also had predictive value when applied across spinal manipulation, craniosacral and other manual therapy interventions. Across 77 studies with a total of 6.276 participants (SMC = -0.53, -0.68 to -0.39 95% CI, $p < 0.0001$, $T^2 = 0.22$, $I^2 = 85.3\%$; combined forest plots not presented), a combined model with all five previously identified covariates predicted 16.9% of between-group variance in effect size, with only the similarity ratings for the number of treatment sessions and the treatment environment adding to the model's predictive power in a noteworthy fashion (Table 7). Including all smaller studies (participants total = 7.138, studies = 110), confirmed the model, again highlighting the outstanding impact on effect sizes when the number of treatment sessions or the treatment environment differ between active and sham control groups (Q(5,104) = 21.65, $p = .0006$, $R^2 = .15$, residual $T^2 = 0.012$, residual $I^2 = 13.9\%$). The predictive value of this model was similar for multidimensional outcome measures (Results not shown in figure or table. Large trials only, 59 studies, 5.193 participants; Meta-analysis of combined effect: SMC = -0.35, -0.47 to -0.23 95% CI, $p < 0.0001$, $T^2 = 0.15$, $I^2 = 74.3$; Meta-regression: Q(5,53) = 17.22, $p = .004$, $R^2 = .23$, residual $T^2 = 0.005$, residual $I^2 = 8.72\%$; Sensitivity analysis including small studies: Q(5,70) = 19.35, $p = .0017$, $R^2 = .18$, residual $T^2 = 0.015$, residual $I^2 = 20.4\%$, participants = 5.654, studies = 76).

Table 7: Multiple meta-regression analysis, predicting between-study variance in unidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in a combined sample of all large manual therapy trials. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).
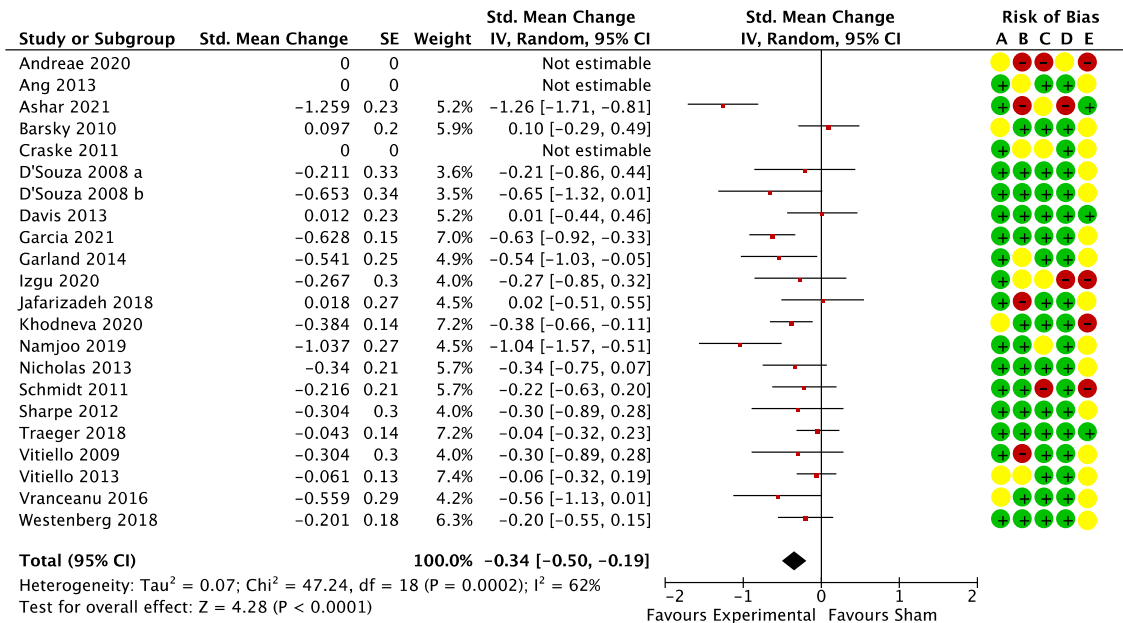
Descriptives								
Mean ES	Mean (SD) similarity rating (number of sessions)	Mean (SD) similarity rating (environment)	Mean (SD) similarity rating (efficacy information)	Mean (SD) similarity rating (application mode)	Mean (SD) similarity rating (body areas addressed)	R^2	k	
-.53	1.88 (0.5)	1.47 (0.8)	1.1 (1.3)	0.34 (1.7)	1.4 (0.8)	.169	77	
Homogeneity Analysis								
	Q	df	*p*	T^2	I^2			
Model	15.94	5	.007	0.009	68.6 %			
Residual	78.24	71	.26	0.006	9.3 %			
Total	94.18	76	.077					
Regression Coefficients								
	B	SE	-95% CI	+95% CI	Z	*p*	Beta	
Constant	-1.57	.31	-2.18	-.95	-5.0	.0000	.0	
Similarity: Number of sessions	..33	.15	.04	.62	2.3	.0239	.24	
Similarity: Environment	.28	.1	.09	.47	2.92	.0035	.31	
Similarity: Efficacy information	.09	.08	-.07	.25	1.13	.26	.12	
Similarity: Application mode	-.04	.04	-.12	.05	-.85	.39	-.09	
Similarity: Body areas	-.03	.05	-.14	.07	-.63	.53	-.31	
Method of Moments Random Effects Variance Component								
v = .33								

Physical therapy interventions, rehabilitation, and exercise

Sixteen studies with a total of 1.554 participants examined rehabilitation and exercise interventions. Aggregated effect sizes were -0.6 (-0.89 to -0.3 95% CI, $p < 0.0001$, $I^2 = 80\%$) for unidimensional measures (Figure 9) and -0.49 (-0.74 to -0.24 95% CI, $p < 0.0001$, $I^2 = 73\%$) for multidimensional outcomes (Figure 10).

Figure 9: Physical therapy and rehabilitation intervention trials. Between-group differences in unidimensional outcome measures are presented as standardized mean change scores alongside risk-of-bias assessments per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Albert 2012	−0.249	0.15	10.1%	−0.25 [−0.54, 0.04]
Arcos–Carmona, 2011	0	0		Not estimable
Barbosa 2019	0.077	0.34	7.1%	0.08 [−0.59, 0.74]
Bennell 2014	−0.37	0.18	9.6%	−0.37 [−0.72, −0.02]
Cacchio 2009	−1.619	0.37	6.7%	−1.62 [−2.34, −0.89]
Costa 2009	−0.5	0.16	9.9%	−0.50 [−0.81, −0.19]
Ganderton 2018	0	0		Not estimable
Holmgren 2012	−0.46	0.22	9.0%	−0.46 [−0.89, −0.03]
Kim 2020	−0.71	0.31	7.6%	−0.71 [−1.32, −0.10]
Krauss 2014	0	0		Not estimable
Messier 2021	0.104	0.14	10.2%	0.10 [−0.17, 0.38]
Mulroy 2011	−1.281	0.29	7.9%	−1.28 [−1.85, −0.71]
Shem 2020	−0.118	0.33	7.3%	−0.12 [−0.76, 0.53]
Wang 2009	0	0		Not estimable
Wang 2010	−0.879	0.27	8.2%	−0.88 [−1.41, −0.35]
Zemadanis 2017	−1.843	0.38	6.5%	−1.84 [−2.59, −1.10]
Total (95% CI)			**100.0%**	**−0.60 [−0.89, −0.30]**

Heterogeneity: Tau^2 = 0.20; Chi^2 = 55.74, df = 11 (P < 0.00001); I^2 = 80%
Test for overall effect: Z = 3.99 (P < 0.0001)

Favours Experimental   Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

Figure 10: Physical therapy and rehabilitation intervention trials. Between-group differences in multidimensional outcome measures are presented as standardized mean change scores alongside risk-of-bias assessments per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Albert 2012	0.03	0.15	9.7%	0.03 [−0.26, 0.32]
Arcos–Carmona, 2011	−0.6	0.28	7.2%	−0.60 [−1.15, −0.05]
Barbosa 2019	0	0		Not estimable
Bennell 2014	−0.01	0.2	8.7%	−0.01 [−0.40, 0.38]
Cacchio 2009	−2.31	0.41	5.1%	−2.31 [−3.11, −1.51]
Costa 2009	−0.36	0.16	9.5%	−0.36 [−0.67, −0.05]
Ganderton 2018	−0.26	0.22	8.3%	−0.26 [−0.69, 0.17]
Holmgren 2012	−0.4	0.21	8.5%	−0.40 [−0.81, 0.01]
Kim 2020	−1.11	0.33	6.3%	−1.11 [−1.76, −0.46]
Krauss 2014	−0.19	0.18	9.1%	−0.19 [−0.54, 0.16]
Messier 2021	0	0		Not estimable
Mulroy 2011	−0.38	0.27	7.4%	−0.38 [−0.91, 0.15]
Shem 2020	−0.22	0.33	6.3%	−0.22 [−0.87, 0.43]
Wang 2009	−0.77	0.33	6.3%	−0.77 [−1.42, −0.12]
Wang 2010	−0.84	0.27	7.4%	−0.84 [−1.37, −0.31]
Zemadanis 2017	0	0		Not estimable
Total (95% CI)			**100.0%**	**−0.49 [−0.74, −0.24]**

Heterogeneity: Tau^2 = 0.14; Chi^2 = 44.36, df = 12 (P < 0.0001); I^2 = 73%
Test for overall effect: Z = 3.90 (P < 0.0001)

Favours Experimental   Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

Surprisingly, in this subgroup similarity regarding the extent of treatment exposure ('number of sessions') was not found to reliably inform a potential regression model to predict between-group effect sizes, even though large variation existed between studies in how similar active and control groups were for the amount of treatment received; differences in the treatment environment were also not predictive. Instead, (dis)similarity in treatment individualisation and the level of fidelity monitoring predicted 86%

of between-group variance in unidimensional outcomes (Table 8). This was confirmed when adding three studies with less than 20 participants per arm (Q(2,11) = 13.9, p = .001, R^2 = .61, residual T^2 = 0.0, residual I^2 = 0.0 %, participants = 1.251, studies = 15). The small number of studies in this subgroup prevented the addition of further variables to a single model.

Table 8: Multiple meta-regression analysis, predicting between-study variance in unidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in a sample of all large physiotherapy and rehabilitation exercise trials. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).
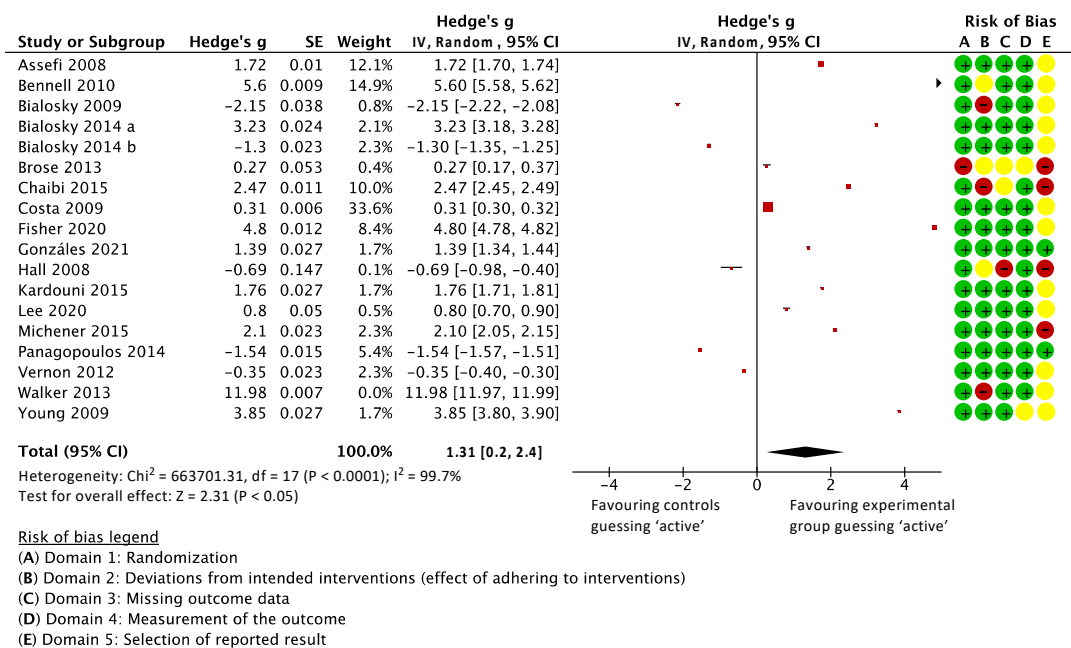
Descriptives							
Mean ES	Mean (SD) similarity rating (individualisation)	Mean (SD) similarity rating (fidelity monitoring)	R^2	k			
-.46	-0.17 (1.3)	1.0 (1.1)	.62	12			
Homogeneity Analysis							
	Q	df	p	T^2	I^2		
Model	12.28	2	.0022	0.048	83.7%		
Residual	7.67	9	.5678	0.00	0%		
Total	19.95	11	.046				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	p	Beta
Constant	-.73	.17	-1.07	-.39	-4.23	.0000	.0
Similarity: Individualisation	-.24	.11	-.45	-.3	-2.2	.0276	-.56
Similarity: Fidelity monitoring	.18	.13	-.07	.42	1.39	.1647	.35
Method of Moments Random Effects Variance Component							
v = .11							

There were more studies that provided multidimensional outcome measures than unidimensional outcomes in the rehabilitation and exercise group, resulting in 14 large trials for this analysis (1.257 participants). The same model with similarity in fidelity monitoring and treatment individualisation predicted a non-significant proportion (7%) of the variance in between-group differences (Q(2,11) = 1.1, p = 0.57, R^2 = .1, residual T^2 = 0.015, residual I^2 = 25.1%, no table provided); The model was further-weakened by removing an outlier [20] (Q(2,10) = 0.0, p = 1.0, R^2 = .00, residual T^2 = 0.00, residual I^2 = 2.7%). Adding the two small trials available in this group made little difference (Q(2,12) = 1.14, p = .57, R^2 = .08, residual T^2 = 0.01, I^2 = 20.8%, participants = 1.314, studies = 16). Similarity in the number of treatment sessions or treatment environment were again not found to provide any predictive value.

Psychological interventions

Nineteen meta-analysed large trials studied psychological or behavioural interventions and employed unidimensional outcomes (effect size -0.34 [-0.50 to -0.19 95% CI], p < 0.0001, I^2 = 62%, participants = 2.085, Figure 11) and multidimensional measures (-0.2 [-0.32 to -0.08 95% CI], p = 0.001, I^2 = 39%, participants = 2.122, studies = 18, Figure 12). One study [59] was removed as an outlier before meta-analysis of multidimensional outcomes.

Figure 11: Psychological and behavioural interventions. Between-group differences in unidimensional outcome measures are presented as standardized mean change scores alongside risk-of-bias assessments per study.

Study or Subgroup	Std. Mean Change	SE	Weight	Std. Mean Change IV, Random, 95% CI
Andreae 2020	0	0		Not estimable
Ang 2013	0	0		Not estimable
Ashar 2021	−1.259	0.23	5.2%	−1.26 [−1.71, −0.81]
Barsky 2010	0.097	0.2	5.9%	0.10 [−0.29, 0.49]
Craske 2011	0	0		Not estimable
D'Souza 2008 a	−0.211	0.33	3.6%	−0.21 [−0.86, 0.44]
D'Souza 2008 b	−0.653	0.34	3.5%	−0.65 [−1.32, 0.01]
Davis 2013	0.012	0.23	5.2%	0.01 [−0.44, 0.46]
Garcia 2021	−0.628	0.15	7.0%	−0.63 [−0.92, −0.33]
Garland 2014	−0.541	0.25	4.9%	−0.54 [−1.03, −0.05]
Izgu 2020	−0.267	0.3	4.0%	−0.27 [−0.85, 0.32]
Jafarizadeh 2018	0.018	0.27	4.5%	0.02 [−0.51, 0.55]
Khodneva 2020	−0.384	0.14	7.2%	−0.38 [−0.66, −0.11]
Namjoo 2019	−1.037	0.27	4.5%	−1.04 [−1.57, −0.51]
Nicholas 2013	−0.34	0.21	5.7%	−0.34 [−0.75, 0.07]
Schmidt 2011	−0.216	0.21	5.7%	−0.22 [−0.63, 0.20]
Sharpe 2012	−0.304	0.3	4.0%	−0.30 [−0.89, 0.28]
Traeger 2018	−0.043	0.14	7.2%	−0.04 [−0.32, 0.23]
Vitiello 2009	−0.304	0.3	4.0%	−0.30 [−0.89, 0.28]
Vitiello 2013	−0.061	0.13	7.4%	−0.06 [−0.32, 0.19]
Vranceanu 2016	−0.559	0.29	4.2%	−0.56 [−1.13, 0.01]
Westenberg 2018	−0.201	0.18	6.3%	−0.20 [−0.55, 0.15]
Total (95% CI)			100.0%	−0.34 [−0.50, −0.19]

Heterogeneity: Tau^2 = 0.07; Chi^2 = 47.24, df = 18 (P = 0.0002); I^2 = 62%
Test for overall effect: Z = 4.28 (P < 0.0001)

Favours Experimental Favours Sham

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

Figure 12: Psychological and behavioural interventions. Between-group differences in multidimensional outcome measures are presented as standardized mean change scores alongside risk-of-bias assessments per study. The study by Namjoo et al. [59] is illustrated here but was excluded from the meta-analysis as an outlier (weight in the model = 0%).
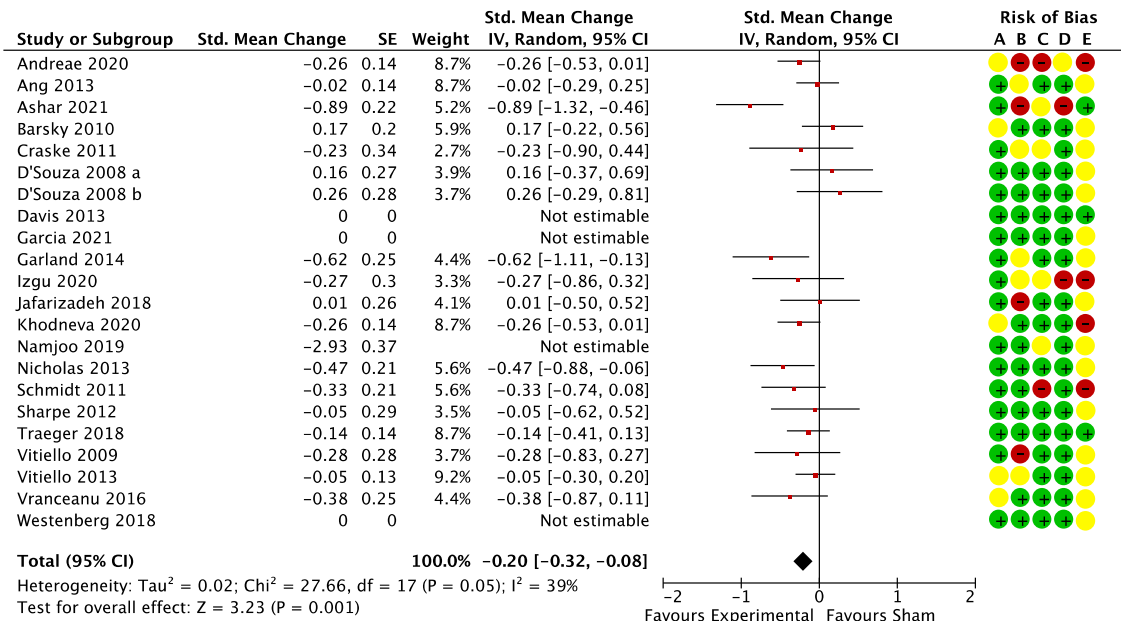
Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

A multiple meta-regression model with the variables 'application mode' and 'attention focus/cognitive function' predicted 41.3% of the remaining variance between studies in unidimensional outcomes (Table 9). The model also predicted between-study variance when a small study was added ($Q(2,17) = 11.69$, $p = .0029$, $R^2 = .41$, residual $T^2 = 0.00$, $I^2 = 0.3\%$, participants = 2.119, studies = 20).

Table 9: Multiple meta-regression analysis, predicting between-study variance in unidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in large trials of psychological and behavioural interventions. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (application mode)	Mean (SD) similarity rating (attention focus / cognitive function)	R^2	k			
-.33	0.84 (1.7)	-0.26 (1.6)	.41	19			
Homogeneity Analysis							
	Q	df	*p*	T^2	I^2		
Model	11.48	2	.0032	0.02	82.6%		
Residual	16.3	16	.4222	0.00	1.8%		
Total	27.78	18	.0654				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	*p*	Beta
Constant	-.4	.08	-.55	-.24	-4.97	.0000	.0
Similarity: Attention	.09	.04	.006	.17	2.1	.036	.41
Similarity: Application mode	.09	.04	0.002	.17	2.	.045	.41
Method of Moments Random Effects Variance Component							
v = .03							

The same model predicted 54% of unexplained heterogeneity in multidimensional outcome measures (Table 10), confirmed by a sensitivity analysis with an additional small study (Q(2,17) = 6.92, *p* = 0.031, residual T^2 = 0.03, residual I^2 = 39.9%, participants = 2.241, studies = 20).

Table 10: Multiple meta-regression analysis, predicting between-study variance in multidimensional outcomes by means of the level of similarity between active and sham control groups for a number of selected features in large trials of psychological and behavioural interventions. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (application mode)	Mean (SD) similarity rating (attention focus / cognitive function)	R^2	k			
-.3	0.95 (1.5)	-0.26 (1.6)	.21	19			
Homogeneity Analysis							
	Q	df	*p*	T^2	I^2		
Model	7.18	2	.0276	0.01	72.1%		
Residual	27.53	16	.036	0.03	41.9 %		
Total	34.71	18	.0103				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	*p*	Beta
Constant	-.49	.13	-.73	-.24	-3.84	.0001	.0
Similarity: Attention	.02	.06	-.1	.15	.36	.7219	.06
Similarity: Application mode	.18	.07	0.04	.31	2.515	.0109	.44
Method of Moments Random Effects Variance Component							
v = .13							

Blinding index subgroup

Bang's blinding index could be calculated for 18 comparisons between an active and a control intervention. In this subgroup, there were nine spinal manipulation, seven other manual therapy trials and one trial each of physiotherapy/rehabilitation and spiritual healing (Reiki). The average sample size at randomization was 64.4 patients (SD 37.4, range 10 to 154).

As per Colagiuri et al. [25], the blinding indices of each trial's two groups were combined as a ratio using Hedge's g, with values larger than 0 indicating that participants in the active group were more likely to correctly guess their allocation to the active group than those in the control group, and values below 0 indicating that patients in the sham control group were more likely to wrongly guess that they received an active treatment compared to those in the active group. Having excluded one outlier where blinding was fully unsuccessful [81], the combined Hedge's g was 1.31 favouring participants in the experimental groups guessing correctly over control participants believing to have received active treatment (unsuccessful or unbalanced blinding (0.2 to 2.43 95% CI, $p = 0.02$, $I^2 = 99.7\%$, participants = 1.013, studies = 17). Data are presented as a forest plot below (Figure 13).

Figure 13: Forest plot showing all studies for which Bang's blinding index (BI) could be calculated. The ratio between BI in active and control groups is presented as Hedge's g, with values near 0 indicating that participants in both groups were likely to make similar guesses as to which treatment they received (i.e., were adequately blinded). Values below 0 indicate that patients in the sham arm thought disproportionately more that they were in the active arm than those in the actual active group (indicating higher sham credibility), and values to the right indicating that the active treatment made more patients think that they had a real treatment than the sham treatment in the same study (in other words, fewer sham-arm patients thought that theirs was a real treatment than active-arm patients did about their intervention, indicating unsuccessful blinding). An extreme outlier (Walker et al., 2013) [69] is shown in the forest plot but does not feed into the meta-analysis (weight = 0%).

Study or Subgroup	Hedge's g	SE	Weight	Hedge's g IV, Random, 95% CI
Assefi 2008	1.72	0.01	12.1%	1.72 [1.70, 1.74]
Bennell 2010	5.6	0.009	14.9%	5.60 [5.58, 5.62]
Bialosky 2009	-2.15	0.038	0.8%	-2.15 [-2.22, -2.08]
Bialosky 2014 a	3.23	0.024	2.1%	3.23 [3.18, 3.28]
Bialosky 2014 b	-1.3	0.023	2.3%	-1.30 [-1.35, -1.25]
Brose 2013	0.27	0.053	0.4%	0.27 [0.17, 0.37]
Chaibi 2015	2.47	0.011	10.0%	2.47 [2.45, 2.49]
Costa 2009	0.31	0.006	33.6%	0.31 [0.30, 0.32]
Fisher 2020	4.8	0.012	8.4%	4.80 [4.78, 4.82]
Gonzáles 2021	1.39	0.027	1.7%	1.39 [1.34, 1.44]
Hall 2008	-0.69	0.147	0.1%	-0.69 [-0.98, -0.40]
Kardouni 2015	1.76	0.027	1.7%	1.76 [1.71, 1.81]
Lee 2020	0.8	0.05	0.5%	0.80 [0.70, 0.90]
Michener 2015	2.1	0.023	2.3%	2.10 [2.05, 2.15]
Panagopoulos 2014	-1.54	0.015	5.4%	-1.54 [-1.57, -1.51]
Vernon 2012	-0.35	0.023	2.3%	-0.35 [-0.40, -0.30]
Walker 2013	11.98	0.007	0.0%	11.98 [11.97, 11.99]
Young 2009	3.85	0.027	1.7%	3.85 [3.80, 3.90]
Total (95% CI)			100.0%	1.31 [0.2, 2.4]

Heterogeneity: Chi^2 = 663701.31, df = 17 (P < 0.0001); I^2 = 99.7%
Test for overall effect: Z = 2.31 (P < 0.05)

Risk of bias legend
(**A**) Domain 1: Randomization
(**B**) Domain 2: Deviations from intended interventions (effect of adhering to interventions)
(**C**) Domain 3: Missing outcome data
(**D**) Domain 4: Measurement of the outcome
(**E**) Domain 5: Selection of reported result

A meta-regression model with the similarity variables 'sensation' and 'application mode' (e.g., manual vs. device) was able to predict 35% of the unexplained variance between studies (Q(2,14) = 8.92, p = .012, R^2 = .35, residual T^2 = 0.003, residual I^2 = 15.1%, participants = 1.013, studies = 17, Table 10).

Notably, this model was primarily driven by differences in the sensations participants were exposed to. This effect that would have been reverted to an emphasis on differences in application modes by including the Walker et al. [81] study (participants = 183), where detuned ultrasound, a hand-held device delivering low-force impulses to paraspinal tissues, and random hand placements on the patient's back were used as a sham control for individualised chiropractic treatments (Q(2,15) = 12.0, R^2 = .54, p = 0.002, residual T^2 = 0.00, residual I^2 = 0%). The apparent importance of choosing similar application modes and producing similar sensations in control interventions to achieve balanced blinding was further-emphasised by excluding another study with largely unsuccessful blinding: Bennell et al. (2010) used detuned ultrasound and a non-therapeutic gel to control for standardised manual therapy and a home exercise programme for shoulder pain in 120 participants (Q(2,13) = 2.25, R^2 = .14, p = 0.32, residual T^2 = 0.001, residual I^2 = 4.8%, participants = 893, studies = 16). It is also noteworthy that the trials in this subgroup exposed participants in either group to the same extent of treatment (e.g., 'number of sessions') and treatment environment, so that these variables could not be tested for.

Table 11: Multiple meta-regression analysis, predicting between-study variance in the success of blinding by means of the level of similarity between active and sham control groups for a number of selected features in all trials where Bang's blinding index could be calculated (8 spinal manipulation trials, 7 other manual therapy, 1 physiotherapy and rehabilitation exercise, 1 Reiki intervention). Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean ES	Mean (SD) similarity rating (sensations)	Mean (SD) similarity rating (application mode)	R^2	k			
1.32	1.24 (1.15)	1.71 (0.99)	.35	17			
Homogeneity Analysis							
	Q	df	p	T^2	I^2		
Model	8.92	2	.0116	0.008	77.6 %		
Residual	16.48	14	.2847	0.003	15.1 %		
Total	25.4	16	.0631				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	p	Beta
Constant	3.0	.87	1.27	4.7	3.42	.0006	.0
Similarity: Sensations	-.97	.52	-2.0	.05	-1.87	.06	-.5
Similarity: Application mode	-.28	.6	-1.46	.91	-.46	.65	-.12
Method of Moments Random Effects Variance Component							
v = 3.09							

In fifteen trials (1.084 participants), both Bang's BI and pain-related outcome data were available (estimated aggregated effect size for unidimensional outcomes: -0.22 [-0.39 to -0.04 95 % CI, p = 0.015, T^2 = 0.05, I^2 = 43.7%]; multidimensional outcomes: -0.23 [-0.36 to -0.09 95% CI, p = 0.001, T^2 = 0.00, I^2 = 9.1%], participants = 1.019, studies = 14). Meta-regression with the BI ratio as a potential predictor of between-group differences showed no relationship between the two for unidimensional outcome measures (Q(1,13) = .01, R^2 = .001, p = .9) or multidimensional outcomes (Q(1,12) = 1.5, R^2 = .12, p = .23), which was unsurprising given the small between-study heterogeneity in outcomes.

Differential attrition and trial outcomes

Differential attrition did not predict significant between-study variance in trial outcomes in the combined sample for unidimensional outcomes ($Q(1,148) = 3.2$, $R^2 = .02$, $p = .07$, residual $T^2 = 0.02$, residual $I^2 = 26.4\%$, participants = 11.009, studies = 150) or multidimensional measures ($Q(1,115) = .0$, $R^2 = .0$, $p = .9$, residual $T^2 = 0.03$, residual $I^2 = 31.3\%$, participants = 9.868, studies = 117).

Differential attrition and similarity between active and sham control interventions

In a random-effects meta-regression model with all studies weighted equally, a model with six similarity ratings was able to significantly predict 12.7 % of the variation in differential dropouts, with differences in fidelity monitoring contributing most to the model's predictive value (Table 12). A sensitivity analysis with all studies irrespective of sample size confirmed the model, with differences in treatment environment now increasing their contribution to the model ($Q(6,143) = 13.2$, $R^2 = .08$, $p = .04$, residual $T^2 = 0.0$, residual $I^2 = 0.0\%$, participants = 11.829, studies = 150, no detail provided in table).

Table 12: Multiple meta-regression analysis, predicting between-study variance in the level of differential attrition by means of the level of similarity between active and sham control groups for a number of selected features. Means and variance of the similarity ratings included in the model are provided, having a possible range of -2 (different for all studies) to 2 (similar across all studies).

Descriptives							
Mean (SD) similarity ratings	n of sessions	Standardisation	Fidelity monitoring	Individualisation	Application mode	Participation	Treatment Environment
	1.69 (0.9)	0.97 (1.2)	0.61 (1.0)	0.5 (1.5)	0.8 (1.8)	0.9 (1.5)	1.2 (1.1)
Mean ES				**R^2**	k		
.3068				.127	111		
Homogeneity Analysis							
	Q	df	*p*	**T^2**	**I^2**		
Model	15.1	6	.0197	0.08	60.2%		
Residual	104.	104	.48	0.00	0%		
Total	119.1	110	.26				
Regression Coefficients							
	B	SE	-95% CI	+95% CI	Z	*p*	Beta
Constant	-1.13	1.5	-4.1	1.83	-.75	.45	.0
Similarity: Fidelity monitoring	-2.14	.68	-3.5	-.8	-3.13	.0017	-.3
Similarity: Application mode	.44	.45	-.45	1.32	1.0	.34	.12
Similarity: Environment	.87	.7	-.51	2.25	1.23	.22	.13
Similarity: N of sessions	.62	.86	-1.06	2.29	.72	.47	.08
Similarity: Standardisation	.57	.59	-.59	1.72	.96	.34	.1
Similarity: Participation	-.16	.53	-1.19	.88	-.29	.77	-.03
Method of Moments Random Effects Variance Component							
v = 46.5							

Discussion

Analysing 194 publications, we found moderate placebo responses across physical, psychological, and self-management therapies for pain, with patients in sham control groups improving with an average effect size of 0.46 from pre-treatment to the first post-treatment assessment. Benefits of experimental interventions over sham controls ranged from effect sizes of 0.34 in psychological interventions (number needed to treat, NNT ≈ 9) to 0.72 in some manual therapies (NNT ≈ 4) [53], with risk-of-bias concerns in most trials. Effect sizes were smaller when multidimensional measures such as disability questionnaires were employed, as opposed to simple pain intensity scales.

In most of the studied intervention subgroups, there was considerable unexplained heterogeneity in trial results between studies. Assessing and rating the level of similarity between experimental and sham control interventions, however, explained some of that heterogeneity. For example, a proportion of variance in trial outcomes was explained by how different tested and control interventions were in terms of the number of treatment sessions, application modes or treatment environment. Further, trials were at a higher risk of differential attrition when there were differences in monitoring of the groups' treatment adherence. In a subgroup of studies where the blinding success of control interventions had been measured, meta-regression analyses were also able to partially predict the risk of unbalanced blinding.

These findings underline the importance of carefully matched experimental and control interventions in efficacy and mechanistic trials. They further provide quantitative evidence that resemblance in some features may be particularly influential. We identified several features that are commonly assessable from trial reports, capture influential aspects of (dis)similarity, and have little conceptual or practical overlap between each other. Namely, these features are: The number of treatment sessions, application mode, intervention individualisation, fidelity monitoring, and the treatment environment. These are presented and discussed in Table 13.

Table13: Overview of sham control design features that have an impact on trial outcomes.

Control intervention feature and description	Influence on trial results and discussion
Number of treatment sessions *The number of times a patient receives the interventions, both through a provider or self-delivered.*	In our preliminary assessments, we found this feature to be highly correlated with, for example, treatment duration and treatment frequency. We thus suspect the number of treatment sessions to be a good proxy for the *extent of treatment* received by patients. It is important to note that differences in the exposure to an intervention often occur not only when patients are asked to attend more clinic appointments than those in the control group, but also when they are provided with home exercise or self-management programmes that expose them to higher 'dosages' of specific and contextual effects.
Application mode *The mode or medium through which an intervention is delivered, including telephone,*	Questioning the suitability of sham devices for blinding in non-device RCTs, differences in the 'tool' via which active and control interventions are applied lead to less credible control interventions. Regarding effects on trial outcomes, the effect can likely go either way, with more elaborate sham controls

online, hands-on or conversation-based delivery. In this sample, device-based active treatments were not eligible for review, but device-based shams (such as detuned ultrasound) were common.	leading to smaller between-group differences [33] or undermining through supposed credibility differences. In other words, sham devices do seem to produce comparable placebo effects to respective active interventions but may compromise trials in other ways. The use of sham devices is still common, especially in physiotherapy and non-spinal manipulation trials (Ref paper 1). In psychological and behavioural intervention, the 'application mode' may differ when control interventions rely on print or web-based materials rather than personal interactions.
Intervention individualisation *The extent to which treatments are personalised or adapted to each patient. May reach from fully standardised to highly individualised.*	Likely somewhat overlapping with the concept of treatment standardisation, we found that the extent to which the active and control interventions are individualised to patients can influence trial outcomes. We can only suspect that the individualisation is communicated to the patient verbally and nonverbally, making the patient feel valued more or suppose that the treatment addresses the root cause of their problem. Contrastingly, patients receiving generic control treatment may wonder why symptomatic areas were not touched or topics not addressed that felt relevant to them. Highly standardised interventions of course offer little room for individualisation.
Patient participation *The level of patients contributing to the intervention, ranging from physically and psychologically passive recipients to largely shaping the intervention themselves.*	Many commonly used control interventions are more passive than the study treatment, especially in physiotherapy, self-management, and exercise trials. Our regression models did not clearly indicate, however, whether different participation levels contribute to differences in pain reports. Whilst patients play an active role in most cognitive, behavioural, and exercise-based interventions for pain, manual therapies can be passive experiences for patients. This allows researchers to employ passive control interventions that match a passive treatment, as reflected by high similarity ratings in this subgroup.
Fidelity monitoring *Monitoring and potentially promoting treatment adherence by patients or therapists' delivery of interventions according to protocols.*	When the treatment adherence or therapists' intervention provision are ensured more in one group than in the other, patients will likely receive different amounts of active and control care. Differences in fidelity monitoring between groups was influential for differential attrition across all trials. In physiotherapy but also in behavioural interventions, patients often perform (parts of) the intervention under their own supervision, for example as home-based exercise programmes. Therefore, fidelity monitoring is more relevant than, for example, in most manual therapy scenarios. Potentially, fidelity monitoring itself also acts as an additional contextual factor, ensuring that patients believe that it is important how and how often the intervention is performed.
Treatment environment	One of the most consistent predictors of between-study variance in pain outcomes, the treatment environment differs, for example, when the active intervention is mainly performed

The physical environment in which interventions take place.	in a healthcare centre and the control intervention at the patient's home, or vice versa [58]. Another example is the study by Arcos-Carmona et al. (2011) where the intervention involved aerobic exercise in a swimming pool whilst the control group received a magnetotherapy sham, lying prone on a treatment bench [2]. Unsurprisingly, this difference in environment comes with large differences in contextual factors that may result in differential placebo effects. Whilst treatment environments were usually similar for all studied therapy types, the consistency with which differences in this feature predicted trial outcomes shows that matching of treatment environments ought to be paramount.

The meta-analysis of similarity features was complicated by insufficient reporting, possibly explaining why some features that are commonly deemed important to be matched between study groups did not appear significant in our analyses. This may apply to the (dis)similarity in co-interventions and concomitant treatments [36] and personal interactions with staff and providers [5,43]. The information provided to participants about the supposed efficacy or rationale of the interventions is also rarely reported or must be inferred. Commonly identified as important in the literature [18,17,77,55], our analyses were nonetheless sensitive to this feature. Further, we were unable to ascertain whether the use of different providers for both groups changed trial results, because most trials used the same providers or did not report this information [41]. Further, there are therapy-specific considerations that cannot be reliably captured in a systematic evidence synthesis, for example the replication of treatment side effects in the control group or the modification of providers' treatment 'styles' to individual patients.

Inconsistency in our meta-regression findings may further be linked to a lack of variability in similarity ratings within a given subgroup, little unexplained heterogeneity in pain-related outcomes, or, of course, the possibility that similarity for a given item did not influence effect sizes sufficiently to detect a link. Relatedly, the patients' experience may be dominated by different treatment aspects depending on therapy type, possibly explaining why certain features only significantly predicted study results in individual therapy types. Lastly, it could be argued that this review's results may be influenced by publication bias, with negative results less likely to be published [65]. However, missing small trials will not have impacted our primary analyses (as these only included trials with more than 20 participants per arm). While we are unable to estimate the impact of missing large trials, we have not made judgements on treatment efficacy so that testing for publication bias was not deemed necessary.

Some of the employed similarity ratings may overlap with supposed 'specific' elements of treatments. If, for example, cognitive distraction is a purportedly integral part of the intervention, then of course the similarity rating will be low for this feature and links to trial outcomes may be found. This may have influenced our findings in the subgroup of psychological and behavioural interventions. In this instance, the question of similarity becomes a mechanistic one, demonstrating that treatment mechanisms need to be considered in the design of control interventions, and ideally reported. Conversely, mechanisms can potentially be studied via our meta-analytic approach of assessing the predictive role of (dis)similarity between active and control interventions for specific features.

Further important insights of our review relate to the nature of pain-related outcomes employed in RCTs. First, distinguishing uni- from multidimensional outcomes enabled us to demonstrate that the latter, more complex outcome measures produce less between-study variance in results, leading to more consistent but smaller effects. Similarly, placebo responses were smaller for multidimensional outcomes. Secondly, multidimensional outcomes appeared less susceptible to contextual effects produced by differences between active and control interventions. In other words, they may allow for less well-matched control interventions. This statement needs to be cautioned, however, because we do not know if there are unidentified confounding factors and because this effect is partly due to the reduced heterogeneity just discussed, weakening our meta-regression models for multidimensional outcome measures.

Apart from the similarity between active and control interventions in efficacy trials, we should consider other sources of heterogeneity in trial outcomes. Patients expectations of benefit with regards to the study treatment and the planned number of provider interactions have been identified as predictors of the placebo response in drug trials [79,69]. Both variables are conceptually related to some of the predictors identified in the present review. Other, psycho-biological predictors are known but have not been tested here, including baseline pain, the nature of the studied painful condition, gender, patient personality traits, and different healthcare settings [51,40,79,55]. Of course, differences in intervention efficacy will have contributed to heterogeneity in effect sizes in our, often diverse, intervention subgroups. The same applies to trial-specific risk of bias, which we illustrated but did not formally include in our modelling. To these known predictors of trial outcomes, we have added the insight that resemblance between test and control interventions matters.

It needs to be noted that the prominent role of blinding in clinical trial research has recently been questioned in opinion papers [1] and by a meta-analysis that did not find differences in outcomes between blinded and non-blinded studies [57]. Our own subgroup analysis in studies reporting on blinding effectiveness indicated a similar direction but small numbers and little between-study heterogeneity prevent firm conclusions. On the other hand, our other analyses clearly demonstrated that trial outcomes partly depended on characteristics of control interventions and their similarity to experimental treatments. Whilst this testing did not consider patients' blinding status, a possible mediator of this finding was the placebo effect, rather than knowledge of group allocation. This is supported also by the fact that features known to drive placebo responses were most predictive, namely the extent of therapeutic interaction, treatment environment, and other features directly affecting the patient experience. Differences in application modes were predictive of pain-related results in some cases, possibly due to differences in placebo responses or because they facilitated unblinding, as was the case in a subgroup of trials where blinding effectiveness was reported. To gain more certainty about the influence of blinding success on trial outcomes, more consistent reporting of blinding effectiveness is required. Nonetheless, we demonstrated that successful blinding is more likely to be achieved with control interventions that resemble experimental treatments. Participant blinding in trials is likely also influenced by factors rarely assessable from trial reports, such as staff's compliance with trial procedures, or contact amongst participants participant recruitment from the same population. Participant blinding is even more challenging in PPS pragmatic and comparative effectiveness trials and is rarely conducted: less than a quarter of these perform participant blinding [42].

Considering our findings and what is known about the power of placebo effects in the absence of blinding ("open-label placebos") [50,12,82,22], it appears that well-matched control interventions are mainly important to prevent skewed trial results in explanatory trials based on different levels of psychosocial contextual factors. Whether this is independent of blinding effectiveness, requires further investigation. A likely mediator of the placebo effect in trials are participants' expectations of benefit

[63]. Here also, our unsuccessful attempt to compare reports of expectancies highlights a need for homogenisation of methods and reporting. The concept needs to be clearly delineated from treatment credibility and satisfaction, appropriate timepoints and methods of expectancy assessment in trials agreed, and outcome data reported [27].

Conclusions

The present review provides quantitative support for the recommendation that experimental and control interventions in efficacy and mechanistic trials should be 'structurally equivalent' [52,64] or 'indistinguishable' [5]. Crucially, this review added the insight that similarity in the extent of intervention exposure, treatment environment, and patient experience are the most important considerations. Across different groups of physical, psychological, and self-management interventions, these factors predicted variability in trial results. Differences in these and several other, sometimes therapy-specific considerations can lead to differences in contextual effects and thus biased trial outcomes. Apart from impacting pain-related outcomes, such differences can undermine participant blinding and promote differential attrition. More work is needed to quantify the effects of blinding failure on pain outcomes – suggesting that there is an urgent need to conduct routine assessments of blinding effectiveness in clinical trials. Efforts to promote adequate reporting of control interventions, such as the TIDieR-Placebo checklist [44], are welcomed. Nonetheless, more work is required to translate the present findings into evidence-based recommendations for the design, testing, and conduct of control interventions in efficacy and mechanistic RCTs of complex physical, psychological, and self-management interventions for people with pain.

Conflict of Interest statement

The authors have no conflict of interest to report regarding this work.

References

[1]   Anand, R, Norrie, J, Bradley, JM, McAuley, DF, Clarke, M. Fool's gold? Why blinded trials are not always best. BMJ 2020:l6228.

[2]   Arcos-Carmona IM, Castro-Sanchez AM, Mataran-Penarrocha GA, Gutierrez-Rubio AB, Ramos-Gonzalez E, Moreno-Lorenzo C. [Effects of aerobic exercise program and relaxation techniques on anxiety, quality of sleep, depression, and quality of life in patients with fibromyalgia: a randomized controlled trial]. Efectos de un programa de ejercicios aerobicos y tecnicas de relajacion sobre el estado de ansiedad, calidad del sueno, depresion y calidad de vida en pacientes con fibromialgia: ensayo clinico aleatorizado 2011;137:398–401.

[3]   Aycock DM, Hayat MJ, Helvig A, Dunbar SB, Clark PC. Essential considerations in developing attention control groups in behavioral research. Res Nurs Health 2018;41:320–328.

[4]   Bang H, Ni L, Davis CE. Assessment of blinding in clinical trials. Controlled Clinical Trials 2004;25:143–156.

[5]   Baskin TW, Tierney SC, Minami T, Wampold BE. Establishing Specificity in Psychotherapy: A Meta-Analysis of Structural Equivalence of Placebo Controls. Journal of Consulting 2003;71:973–979.

[6]   Beard DJ, Campbell MK, Blazeby JM, Carr AJ, Weijer C, Cuthbertson BH, Buchbinder R, Pinkney T, Bishop FL, Pugh J, Cousins S, Harris IA, Lohmander LS, Blencowe N, Gillies K, Probst P, Brennan C, Cook A, Farrar-Hockley D, Savulescu J, Huxtable R, Rangan A, Tracey I, Brocklehurst P, Ferreira ML, Nicholl J, Reeves BC, Hamdy F, Rowley SC, Cook JA. Considerations and methods for placebo controls in surgical trials (ASPIRE guidelines). The Lancet 2020;395:828–838.

[7]   Benedetti F, Amanzio M. Mechanisms of the placebo response. Pulmonary Pharmacology & Therapeutics 2013;26:520–523.

[8]   Bennell K, Wee E, Coburn S, Green S, Harris A, Staples M, Forbes A, Buchbinder R. Efficacy of standardised manual therapy and home exercise programme for chronic rotator cuff disease: randomised placebo controlled trial. BMJ 2010;340:c2756.

[9]   Bernal-Utrera C, Gonzalez-Gerez JJ, Saavedra-Hernandez M, Lerida-Ortega MA, Rodriguez-Blanco C. Manual therapy versus therapeutic exercise in non-specific chronic neck pain: study protocol for a randomized controlled trial. Trials 2019;20:487.

[10]   Bialosky JE, Bishop MD, Price DD, Robinson ME, Vincent KR, George SZ. A randomized sham-controlled trial of a neurodynamic technique in the treatment of carpal tunnel syndrome. The Journal of orthopaedic and sports physical therapy 2009;39:709–23.

[11]   Bialosky JE, George SZ, Horn ME, Price DD, Staud R, Robinson ME. Spinal manipulative therapy-specific changes in pain sensitivity in individuals with low back pain (NCT01168999). The journal of pain : official journal of the American Pain Society 2014;15:136–48.

[12]   Blease CR, Bernstein MH, Locher C. Open-label placebo clinical trials: is it the rationale, the interaction or the pill? BMJ Evidence-Based Medicine 2020;25:159–165.

[13] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. John Wiley & Sons, 2021.

[14] Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P, for the CONSORT NPT Group. CONSORT Statement for Randomized Trials of Nonpharmacologic Treatments: A 2017 Update and a CONSORT Extension for Nonpharmacologic Trial Abstracts. Ann Intern Med 2017;167:40.

[15] Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, for the CONSORT Group. Extending the CONSORT Statement to Randomized Trials of Nonpharmacologic Treatment: Explanation and Elaboration. Ann Intern Med 2008;148:295.

[16] Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. Journal of Clinical Epidemiology 2004;57:543–550.

[17] Braithwaite FA, Walters JL, Moseley GL, Williams MT, McEvoy MP. Towards more credible shams for physical interventions: A Delphi survey. Clinical Trials 2020:1740774520910365.

[18] Braithwaite FA, Walters JL, Moseley GL, Williams MT, McEvoy MP. Towards more homogenous and rigorous methods in sham-controlled dry needling trials: Two Delphi surveys. Physiotherapy 2019. doi:10.1016/j.physio.2019.11.004.

[19] Brose SW, Jennings DC, Kwok J, Stuart CL, O'Connell SM, Pauli HA, Liu B. Sham Manual Medicine Protocol for Cervical Strain-Counterstrain Research. PM&R 2013;5:400–407.

[20] Cacchio A, De Blasis E, De Blasis V, Santilli V, Spacca G. Mirror therapy in complex regional pain syndrome type 1 of the upper limb in stroke patients. Neurorehabilitation and Neural Repair 2009;23:792–9.

[21] Campbell MK, Entwistle VA, Cuthbertson BH, Skea ZC, Sutherland AG, McDonald AM, Norrie JD, Carlson RV, Bridgman S, KORAL study group. Developing a placebo-controlled trial in surgery: issues of design, acceptability and feasibility. Trials 2011;12:50.

[22] Carvalho C, Pais M, Cunha L, Rebouta P, Kaptchuk TJ, Kirsch I. Open-label placebo for chronic low back pain: a 5-year follow-up. PAIN 2021;162:1521–1527.

[23] Cerritelli F, Verzella M, Cicchitti L, D'Alessandro G, Vanacore N. The paradox of sham therapy and placebo effect in osteopathy. Medicine (Baltimore) 2016;95. doi:10.1097/MD.0000000000004728.

[24] Chaibi A, Šaltytė Benth J, Bjørn Russell M. Validation of Placebo in a Manual Therapy Randomized Controlled Trial. Sci Rep 2015;5:11774.

[25] Colagiuri B, Sharpe L, Scott A. The Blind Leading the Not-So-Blind: A Meta-Analysis of Blinding in Pharmacological Trials for Chronic Pain. The Journal of Pain 2019;20:489–500.

[26] Colloca L. The Placebo Effect in Pain Therapies. Annu Rev Pharmacol Toxicol 2019;59:191–211.

[27] Constantino MJ, Vîslă A, Coyne AE, Boswell JF. A meta-analysis of the association between patients' early treatment outcome expectation and their posttreatment outcomes. Psychotherapy 2018;55:473–485.

[28] Coste J, Medkour T, Maigne J-Y, Pérez M, Laroche F, Perrot S. Osteopathic medicine for fibromyalgia: a sham-controlled randomized clinical trial. Therapeutic Advances in Musculoskeletal 2021;13:1759720X211009017.

[29] Curtis P, Gaylord SA, Park J, Faurot KR, Coble R, Suchindran C, Coeytaux RR, Wilkinson L, Mann JD. Credibility of low-strength static magnet therapy as an attention control intervention for a randomized controlled study of CranioSacral therapy for migraine headaches. Journal of alternative and complementary medicine (New York, NY) 2011;17:711–21.

[30] Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ 2013;346. doi:10.1136/bmj.f2304.

[31] Eccleston C, Hearn L, Williams AC de C. Psychological therapies for the management of chronic neuropathic pain in adults. Cochrane Database of Systematic Reviews 2015. doi:10.1002/14651858.CD011259.pub2.

[32] Elleuch M, Ghroubi S. Vertebral manipulation in chronic low back pain: a prospective randomised study of 85 cases. International Musculoskeletal Medicine 2009;31:57–62.

[33] Faltinsen E, Todorovac A, Bruun LS, Hróbjartsson A, Gluud C, Kongerslev MT, Simonsen E, Storebø OJ. Control interventions in randomised trials among people with mental health disorders. Cochrane Database of Systematic Reviews 2022. doi:10.1002/14651858.MR000050.pub2.

[34] Felson DT, Redmond AC, Chapman GJ, Smith TO, Hamilton DF, Jones RK, Holt CA, Callaghan MJ, Mason DJ, Conaghan PG. Recommendations for the conduct of efficacy trials of treatment devices for osteoarthritis: a report from a working group of the Arthritis Research UK Osteoarthritis and Crystal Diseases Clinical Studies Group. Rheumatology (Oxford) 2016;55:320–326.

[35] González ÁC, Berenguer SB, Luque Mañas JM, Martin-Pintado-Zugasti A. Validation of a sham novel neural mobilization technique in patients with non-specific low back pain: A randomized, placebo-controlled trial. Musculoskeletal Science and Practice 2021;53:102378.

[36] Grøvle L, Hasvik E, Haugen AJ. Rescue and concomitant analgesics in placebo-controlled trials of pharmacotherapy for neuropathic pain and low back pain. PAIN 2020;161:3–10.

[37] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams JW, Meerpohl J, Norris SL, Akl EA, Schünemann HJ. GRADE guidelines: 5. Rating the quality of evidence—publication bias. Journal of Clinical Epidemiology 2011;64:1277–1282.

[38] Hall S, Lewith G, Brien S, Little P. An exploratory pilot study to design and assess the credibility of a sham kinesiology treatment. Forschende Komplementarmedizin (2006) 2008;15:321–6.

[39] Hart T, Bagiella E. Design and Implementation of Clinical Trials in Rehabilitation Research. Archives of Physical Medicine and Rehabilitation 2012;93:S117–S126.

[40] Häuser W, Bartram-Wunn E, Bartram C, Reinecke H, Tölle T. Systematic review: Placebo response in drug trials of fibromyalgia syndrome and painful peripheral diabetic neuropathy—magnitude and patient-related predictors. PAIN® 2011;152:1709–1717.

[41]   Hohenschurz-Schmidt D, Draper-Rodi J, Vase L, Scott W, McGregor A, Soliman N, MacMillan A, Olivier A, Cherian CA, Corcoran D, Abbey H, Freigang S, Chan J, Phalip J, Sørensen LN, Delafin M, Baptista M, Medforth N, Ruffini N, Andresen SS, Ytier S, Ali D, Hobday H, Santosa AA, Vollert J, Rice AS. Blinding and sham control methods in trials of physical, psychological, and self-management interventions for pain (article I): a systematic review and description of methods. PAIN n.d.;XXXX:XXX–XXX.

[42]   Hohenschurz-Schmidt D, Kleykamp BA, Draper-Rodi J, Vollert J, Chan J, Ferguson M, McNicol E, Phalip J, Evans SR, Turk DC, Dworkin RH, Rice ASC. Pragmatic trials of pain therapies: a systematic review of methods. PAIN 2021. doi:10.1097/j.pain.0000000000002317.

[43]   Howe LC, Goyer JP, Crum AJ. Harnessing the placebo effect: Exploring the influence of physician characteristics on placebo response. Health Psychol 2017;36:1074–1082.

[44]   Howick J, Webster RK, Rees JL, Turner R, Macdonald H, Price A, Evers AWM, Bishop F, Collins GS, Bokelmann K, Hopewell S, Knottnerus A, Lamb S, Madigan C, Napadow V, Papanikitas AN, Hoffmann T. TIDieR-Placebo: A guide and checklist for reporting placebo and sham controls. PLOS Medicine 2020;17:e1003294.

[45]   Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol 2014;43:1272–1283.

[46]   Kaptchuk TJ, Stason WB, Davis RB, Legedza ART, Schnyer RN, Kerr CE, Stone DA, Nam BH, Kirsch I, Goldman RH. Sham device v inert pill: randomised controlled trial of two placebo treatments. BMJ 2006;332:391–397.

[47]   Kawchuk GN, Haugen R, Fritz J. A true blind for subjects who receive spinal manipulation therapy. Archives of Physical Medicine and Rehabilitation 2009;90:366–8.

[48]   Koshi EB, Short CA. Placebo theory and its implications for research and clinical practice: a review of the recent literature. Pain Pract 2007;7:4–20.

[49]   Licciardone JC, Buchanan S, Hensel KL, King HH, Fulda KG, Stoll ST. Osteopathic manipulative treatment of back pain and related symptoms during pregnancy: a randomized controlled trial. American Journal of Obstetrics and Gynecology 2010;202:43.e1–8.

[50]   Locher C, Frey Nascimento A, Kirsch I, Kossowsky J, Meyer A, Gaab J. Is the rationale more important than deception? A randomized controlled trial of open-label placebo analgesia. PAIN 2017;158:2320–2328.

[51]   Macedo A, Farré M, Baños J-E. A meta-analysis of the placebo response in acute migraine and how this response may be influenced by some of the characteristics of clinical trials. Eur J Clin Pharmacol 2006;62:161–172.

[52]   Machado LAC, Kamper SJ, Herbert RD, Maher CG, McAuley JH. Imperfect placebos are common in low back pain trials: a systematic review of the literature. Eur Spine J 2008;17:889–904.

[53]   Magnusson K. Interpreting Cohen's d effect size: An interactive viusualization (Version 2.5.0) [Web App]. R Psychologist 2021. Available: https://rpsychologist.com/cohend/. Accessed 30 Jun 2021.

[54] Meissner K, Fässler M, Rücker G, Kleijnen J, Hróbjartsson A, Schneider A, Antes G, Linde K. Differential Effectiveness of Placebo Treatments: A Systematic Review of Migraine Prophylaxis. JAMA Intern Med 2013;173:1941–1951.

[55] Meissner K, Linde K. Chapter Fourteen - Are Blue Pills Better Than Green? How Treatment Features Modulate Placebo Effects. In: Colloca L, editor. International Review of Neurobiology. Neurobiology of the Placebo Effect Part II. Academic Press, 2018, Vol. 139. pp. 357–378. doi:10.1016/bs.irn.2018.07.014.

[56] Michener LA, Kardouni JR, Sousa CO, Ely JM. Validation of a sham comparator for thoracic spinal manipulation in patients with shoulder pain. Manual Therapy 2015;20:171–5.

[57] Moustgaard H, Clayton GL, Jones HE, Boutron I, Jørgensen L, Laursen DLT, Olsen MF, Paludan-Müller A, Ravaud P, Savović J, Sterne JAC, Higgins JPT, Hróbjartsson A. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ 2020;368. doi:10.1136/bmj.l6802.

[58] Mulroy SJ, Thompson L, Kemp B, Hatchett PP, Newsam CJ, Lupold DG, Haubert LL, Eberly V, Ge T-T, Azen SP, Winstein CJ, Gordon J, Physical Therapy Clinical Research N. Strengthening and optimal movements for painful shoulders (STOMPS) in chronic spinal cord injury: a randomized controlled trial. Physical Therapy 2011;91:305–24.

[59] Namjoo S, Borjali A, Seirafi M, Assarzadegan F. Use of Mindfulness-based Cognitive Therapy to Change Pain-related Cognitive Processing in Patients with Primary Headache: A Randomized Trial with Attention Placebo Control Group. Anesthesiology and Pain Medicine 2019;9:e91927.

[60] Nicholas MK, Asghari A, Blyth FM, Wood BM, Murray R, McCabe R, Brnabic A, Beeston L, Corbett M, Sherrington C, Overton S. Self-management intervention for chronic pain in older adults: A randomised controlled trial. PAIN® 2013;154:824–835.

[61] Ojanen M. Can the true effects of exercise on psychological variables be separated from placebo effects? International Journal of Sport Psychology 1994;25:63–80.

[62] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Journal of Clinical Epidemiology 2021;0. doi:10.1016/j.jclinepi.2021.03.001.

[63] Peerdeman KJ, van Laarhoven AIM, Keij SM, Vase L, Rovers MM, Peters ML, Evers AWM. Relieving patients' pain with expectation interventions: a meta-analysis. PAIN 2016;157:1179–1191.

[64] Puhl AA, Reinhart CJ, Doan JB, Vernon H. The quality of placebos used in randomized, controlled trials of lumbar and pelvic joint thrust manipulation-a systematic review. Spine J 2017;17:445–456.

[65] Roest AM, de Jonge P, Williams CD, de Vries YA, Schoevers RA, Turner EH. Reporting Bias in Clinical Trials Investigating the Efficacy of Second-Generation Antidepressants in the Treatment of Anxiety Disorders: A Report of 2 Meta-analyses. JAMA Psychiatry 2015;72:500–510.

[66] Rosario JL. Relief from Back Pain Through Postural Adjustment: a Controlled Clinical Trial of the Immediate Effects of Muscular Chains Therapy (MCT). International journal of therapeutic massage & bodywork 2014;7:2–6.

[67] do Rosario JLP, Orcesi LS, Kobayashi FN, Aun AN, Diolindo Assumpcao IT, Blasioli GJ, Hanada ES. The immediate effects of modified Yoga positions on musculoskeletal pain relief. Journal of Bodywork and Movement Therapies 2013;17:469–74.

[68] Safer DL, Hugo EM. Designing a Control for a Behavioral Group Therapy. Behavior Therapy 2006;37:120–130.

[69] Sanders AE, Slade GD, Fillingim RB, Ohrbach R, Arbes SJ Jr, Tchivileva IE. Effect of Treatment Expectation on Placebo Response and Analgesic Efficacy: A Secondary Aim in a Randomized Clinical Trial. JAMA Network Open 2020;3:e202907–e202907.

[70] Snodgrass SJ, Rivett DA, Sterling M, Vicenzino B. Dose optimization for spinal treatment effectiveness: a randomized controlled trial investigating the effects of high and low mobilization forces in patients with neck pain. The Journal of orthopaedic and sports physical therapy 2014;44:141–52.

[71] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM, Emberson JR, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.

[72] The Cochrane Collaboration. Review Manager 5 (RevMan 5). The Cochrane Collaboration, 2020 Available: https://training.cochrane.org/online-learning/core-software-cochrane-reviews/revman/revman-5-download. Accessed 6 Jan 2021.

[73] Tough EA, White AR, Richards SH, Lord B, Campbell JL. Developing and Validating a Sham Acupuncture Needle. Acupunct Med 2009;27:118–122.

[74] Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, Cleeland C, Dionne R, Farrar JT, Galer BS, Hewitt DJ, Jadad AR, Katz NP, Kramer LD, Manning DC, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robinson JP, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Witter J. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106:337–345.

[75] Turner JA, Deyo RA, Loeser JD, Von Korff M, Fordyce WE. The importance of placebo effects in pain treatment and research. JAMA 1994;271:1609–1614.

[76] Vase L. Can Insights From Placebo and Nocebo Mechanism Studies Help Improve Randomized Controlled Trials? Clinical Pharmacology & Therapeutics 2019;0. doi:10.1002/cpt.1580.

[77] Vase L, Baram S, Takakura N, Takayama M, Yajima H, Kawase A, Schuster L, Kaptchuk TJ, Schou S, Jensen TS, Zachariae R, Svensson P. Can Acupuncture Treatment Be Double-Blinded? An Evaluation of Double-Blind Acupuncture Treatment of Postoperative Pain. PLOS ONE 2015;10:e0119612.

[78]   Vase L, Petersen GL, Riley JLI, Price DD. Factors contributing to large analgesic effects in placebo mechanism studies conducted between 2002 and 2007. PAIN 2009;145:36–44.

[79]   Vase L, Vollert J, Finnerup NB, Miao X, Atkinson G, Marshall S, Nemeth R, Lange B, Liss C, Price DD, Maier C, Jensen TS, Segerdahl M. Predictors of the placebo analgesia response in randomized controlled trials of chronic pain: a meta-analysis of the individual data from nine industrially sponsored trials. PAIN 2015;156:1795–1802.

[80]   Vernon HT, Triano JJ, Ross JK, Tran SK, Soave DM, Dinulos MD. Validation of a novel sham cervical manipulation procedure. The spine journal : official journal of the North American Spine Society 2012;12:1021–8.

[81]   Walker BF, Hebert JJ, Stomski NJ, Losco B, French SD. Short-term usual chiropractic care for spinal pain: a randomized controlled trial. Spine 2013;38:2071–8.

[82]   von Wernsdorff M, Loef M, Tuschen-Caffier B, Schmidt S. Effects of open-label placebos in clinical trials: a systematic review and meta-analysis. Scientific Reports 2021;11:3855.

[83]   Whalley B, Hyland ME, Kirsch I. Consistency of the placebo effect. Journal of Psychosomatic Research 2008;64:537–541.

[84]   Williams AC de C, Fisher E, Hearn L, Eccleston C. Evidence-based psychological interventions for adults with chronic pain: precision, control, quality, and equipoise. PAIN 2021;160:2149–2153.

[85]   Williams AC de C, Fisher E, Hearn L, Eccleston C. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database of Systematic Reviews 2020. doi:10.1002/14651858.CD007407.pub4.

[86]   Wilson DB. SPSS for Windows Meta-Analysis Macros. 2005 Available: http://mason.gmu.edu/~dwilsonb/ma.html.

[87]   Wolny T, Linek P. Neurodynamic Techniques Versus "Sham" Therapy in the Treatment of Carpal Tunnel Syndrome: A Randomized Placebo-Controlled Trial. Arch Phys Med Rehabil 2018;99:843–854.

[88]   Zemadanis K, Betsos T, Mandalidis D. The short and long-term effect of weight-bearing mobilization-with-movement (MWM) and automobilization-MWM techniques on pain and functional status in patients with hip osteoarthritis. INTERNATIONAL JOURNAL OF PHYSIOTHERAPY 2017;4:160–167.

[89]   Zhang W, Robertson J, Jones AC, Dieppe PA, Doherty M. The placebo effect and its determinants in osteoarthritis: meta-analysis of randomised controlled trials. Annals of the Rheumatic Diseases 2008;67:1716–1723.