Title: Investigating the veracity of a sample of divergent published trial data in spinal pain.

Open Science Framework Registration: https://osf.io/345vq

Neil E O'Connell, Centre for Health and Wellbeing Across the Lifecourse, Department of Health Sciences, Brunel University London, Uxbridge, UK

R Andrew Moore, Court Road, Newton Ferrers, Plymouth, UK

Gavin Stewart, School of Natural and Environmental Sciences, University of Newcastle upon Tyne, Newcastle, UK

Emma Fisher, Centre for Pain Research, University of Bath, Claverton Down, Bath, UK

Leslie Hearn, Cochrane Pain, Palliative and Supportive Care Review Group, Oxford University Hospitals, Oxford, UK

Christopher Eccleston, Centre for Pain Research, University of Bath, Claverton Down, Bath, UK

Amanda C de C Williams, Research Dept of Clinical, Educational & Health Psychology, University College London, London, UK

No. of text pages of the entire manuscript: 16

No. of figures: 8

No. of tables: 5

Corresponding Author: Neil O'Connell, neil.oconnell@brunel.ac.uk, Department of Health Sciences, Brunel University, Uxbridge, UB8 3PH, UK. Telephone 01895 268814

Abstract

Evidence-based medicine is replete with studies assessing quality and bias, but few evaluating research integrity or trustworthiness. A recent Cochrane review of psychological interventions for chronic pain identified trials with a shared lead author with highly divergent results. We sought to systematically identify all similar trials from this author to explore their risk of bias, governance procedures, and trustworthiness.

We searched OVID MEDLINE, EMBASE, CENTRAL and PEDro to 22/12/2021 for trials. We contacted the authors requesting details of trial registration, ethical approval, protocol, and access to the trial data for verification. We used the Cochrane Risk of Bias tool and the Cochrane Pregnancy and Childbirth group's Trustworthiness Screening Tool to guide systematic exploration of trustworthiness.

Ten trials were included: nine compared cognitive behavioural therapy (CBT) and physical exercise to usual care, exercise alone, or physiotherapy, and one compared two brief CBT programmes. Eight trials reported results divergent from the evidence base. Assessment of risk of bias and participant characteristics identified no substantial concerns. Responses from the lead author did not satisfactorily explain this divergence. Trustworthiness screening identified concerns about research governance, data plausibility at baseline, the results, and apparent data duplication.

We discuss the findings within the context of methods for establishing the trustworthiness of research findings generally. Important concerns regarding the trustworthiness of these trials reduce our confidence in them. They should probably not be used to inform the results and conclusions of systematic reviews, in clinical training, policy documents, or any relevant instruction regarding adult chronic pain management.

Introduction

Trust is the foundation on which medicine is built. Patients trust that health professionals have based their practice on the best available evidence, and health professionals trust that researchers have accurately and honestly undertaken and reported their research according to best methods. Evidence-based medicine (EBM) has numerous tools and methods to assess and manage quality and bias in research but few addressing the important question of trust. Accordingly, while EBM is replete with studies assessing quality and bias, there are few examining the integrity or trustworthiness of research.

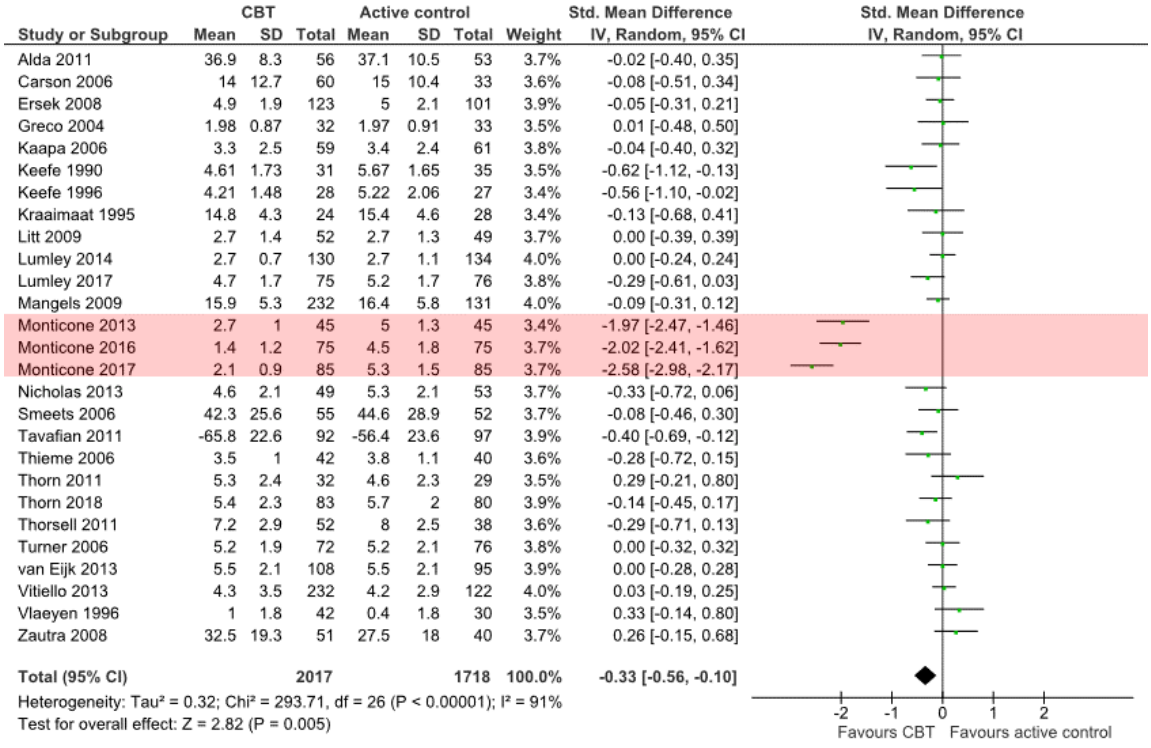
Trustworthiness incorporates research integrity, governance, and potential research misconduct. The latter might include fabrication or falsification of research results, or plagiarism [3], but, importantly, does not include error. We are aware of no consensus on the characteristics of studies that act as possible warning signs for untrustworthiness. Although a variety of methods have been used to assess research misconduct, there are few validated methods beyond the approaches that identify textual plagiarism [3]. Any single method is likely to be insufficient and investigators are recommended to use multiple methods considering aspects of research governance (pre-registration, ethical approval and quality of reporting), with close scrutiny of reported data and sight of the raw data, if deemed necessary.

Several tools have been proposed to formally explore the integrity and trustworthiness of research. The REAPPRAISED checklist [9], for example, was developed to identify possible problems with research integrity and includes items relating to research governance, ethics, authorship, productivity, plagiarism, research conduct, analyses and methods, possible image manipulation,

statistical considerations, errors, and data duplication. The Cochrane Pregnancy and Childbirth review group developed a Trustworthiness Screening Tool (CPC-TST) [6] specifically for clinical trials. This tool is applied to all trials eligible for inclusion in systematic reviews published by their group, exploring scientific integrity and trustworthiness with items relating to aspects of research governance, participant characteristics, feasibility, and study results.

Our research began with an attempt to determine the veracity and completeness of the Cochrane library entry on the effectiveness and safety of psychological interventions for the treatment of chronic pain in adults [34], conducted by some of the authors of this paper (EF, LH, CE, AW). The review included three trials [16,19,21], with a common lead author, whose results diverged substantially from the rest of the field, both at post-treatment and follow-up, with effect sizes and confidence intervals that did not overlap with outcome data from 24 other included trials (see Figure 1).

Figure 1: Forest plot for the analysis "CBT vs active control" from Williams et al. [33] with the divergent trials highlighted.



After assessment of error, and of uniqueness of treatment characteristics, the lead author (Dr M Monticone) was contacted for insight. Details of that correspondence can be found in Supplementary Information. The review author team concluded that these trials were unreliable and excluded them from their primary analyses, including them only in sensitivity analyses. Other similar publications from Dr Monticone's research group became the focus of this study. We determined to recover all recent published study reports, assessing their quality and bias, their governance procedures such as registration and ethical review, and finally their trustworthiness.

Methods

We registered a protocol for this review on the Open Science Framework: https://osf.io/345vq

Searches

We conducted a search of the CENTRAL, PEDro, OVID MEDLINE and EMBASE databases from 2010 to 22/12/21 for all randomised clinical trials (RCTs) published since 2010 in subacute or persistent spinal pain in which Dr Monticone was lead author. We excluded non-randomised studies and studies that did not investigate the effectiveness of an intervention for subacute or persistent spinal pain. Three reviewers (NOC, AW, LH) independently screened the searches and any disagreements were resolved through discussion.

Procedures to explore veracity

We used the Cochrane Risk of Bias (RoB) tool [11] to evaluate the risk of bias for each included study. Two reviewers (NOC, EF) independently applied the screening tool to the included trials, with any disagreements resolved through discussion.

We used the following key domains and items of the Cochrane Pregnancy and Childbirth review group's TST (CPC-TST) [6] to guide our exploration of the included papers (Table 1). We chose this tool as it was specifically developed to evaluate RCTs. Three reviewers (NOC, AW, LH) independently applied the screening tool to each included study, with any disagreements resolved through discussion.

Table 1. Key domains from the CPC-TST used to explore the sample of studies

Research Governance
Are there any retraction notices or expressions of concern listed on the Retraction Watch Database relating to this study?
Was the study prospectively registered (for those studies published after 2010)? If not, have the authors provided a plausible reason?
When requested, did the trial authors provide/share the protocol and/or ethics approval letter?
Did the trial authors engage in communication with the Cochrane Review authors within the agreed timelines?
Did the trial authors provide Individual Patient Data (IPD) upon request? If not, was there a plausible reason?
Baseline characteristics
Is the study free from characteristics of the study participants that appear too similar?
Feasibility
Is the study free from characteristics that could be implausible?
In cases with (close to) zero losses to follow-up, is there a plausible explanation?
Results
Is the study free from results that could be implausible?
Do the numbers randomised to each group suggest that adequate randomisation methods were used?

Research Governance

We reviewed the included trials for details of registration. Where details were provided, we checked the trial registry record to ascertain whether the trial was prospectively or retrospectively registered. We contacted the lead author of the included trials to request details of registration for all included trials, evidence of local ethical approval, full details of the intervention content and delivery, and full individual patient datasets for all included trials. We extracted any information regarding ethical approval processes from study reports. We searched Retraction Watch (https://retractionwatch.com/) for any retraction notices related to the included trials.

Baseline Characteristics

To explore the similarity of baseline data, we extracted these for treatment and control groups (means and standard deviations (SDs)) for all reported continuous variables and calculated p values with unpaired *t*-tests. For categorical variables, we used Fisher's exact test. We conducted separate tests for variables where there were multiple independent levels for which participants may have events in more than one level (e.g. use of different medications), and single multi-level tests (e.g. Fisher's exact test 2x3 or 2x4, χ^2 test where n > 120) where there were multiple levels but participants could only be represented at a single level (e.g. highest level of education or employment status). For each trial, we plotted the distribution of p values and calculated the pooled p value using Stouffer's z-score method [32]. This method calculates a p value by summing the z-scores corresponding to each variable and dividing them by the square root of the number of variables [5]. The pooled p value represents the combined probability across multiple independent comparisons of observing a difference between groups as large as that observed where the null hypothesis is true.

We also used Stouffer's method to calculate a pooled p value using all p values of all included trials and plotted the distribution of all p values from all the included trials combined. In the case of simple randomisation, we might expect baseline p values to display a uniform distribution between 0 and 1. Combined p values close to 1.0 would indicate more similar baseline mean values and close to 0 would indicate more dissimilar means. We classified pooled p values of ≤ 5% from 0 or 1.0 as likely to be inconsistent with random allocation [3]. Distributions were plotted in Jamovi [33] and Microsoft Excel 2019.

Feasibility

To explore the feasibility of participant characteristics, we compared baseline data for pain intensity, disability and health-related quality of life (HRQoL) with published normative data from a clinical population of > 6000 people with persistent back and neck pain seen in a Pain Management and Research Centre [27], to identify unexplained divergence. We extracted and explored the amount of participant attrition for all groups in each study.

Plausibility of results

To enable combination and comparison of effect sizes, we calculated the standardised mean difference (SMD) (Hedge's g) using Revman 5.4 [30] for the outcome measures of pain and disability for all time points in all trials. We calculated pooled effect sizes for immediate, 3 months, 12 months and 24 months post-intervention time-points, using a random effects model. For the mean difference (MD), all pain scales were normalised to a 0-10 scale. We plotted the combined

distribution of both SMDs and p values for pain, disability and HRQoL outcomes for all subscales of the tools used in the trials (the 36-item short form survey (SF-36) or the Scoliosis Research Society-22 patient questionnaire (SRS-22)) for all trials.

We explored the plausibility of these results in several ways: by comparing pain and disability effect sizes with those of the other included trials in the Cochrane review of psychological therapies for persistent pain (Williams et al.) [34]; by examining the level of statistical significance in the results of the included trials; by comparing pain effect sizes for all included trials with other interventions for chronic pain; and by examining conversions of MD and SMD to number needed to treat (NNT) and comparing indicated NNTs with other interventions for chronic pain .

We formally examined the baseline and outcomes data across all included trials for the presence of duplicate or similar data, using an approach modified from that of Bordewijk et al. [2]. Identical data between trials were counted where the means and SDs matched for the same outcome. Similar data were counted where values for the same outcome differed by less than 1.

We explored potential concerns with the randomisation process by reviewing the description of the randomisation method and by scrutinising the number of participants allocated to each group. Identical numbers allocated to each group in the absence of a block approach to randomisation was considered as cause for concern. We extracted and examined the data relating to participant attrition. The observation of zero or nearly zero loss to follow-up, particularly in the longer term, was considered as cause for concern. In scrutinising data from the published records of included trials, we aimed to identify any further errors or apparent inconsistencies.

Results

Supplementary Figure 1 shows a flow diagram of the search process. Our searches identified 10 RCTs of interventions for subacute or persistent spinal pain, randomising 1100 participants [15-24]. Recruitment took place between December 2007 and December 2015, and the trials were published between 2012 and 2021. Trial sizes ranged from 20 to 170 participants randomised (mean (SD) 110 (53)); Table 2 provides a summary of study characteristics.

Table 2. Characteristics of included trials.

Study ID	Journal of publication	Setting	Participants	N randomised	Details of experimental intervention	Details of control intervention
Monticone 2012 [15]	Eur Spine J	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Chronic non-specific neck pain	80	CBT Physiotherapy including exercise (posture, strength, stretching), ergonomic advice, manual therapy (≤ 12 sessions; x1-2 weekly)	Physiotherapy including exercise (posture, strength, stretching), ergonomic advice, manual therapy (≤ 12 sessions; x1-2 weekly)

Monticone 2013 [16]	Clin J Pain	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Chronic non-specific low back pain	90	CBT (x1 weekly; 5 weeks; then x1 monthly for 1 year) Exercise (posture, strength, stretching), ergonomic advice, manual therapy (10 sessions; x2 weekly; telephone reminders to exercise for 1 year)	Exercise (posture, strength, stretching), ergonomic advice, manual therapy (10 sessions; x2 weekly; telephone reminders to exercise for 1 year)
Monticone 2014a [17]	Eur Spine J	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Chronic non-specific low back pain	20	CBT (x1 weekly; 8 weeks) Exercise (motor control focused) (x2 weekly; 8 weeks)	Exercise (posture, strength, stretching), manual therapy. (x2 weekly; 8 weeks)
Monticone 2014b [18]	Eur Spine J	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Lumbar fusion for degenerative or isthmic spondylolisthesis	130	CBT (x2 weekly; 4 weeks) Exercise (posture, strength, stretching, walking), ergonomic advice (x5 weekly; 4 weeks)	Exercise (posture, strength, stretching, walking), ergonomic advice (x5 weekly; 4 weeks)
Monticone 2016a [19]	Eur Spine J	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Chronic non-specific low back pain	150	CBT (x1 weekly; 5 weeks) Task-based exercise (motor control training, task-oriented exercises, coordination/balance exercises) (x2 weekly; 5 weeks)	Exercise (posture, strength, stretching, walking), ergonomic advice (x2 weekly; 5 weeks)
Monticone 2016b [20]	Eur Spine J	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Adult idiopathic scoliosis	130	Active self-correction and scoliosis alignment exercises with cognitive-behavioural strategies and ergonomic advice (x1 weekly; 20 weeks)	Physiotherapy including exercise (postural, strength, stretching), manual therapy (x1 weekly; 20 weeks)

Monticone 2017 [21]	Clin Rehabil	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Chronic non-specific neck pain	170	CBT (x1 weekly; 10 weeks) Exercises (graded exposure, mobility, postural, strength, stretching) (x1 weekly; 10 weeks)	Physiotherapy exercises (strength, stretching, mobilisation) (x1 weekly; 10 weeks)
Monticone 2018 [22]	Eur J Phys Rehabil Med	Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone	Chronic non-specific neck pain	30	NeckPix©* CBT (x4 weekly; 1 week) Exercise (mobility, strength, stretching, motor control, task-oriented) (x2 weekly; 5 weeks)	CBT (x4 weekly; 1 week). Exercise (mobility, strength, stretching, posture) (x2 weekly; 5 weeks)
Monticone 2020 [23]	Disabil Rehabil	Unclear	Failed back surgery syndrome	150	CBT (x1 weekly; 10 weeks) Exercise (mobility, motor control, task-oriented, stretching, balance, proprioception), ergonomic advice (x2 weekly; 10 weeks)	Physiotherapy: exercise (mobility, stretching, strength, posture), manual therapy ergonomic advice (x2 weekly; 10 weeks).
Monticone 2021 [24]	Eur J Phys Rehabil Med	Unclear	Subacute low back pain	150	CBT (x1 weekly; 10 weeks) Exercise (mobility, motor control, task-oriented, postural, proprioception) (x2 weekly; 10 weeks)	Physiotherapy: Exercise (strength, stretching, postural), manual therapy (x2 weekly;10 weeks)

*Footnotes: * "a multi-image instrument developed to assess daily activities in the context of pain-related fear"*

Nine included trials compared a form of CBT and physical exercise to either usual care, exercise alone, or physiotherapy, and one trial [22] compared two different brief CBT programmes Eight trials were conducted in the same clinical centre in Lissone, Italy, while two trials did not specify the setting. There was some variation across trials in the description of the specific populations studied, with conditions including chronic low back pain [16,17,19], subacute low back pain [24], chronic

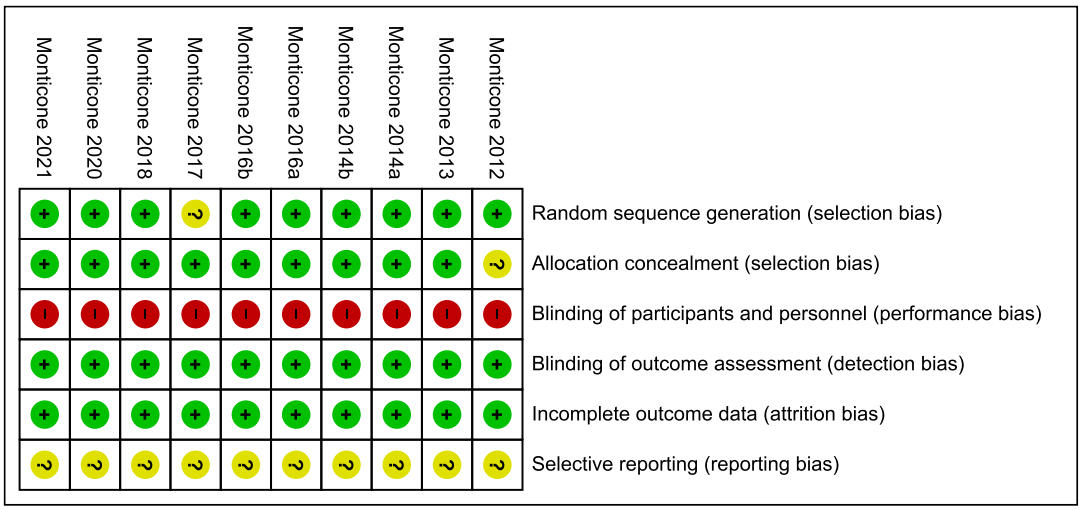
neck pain [15,21,22], persistent pain in adults with idiopathic scoliosis [20], failed back surgery syndrome [23] and pain after lumbar fusion surgery [18].

E-mail correspondence between members of the review [34] team and the lead author and two co-authors of these trials elicited data as requested that allowed their inclusion in the meta-analyses. However, the explanations of treatment content and process revealed nothing unusual, apart from monthly telephone reminders to participating patients which were offered as an explanation for zero attrition at two-year follow-up. At the time of that review, the team enquired further but received no information.

Risk of bias

The included trials were mostly rated as low RoB for randomisation, allocation concealment, blinding of outcome assessors and attrition, with one trial judged as unclear for randomisation and one for allocation concealment (due to lack of detail in the reported methods). All trials were rated as unclear RoB for selective outcome reporting due to the lack of pre-registration or available protocols. All trials were rated as high RoB for blinding of participants and personnel as it was not possible to blind clinicians or participants for these interventions and comparisons. This is the norm for most trials of psychological interventions. Figure 2 presents the RoB judgements on each domain for each study.

Figure 2. Risk of bias summary: review authors' judgements about each RoB item for each included study.



Trustworthiness screening tool.

Table 3 illustrates the results of the CPC-TST tool assessment. There follows a more detailed description of the findings of that screening process.

Table 3. Results of the CPC-TST assessment

DOMAIN	2012	2013	2014a	2014b	2016a	2016b	2017	2018	2020	2021
Retraction notices?										
Prospectively registered?										
Did authors engage with requests for information?										
Protocols or ethics approval shared on request										
IPD shared on request?										
Baseline similarity (continuous data only)										
Baseline similarity all variables										
Participant characteristics (feasibility)										
Attrition feasibility										
Randomisation concerns										
Results plausibility										

Footnotes: Red = Some concerns; Green = No concerns

Research governance

We contacted the ethics committee of the Instituti Clinici Scientifici Maugeri on 7 January 2022 to request confirmation and evidence that ethical approval was sought and granted for the studies but, at the time of writing (5 April 2022), had not received a response. We contacted the lead author of the trials by email on 1 December 2021, with a reminder sent on 6 January 2022. For the 10 included trials, we asked: whether a clinical trial protocol was developed for the trials and for a copy of any such protocols; for information relating to trial registration or an explanation for non-registration; whether ethical approval was obtained for the included trials and for evidence of such; for access to IPD for each trial; and for an explanation of observed anomalies regarding randomisation, specific apparent errors in baseline p values and instances of duplicate and highly similar data between trials. We received an email response from Dr Monticone on 12 January 2022. Supplementary information has the full details of our enquiries and of Dr Monticone's responses. We also contacted co-author Dr Barbara Rocca on 14 January 2022 requesting this information but at the time of writing (5 April 2022) had not received a response.

None of the identified trials was pre-registered, though three [21,23,24] reported a trial registration number. These latter were registered retrospectively between 2 and 5 years after recruitment was reported to have ended. One trial was registered after the manuscript [21] had been submitted for

publication (ISRCTN14581536), while the other two were registered 12 and 16 months before submission. We identified no retraction notices for any of the included trials.

In his response, Dr Monticone confirmed that none of the trials had been pre-registered. The reasons given for this were that either they started before this issue was strictly required by journals or because the journals had not required it. In three cases [21,23,24], trials were retrospectively registered at the recommendation of the relevant journals.

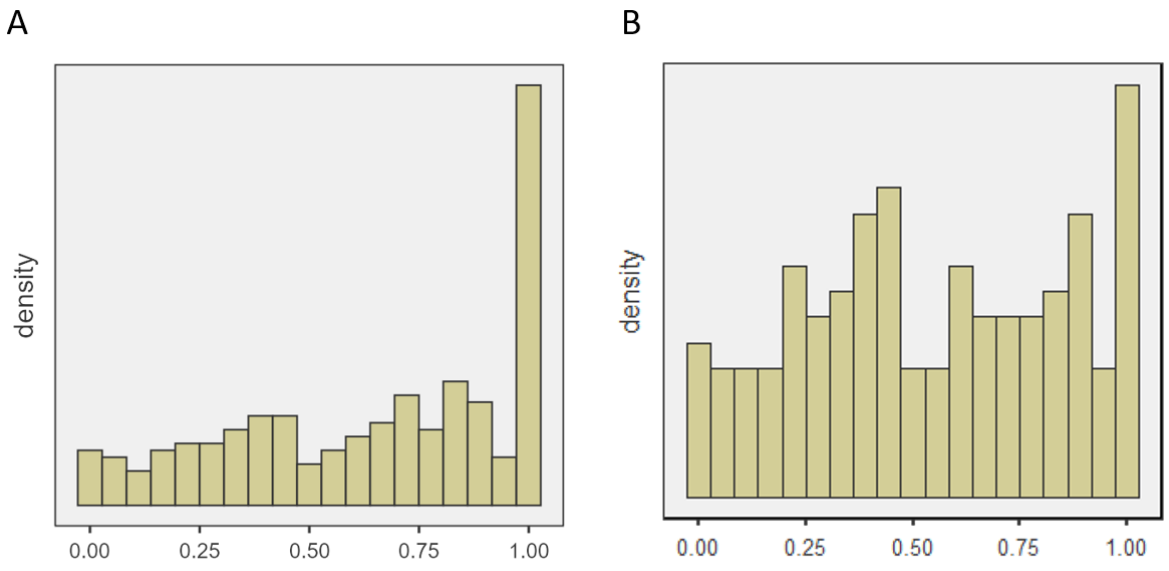
All trial reports included a statement that they had been approved by the Hospital's Institutional Review Board. Six stated that the trial was conducted in conformity with ethical and humane principles of research and one stated that the study was conducted in accordance with the principles of the Helsinki Declaration. In his response, Dr Monticone stated "There was the approval of our Institutional Review Board at the Hospital where the studies were performed. I would prefer to avoid sending these documents." Dr Monticone also responded that there were no trial protocols, giving as a reason that "*the intervention groups always belonged to our clinical practice*". It is therefore unclear what information was submitted to the ethics review board if there were no trial protocols. Dr Monticone also stated that he "*would prefer to avoid sending databases*" which we understood as a decision not to share full IPD sets for the included trials.

Baseline Characteristics

Reviewing the distribution of baseline p values for all variables across all included trials revealed a non-uniform distribution (see Figure 3). The median p value was 0.713 (interquartile range (IQR) 0.377 to 0.943). Stouffer's method revealed pooled baseline p values for each individual trial as within 5% of 0 or 1 for 9 of the 10 included trials.

The distribution of p values for each individual trial is presented in the Supplementary information (Supplementary Figure 2). Table 4 presents pooled p values for baseline comparisons for each trial.

Figure 3. Distribution of p values across all trials. A: for all baseline variables across all included trials. B: for continuous outcomes only



The baseline characteristics tables included several categorical variables with low numbers of events. We considered that this might skew our analysis and so conducted a sensitivity analysis including only continuous outcomes. In that analysis, the median p value was 0.623 (IQR 0.384 to 0.848). Stouffer's method resulted in pooled baseline p values for each trial as within 5% of 0 or 1 for 5 of the 10 included trials which we judged as likely to be inconsistent with random allocation. The distribution of p values remained non-uniform.

Table 4: Pooled p values for baseline comparisons for each included study.

	Study ID									
	Monticone 2012	Monticone 2013	Monticone 2014a	Monticone 2014b	Monticone 2016a	Monticone 2016b	Monticone 2017	Monticone 2018	Monticone 2020	Monticone 2021
All variables (continuous and categorical)	0.0002	1.0000	1.0000	0.99787	0.9342	1.0000	1.0000	1.0000	0.9933	0.9972
Continuous variables only	0.0000	0.9999	1.0000	0.8538	0.0834	1.0000	0.9223	0.6461	0.8161	0.9917

Feasibility of participant characteristics

When comparing the baseline characteristics of participants in the trials of spinal pain with published norms [27], we observed that baseline pain intensity was frequently higher than norms, despite no study reporting a minimum threshold for pain intensity in their inclusion criteria. The median reported baseline intensity was 6/10 (range 4.8-7.0), compared to published norms of mean (SD) 4.1 (1.2) for persistent neck pain and 4.2 (1.0) for persistent low back pain. In some included trials, average baseline scores for HRQoL on the SF-36 subdomains of Role Function, Social Function and Vitality [14-19] and Physical Function [15,16,19] were notably higher than those observed in the published norms (Supplementary Table 1). However, we judged that these observations were not sufficiently remarkable to warrant a positive risk judgement on the CPC-TST for any trial.

Randomisation concerns

Treatment groups were of equal size after allocation in all studies. Four of the 10 trials [18,21,23,24] reported a block method for randomisation that might increase the chances of equal numbers emerging in treatment groups. Of these, one trial reported using random permuted blocks and random block length. The other three trials reported using a "*permuted block randomisation process*" but did not add further detail. In his response, Dr Monticone reported that "*the number of patients randomised was generated by chance based on the patients that were excluded*". We

judged that it was unlikely that equal group numbers in all 10 trials would result from a random process of allocation.

Plausibility of results

Effect sizes for all outcomes were large or extremely large in 8 of the 10 trials. All 8 of these trials compared a form of CBT and physical exercise with either usual care, exercise alone or physiotherapy. These large effect sizes were seen at both short- and long-term follow-up, with larger median effect sizes observed at long-term follow-up.

Figure 4 summarises the effect sizes for pain and disability for all trials. In addition, we present the distribution of effect sizes for pain, disability and all HRQoL subscales at both the short-term (immediate and 3 months post-intervention) and long-term (1 and 2 years post-intervention) follow-up time points. For this purpose, all were converted to positive values.

Figure 4. A summary of effect sizes and the distribution of effect sizes across all trials and follow-up points. *Effect sizes for A. Pain intensity; B. Disability; C/D. The distribution of effect sizes from the outcomes pain, disability, HRQoL subscales, combined at post-intervention (n = 81) and long-term follow-up (n = 80).*

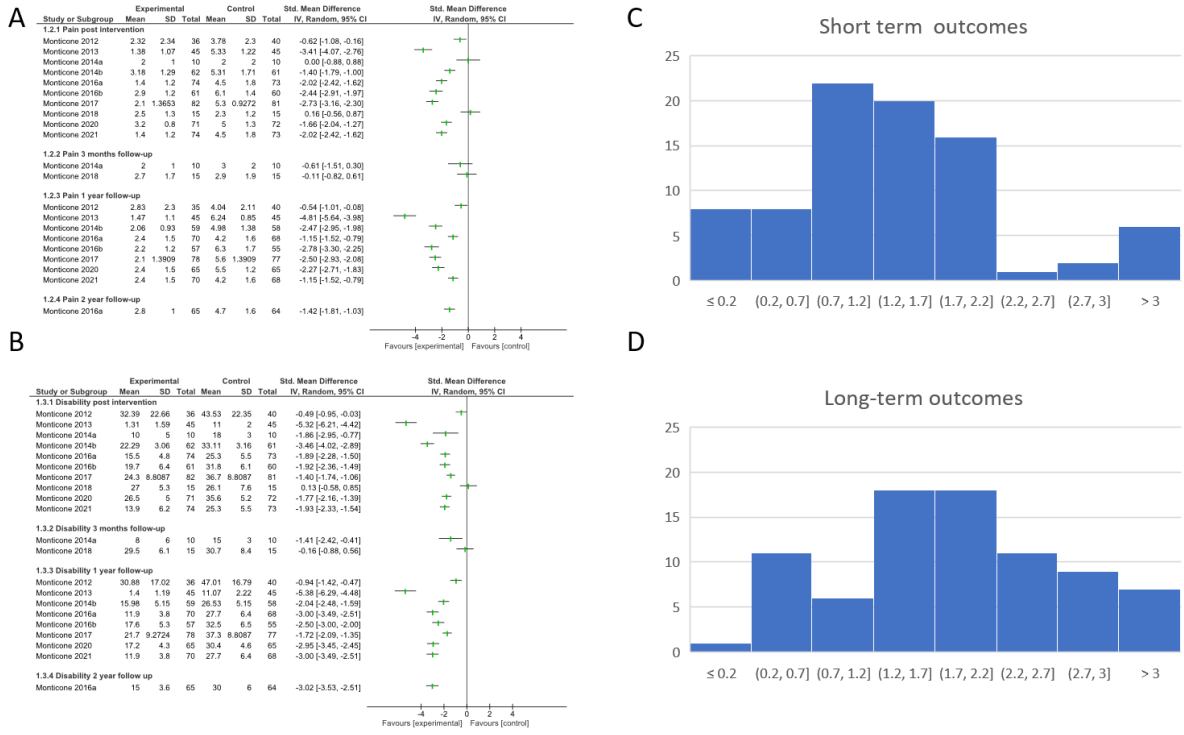


Table 5 presents the pooled effect sizes across the included trials for pain and disability for short- and long-term follow-up. In this sample of trials, the pooled effect size for pain intensity was SMD -1.65 (95% confidence interval (CI) -2.21, -1.09) at end of treatment and -2.17 (95% CI -2.89, -1.45) at long-term follow-up. This represents a six-fold difference between the lower confidence interval of the Monticone studies and the upper confidence interval of all others combined. Supplementary Figure 3 shows SMD values for pain and disability in Williams 2020 [34] excluding the three

previously included trials, and the SMD values for pain and disability from the 10 trials included in this analysis. There is little overlap.

On a 0-10 pain numerical rating scale (NRS), this equates to a pooled effect of -2.29 (95% CI -2.94, -1.65) at end of treatment and -2.93 (95% CI -3.73, -2.14) at one year follow-up. In comparison, a systematic review [29] comparing combined physical and psychological rehabilitation with physical rehabilitation alone reported a mean difference in pain intensity of -0.52 (95% CI 0.16-0.88) at short-term and -0.47 (95% CI 0.13, 0.81) at long-term follow-up.

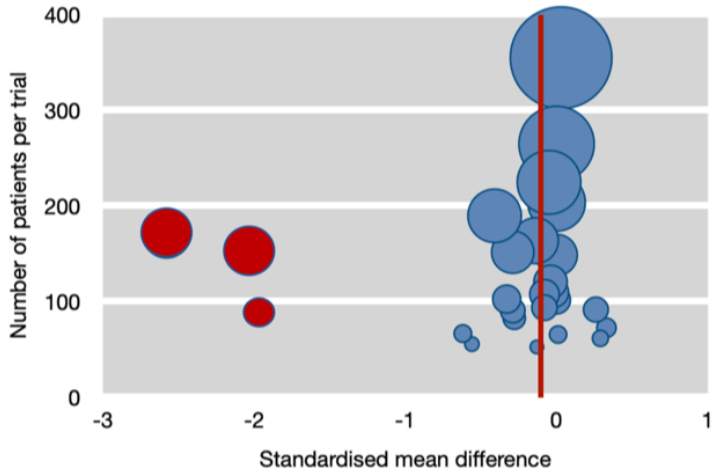
Table 5: Comparison of effect sizes (all SMD with 95% confidence interval using random effects) for pain intensity and disability after treatment and at 1 year follow-up

	At end of treatment period			At 1 year follow-up		
Outcome	**Number of studies**	**Results from studies currently examined**	**Results from the Cochrane review**	**Number of studies**	**Results from studies currently examined**	**Results from the Cochrane review**
Pain intensity	10	-1.65 (-2.21 to –1.09)	-0.09 (-0.17 to -0.01)	8	-2.17 (-2.89 to -1.45)	-0.08 (-0.19 to 0.04)
Disability	10	-1.96 (-2.60 to -1.32)	0.12 (-0.20 to -0.04)	8	-2.64 (-3.32 to -1.95)	-0.12 (-0.26 to 0.02)

Comparisons within Williams 2020

To place these results in context, in the most recent Cochrane systematic review of psychological interventions for persistent pain [34], the pooled effect size (SMD) for the comparison 'CBT vs active care', derived from 23 RCTs with 3235 participants, was -0.09 (95% CI -0.17, -0.01) for pain at the end of treatment. Figure 5 demonstrates the magnitude of the difference using the SMDs calculated for each of the included studies in the Williams et al. primary analysis [34], and the three Monticone trials [16,19,21] excluded from the primary analysis.

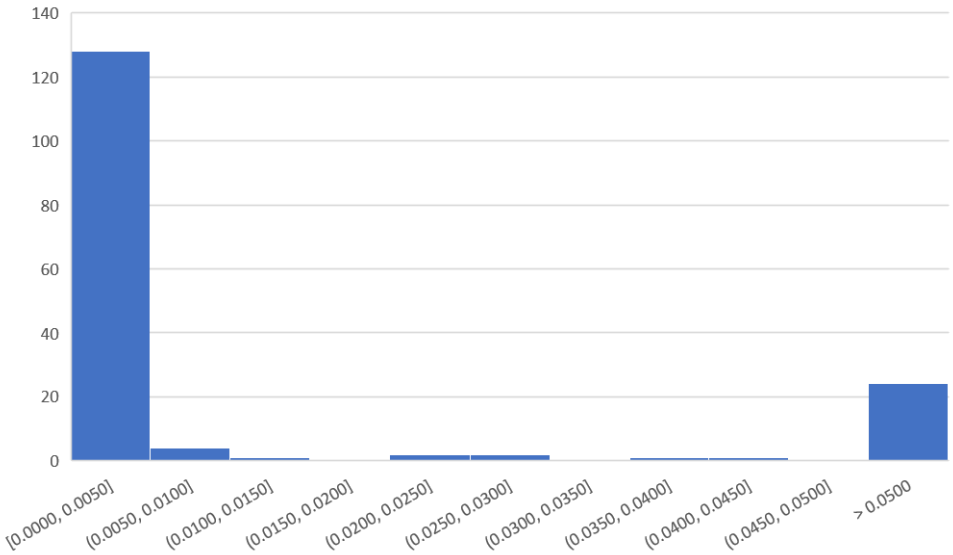
Figure 5: Individual study pain reduction SMDs plotted against the total number of patients in trial (from Williams et al.) [33]. Symbol diameter is proportional to the total number of participants. Red filled circles represent studies by Monticone et al. Blue filled circles represent all other trials in that analysis.



Examination of statistical significance

Figure 6 displays the distribution of p values for all pain intensity, disability, and HRQoL subscales from all post-intervention and long-term follow-up comparisons, from all 10 included trials (n = 163 comparisons). Most had p values of < 0.001.

Figure 6. Distribution of p values for pain, disability, and HRQoL comparisons from included spinal pain trials (from short-term and long-term follow-up): n=163 comparisons.



The reported effect estimates in the 10 included Monticone trials in this analysis are both extreme in size and precise, as reflected by the extremely high rate of reported p values of < 0.001. The extent of the divergence is stark, illustrated by the six-fold difference between the lower confidence interval of the Monticone studies and the upper confidence interval of all others, and the more than 18-fold difference between point estimates of the SMD. Reflecting the MD in pain intensity as a proportion of baseline levels, results show median reductions in pain intensity of 40% (IQR 28-52) in the short term (ST) and 44% (IQR 28-53) in the long term (LT), attributable to the interventions. One trial [17] found no evidence for an effect and one trial [15] found medium size effects on pain and disability at short-term follow-up and a medium size effect on pain and a large effect on disability at long-term follow-up. Seven of the 8 included trials with long term (≥ 1 year) follow-up reported an average long-term mean difference for pain intensity greater than 1.5/10 (range 0-4.1). A similar pattern was also found for disability.

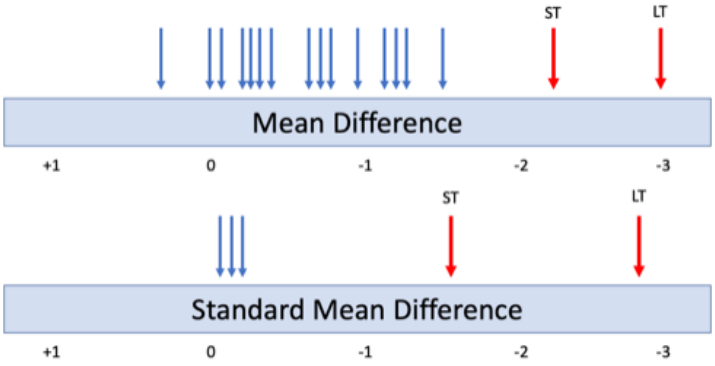
Dr Monticone responded that he would prefer not to provide us with access to full IPD. He stated that the large effect sizes "*were due to the specific characteristics of the CBT group. Indeed, relevant efforts were made in order to strongly improve patients 'health conditions. This is also demonstrated by the between-group clinical differences achieved as well as by the level of satisfaction found in CBT groups*." No further explanation was offered.

Comparing effect sizes from included studies with those of other interventions for chronic pain, using SMD and MD

Supplementary Table 2 shows 10 systematic reviews with 14 interventions, mainly reporting MDs using a 0-10 pain measure. Interventions include NSAIDs, antidepressants, anticonvulsants, cannabinoids, opioids, psychological therapies, acupuncture, magnetic stimulation, and therapeutic ultrasound. Pain conditions include chronic pain, low back pain, osteoarthritis and rheumatoid arthritis, fibromyalgia, and neuropathic pain. Most reported outcomes were compared with placebo at around three months, but some after shorter times.

Results are shown graphically in Figure 7. The pooled effect size for the included studies is considerably greater than for any other intervention. Few interventions have a greater average effect size than that often considered a clinically important difference, approximately equivalent to a mean difference of -1.

Figure 7: MD and SMD for a range of interventions for chronic pain (blue), and the pooled analysis of all Monticone chronic pain studies (red) post-intervention (ST) and at follow-up (LT). The more negative, the larger the effect size.



Comparing effect sizes from included studies with those of other interventions for chronic pain, using NNT

As stated above, the included trials show median reductions in pain intensity of 40% (ST) and 44% (LT) (Supplementary Figure 4). For placebo, average initial pain intensity of about 5.9/10 fell to 4.7 post-intervention and 4.6 at long-term follow-up. For patients receiving experimental treatment, the values were 5.9, 2.4, and 2.2, respectively; this average LT reduction by 63% implies than most patients would experience pain reduction of more than 50%.

Using the method proposed by Faraone [8], we converted SMD to number needed to treat (NNT). When the SMD = 1, the NNT = 2, but as the SMD approaches zero the NNT becomes very large, so an SMD of 0.5 becomes an NNT of about 5, and an SMD of 0.25 is equivalent to an NNT of above 15. The average SMD for the Monticone studies produces a NNT **below** 2.

A linear relationship between MD and NNT can be shown up to a MD of about 1.5 for NSAIDs in OA, equivalent to an NNT of about 4.5 [26] (see Supplementary Figure 5). It is not possible to predict

accurately the shape of the curve beyond that, but the average MD of 2.7 for all 10 Monticone trials could plausibly imply NNTs of 2 or below. This level of effectiveness is highly unusual. There are almost no examples of NNT values for chronic pain interventions of any sort that are below 3 or even 4.

Data duplication/ similarity

There was no consistent evidence of large-scale data duplication across the included trials (see Supplementary Table 3). However, there were specific examples of identical or similar data, and Figure 8 shows tables from three publications with highlighted key examples of identical, or highly similar, data.

For baseline variables, most cases of similar or identical numbers arose from categorical variables with very low numbers (for example the number of participants taking specific types of medication) which might reasonably be expected to occur through chance. However, there were striking cases of similarity between outcome data in a trial (n = 150) published in 2016 in patients with chronic low back pain [19] and two trials published in 2020 and 2021 (both n = 150) in patients with failed back surgery syndrome and subacute low back pain, respectively [23,24]. In his response, Dr Monticone reported "*I checked the tables and you are right as for the NRS, the ODI and the TSK. I was surprised but these are the data the staff collected. I think that values on catastrophizing differ, because I used another scale in my 2021 paper (the CSQ-R)*."

Figure 8. Examples of identical and highly similar data in the results of three included trials (table excerpts copied with permission). *Footnotes: Red shading = identical data; orange shading = highly similar data*

Monticone 2016a

Table 2 Changes over time within and between the control and experimental group (n = 150)

	Group	Pre-training*	Post-training*	12M Follow-up*	24M Follow-up*	
Primary outcome						
ODI (0–100)	Experimental	34.4 (4.6)	15.5 (4.8)	11.9 (3.8)	15.0 (3.6)	
	Control	32.4 (5.4)	25.3 (5.5)	27.7 (6.4)	30.0 (6.0)	
Secondary outcomes						
TSK (13–52)	Experimental	27.5 (4.7)	17.6 (5.3)	15.5 (4.8)	14.4 (4.1)	
	Control	28.5 (5.5)	26.8 (5.9)	29.6 (5.3)	29.9 (4.3)	
PCS (0–52)	Experimental	27.8 (4.6)	12.7 (5.3)	10.0 (5.1)	11.4 (3.7)	
	Control	26.4 (5.4)	25.1 (5.0)	26.7 (4.9)	27.1 (4.7)	
NRS (0–10)	Experimental	6.4 (1.7)	1.4 (1.2)	2.4 (1.5)	2.8 (1.0)	
	Control	6.1 (1.6)	4.5 (1.8)	4.2 (1.6)	4.7 (1.6)	

ODI, Oswestry Disability Index; TSK, Tampa Scale of Kinesiophobia; PCS, Pain Catastrophizing S

Monticone 2020

Table 2. Changes over time within and between the experimental and the control group in terms

	Group	Pre-training*	Post-training*	12M Follow-up*	
Primary outcome					
ODI (0–100) ↓	Exp	51.4 (10)	26.5 (5)	17.2 (4.3)	
	Control	48.9 (10.3)	35.6 (5.2)	30.4 (4.6)	
Secondary outcomes					
TSK (13–52) ↓	Exp	36.5 (4.7)	17.7 (5.4)	15.3 (4.6)	
	Control	35 (4.6)	26.7 (5.9)	29.5 (5.3)	
PCS (0–52) ↓	Exp	34.6 (3.9)	12.6 (5.3)	10.8 (4.5)	
	Control	32.8 (4.2)	25.1 (5)	26.6 (4.9)	
NRS (0–10) ↓	Exp	6.6 (1.5)	3.2 (0.8)	2.4 (1.5)	
	Control	6.5 (1.4)	5 (1.3)	5.5 (1.2)	

*mean value (standard deviation).
†Between-group change (95% confidential interval).
ODI: Oswestry Disability Index; TSK: Tampa Scale of Kinesiophobia; PCS: Pain Catastrophising Scale;

Monticone2021

	Group	Pre-training[a]	Post-training[a]	12M Follow-up[a]	Change at post-training[b]	
Primary outcome						
ODI (0-100) ↓	Experimental	23.9 (12.1)	13.9 (6.2)	11.9 (3.8)	11.5 (1.0)	
	Control	23.7 (13.6)	25.3 (5.5)	27.7 (6.4)		
Secondary outcomes						
NRS (0-10) ↓	Experimental	5.5 (2.2)	1.4 (1.2)	2.4 (1.5)	3.1 (0.3)	
	Control	4.8 (2.5)	4.5 (1.8)	4.2 (1.6)		
TSK						

Data anomalies/ errors

Beyond the apparent duplication of data, we identified examples of anomalous or erroneous data. Specifically, there were two instances of reported baseline p values that did not match the

presented means/SDs [17,20]. These were the variables step length, step time, single support time (left and right) in Monticone 2014a [17], and all SRS-22 subscales for Monticone 2016b [20]. In these instances, baseline mean values/SDs between the treatment groups were identical to or differed by a maximum of one decimal place unit but the presented p value ranged from 0.161 to 0.884. In his response, Dr Monticone stated that he had "*checked again the data and they are OK as presented*".

Attrition

Levels of attrition varied across the included studies (median (range) 9% (0-17)) at the end of follow-up, but three trials [16,17,22] reported no attrition at any follow-up point. Of these, two were small trials [17,22] but one [16] randomised 90 participants, with a 12-month intervention followed up for 24 months. A further three trials reported < 10% attrition at 1-year follow-up [15,21,24]. We considered the trial with zero attrition at 24-month follow-up [16] to be at high risk on this item of the CPC-TST. While we did not rate the other trials at high risk on this item, it should be noted that < 10% attrition at 1-year follow-up might be considered unusual.

Discussion

We wished to confirm the conclusions of the Cochrane review of psychological interventions for chronic pain [34]. Given the divergence identified in the results of three trials led by Dr Monticone [16,19,21], we assessed a total of 10 trials from the same research group examining chronic spinal pain. Eight reported very large effect sizes for pain, disability and HRQoL for comparisons of CBT and physical rehabilitation versus physical rehabilitation alone. In context, these are about 20 times the standard effect size of the comparison 'CBT vs active care'. This level of effectiveness is highly unusual in a single trial, let alone a group of trials. Expressed as an NNT of 2 or below, they are not only outliers in comparison to other CBT trials but, if treated separately as a specific treatment, they would give the best NNTs ever recorded, a 'best in class' treatment compared with any other psychological, physical, rehabilitative, or pharmacological treatment examined in any chronic pain condition.

There are no data in the 10 published reports to suggest that the treatments in these 8 trials are more potent than the norm. There is no indication of any aspects of the experimental treatment uniquely different to the CBT and rehabilitation provided in other trials: staff training and experience, treatment content, intensity, and mode of delivery were unremarkable. Similarly, there are no obvious reasons from the published reports for the excellent participant retention data. Using the Cochrane Risk of Bias tool, the trials have a normal (for this field) RoB profile.

By contrast, the analysis of trustworthiness, using the CPC-TST tool, revealed several anomalies. First, on governance: none of the trials was pre-registered, despite the International Committee of Medical Journal Editors (ICMJE) requirement for pre-registration, in place since 2005. Pre-registration protects integrity and increases trustworthiness by requiring a record of core methodological features of the trial, changes from which after study completion require justification. Second, on randomisation: the distribution of p values deviated from that expected with simple randomisation, with a skew towards higher p values, indicating that baseline average scores were broadly more similar than might be expected from simple randomisation. This is reflected across all trials combined, and, for many individual trials, combining p values for each trial produced p values that deviated substantially from 0.5. Third, all studies achieved groups of exactly equal size post-randomisation despite only four trials reporting a block method for randomisation. Fourth, there were identical or highly similar outcome data reported in trials presented as independent trials. Transposition error is possible within trials but hard to understand between

different trial reports. Errors in reporting p values raised further concern. In conclusion, data error, data similarity (or duplication), randomisation oddity, and p-value error, coupled with a failure to pre-register, are likely contributors to explaining the extreme positivity of these data. Dr Monticone has not shared evidence of ethical approval or IPD with us to allow independent scrutiny of these results. We did not consider that his responses to specific queries regarding randomisation, duplicate and highly similar data, or anomalous baseline p values adequately explained the issues raised. Overall, based on this analysis, we judge these trial data to be untrustworthy.

Our focus is on the reports of trials appearing in peer-reviewed scientific publications. We have no data on which to comment on the conduct or integrity of individual investigators. Data fabrication and alteration have occurred previously in pain research. In some cases, the evidence has been overwhelming, as in the cases of anaesthesia researchers Yoshitaka Fujii, Joachim Boldt, and others [1,4,14,35]. In other cases, the overall patterns of data put any natural explanation out of reach, leaving reasonable doubt about investigator conduct. None of the included trials were published in journals suspected or presumed to be "predatory" in nature (see Supplementary Table 4). Hayden et al. [10] recently explored aspects of publication integrity in a large cohort of clinical trials of exercise for low back pain. They found a growing number of trials published in presumed predatory journals. While publication in a predatory journal was not associated with reported outcomes, it was associated with a range of quality, reporting and integrity issues. Our results suggest that there is a need to carefully scrutinise trials in more trusted publications.

The scale of the problem of untrustworthy trials in pain is unknown. In a systematic review of surveys of researchers, 2% of researchers across scientific disciplines admitted to fabricating, falsifying or modifying data at least once themselves, and 14% believed that colleagues had falsified data [7]. In Norway, < 1% of researchers admit fabrication, falsification and plagiarism, but 40% admit questionable research practices [13]. As a community, we need to establish clearer routines of looking beyond bias to broader questions about the trustworthiness of evidence; one cofounder of the Committee on Publication Ethics and former BMJ editor, Richard Smith, suggested that we have reached the point where systematic reviewers should start by assuming that a study is fraudulent until they have evidence to the contrary [31]. "*A lot of what is published is incorrect*" (p1380) [12], and inclusion of untrustworthy studies in systematic reviews is not a trivial matter.

In current scientific editorial practice, where automatic integrity checks are not the norm, evaluating and raising concerns regarding the trustworthiness of studies lie in the hands of individual editors, peer reviewers, the broader research and clinical community, and initiatives such as Retraction Watch. Formal mechanisms and validated processes are currently lacking. Here we have used one of the developing approaches, strongly informed by the work of the Cochrane Pregnancy and Childbirth review group [6]. That group now applies its screening tool routinely to all trials identified in their systematic reviews, excluding from subsequent analyses trials considered to present any concerns. While there is some risk of losing potentially valuable evidence, such an approach would reduce the risk of reviews being distorted by untrustworthy data and should be actively considered. We might start by making pre-registration a prerequisite for the inclusion of trials in systematic reviews.

We conducted our review using a formal protocol published on the Open Science Framework [28]. It has some limitations. No available tools for exploring research integrity or trustworthiness have been formally validated and we selected the CPC-TST on the basis of face validity and perceived usability for the type of trials. In using Stouffer's method to combine p values, we acknowledge that the assumption of independence between pooled values is unlikely to be met for all variables and that this may have contributed to the observation of extreme combined p values. However, it does

not adequately explain the peak of p values of 1.0 in the observed distribution. We did not conduct further sensitivity analyses to explore correlations and exclude correlated variables. This is due to the lack of power in any such analyses and the issues it would raise as regards multiple testing. We focused on a sample of trials from a single author group evaluating similar interventions in similar patient groups. Our reasons for this were based on the prior observation of consistently extreme results in three trials from that group. This led us to consider whether this represented a broader pattern observable from other trials published from that group. We have not applied the same assessment to the broader evidence base on this topic or to other trials from this author group for conditions other than persistent spinal pain.

In summary, the results of eight of the included trials are highly divergent from norms in the evidence base for psychological therapies for persistent pain. Replication of these results outside a single institution would represent a substantial advance for pain medicine and very good news for patients living with pain. However, we have not found satisfactory plausible explanations for that divergence in either the details of the interventions themselves or how they were delivered, nor has reporting error been retrospectively declared. Our exploration of these studies has raised concerns in specific cases regarding trustworthiness, particularly relating to research governance and to the plausibility, integrity and accuracy of the data. Possible explanations for the latter include error, data manipulation, or data fabrication. Taken together or alone, we have no confidence in the veracity of these trial results and assert that these studies should be excluded from evidence syntheses on this topic and from clinical practice guidelines.

Acknowledgements

We would like to thank Professor Ben Mol for sharing early versions of the Cochrane Pregnancy and Childbirth TST**.**

Declarations of interest

NOC: None known.

RAM: None known.

AW: AW was an author on the most recent update of the Williams review [21]

EF: EF is funded by Versus Arthritis and was an author on the most recent update of the Williams review [21]. EF has no other declarations of interest in relation to this manuscript.

LH: None known; LH was an author on the most recent update of the Williams review [21]

GS: None known

AE: None known

CE: CE was an author on the most recent update of the Williams review [21]

References

1. Adam M. Japanese PONV Researcher Probed in Sweeping Research Fraud Case. Anesthesiol News 2012;7 March issue.

2. Bordewijk EM, Wang R, Askie LM, Gurrin LC, Thornton JG, van Wely M, Li W, Mol BW. Data integrity of 35 randomised controlled trials in women' health. Eur J Obstet Gynecol Reprod Biol. 2020;249:72-83.

3. Bordewijk EM, Li W, van Eekelen R, Wang R, Showell M, Mol BW, van Wely M. Methods to assess research misconduct in health-related research: A scoping review. J Clin Epidemiol. 2021;136:189-202.

4. Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia. 2012;67(5):521-37. doi: 10.1111/j.1365-2044.2012.07128.x.

5. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia 2017;72:944–52.

6. Cochrane Pregnancy and Childbirth Review Group. Identifying and handling potentially untrustworthy trials in Pregnancy and Childbirth Cochrane Reviews. https://pregnancy.cochrane.org/news/identifying-and-handling-potentially-untrustworthy-trials-pregnancy-and-childbirth-cochrane. Accessed 16/9/21

7. Fanelli D. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLoS ONE 2009;4(5):e5738.

8. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. P T 2008;33(12):700-11. PMID: 19750051; PMCID: PMC2730804.

9. Grey A, Bolland MJ, Avenell A, Klein AA, Gunsalus CK. Check for publication integrity before misconduct. Nature. 2020;577(7789):167-169. doi: 10.1038/d41586-019-03959-6.

10. Hayden JA, Ellis J, Ogilvie R, Boulos L, Stanojevic S. Meta-epidemiological study of publication integrity, and quality of conduct and reporting of randomized trials included in a systematic review of low back pain. J Clin Epidemiol. 2021;134:65-78.

11. Higgins JPT, Altman DG, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS, editors(s). Cochrane Handbook for Systematic Reviews of Interventions version 5.2.0 (updated June 2017). Cochrane, 2017. [Available from www.training.cochrane.org/handbook]

12. Horton R. What is medicine’s 5 Sigma? Lancet 2015;385:1380.

13. Kaiser M, Drivdal L, Hjellbrekke J, Ingierd H, Rekdal OB. Questionable Research Practices and Misconduct Among Norwegian Researchers. Sci Eng Ethics 2022;28:2.

14. Kharasch ED. Scientific Integrity and Misconduct - Yet Again. Anesthesiology 2021;135(3):377-9.

15. Monticone M, Baiardi P, Vanti C, Ferrari S, Nava T, Montironi C, et al. Chronic neck pain and treatment of cognitive and behavioural factors: results of a randomised controlled clinical trial. Eur Spine J 2012;21(8):1558-66.

16. Monticone M, Ferrante S, Rocca B, Baiardi P, Dal Farra F, Foti C. Effect of a long-lasting multidisciplinary program on disability and fear-avoidance behaviors in patients with chronic low back pain: results of a randomized controlled trial. Clin J Pain 2013;29(11):929-38. [CRSREF: 14279860]

17. Monticone M, Ambrosini E, Rocca B, Magni S, Brivio F, Ferrante S. A multidisciplinary rehabilitation programme improves disability, kinesiophobia and walking ability in subjects with chronic low back pain: results of a randomised controlled pilot study. Eur Spine J 2014a;23(10):2105-13.

18. Monticone M, Ferrante S, Teli M, Rocca B, Foti C, Lovi A, et al. Management of catastrophising and kinesiophobia improves rehabilitation after fusion for lumbar spondylolisthesis and stenosis. A randomised controlled trial. Eur Spine J 2014b;23(1):87-95.

19. Monticone M, Ambrosini E, Rocca B, Cazzaniga D, Liquori V, Foti C. Group-based task-oriented exercises aimed at managing kinesiophobia improved disability in chronic low back pain. Eur J Pain 2016a;20:541-51. [CRSREF: 14279862]

20. Monticone M, Ambrosini E, Cazzaniga D, Rocca B, Motta L, Cerri C, et al. Adults with idiopathic scoliosis improve disability after motor and cognitive rehabilitation: results of a randomised controlled trial. Eur Spine J 2016b;25(10):3120-9.

21. Monticone M, Ambrosini E, Rocca B, Cazzaniga D, Liquori V, Pedrocchi A, et al. Group-based multimodal exercises integrated with cognitive-behavioural therapy improve disability, pain and quality of life of subjects with chronic neck pain: a randomized controlled trial with one-year follow-up. Clin Rehabil 2017;31(6):742-52. [CRSREF: 14279864]

22. Monticone M, Ambrosini E, Vernon H, Rocca B, Finco G, Foti C, et al. Efficacy of two brief cognitive-behavioral rehabilitation programs for chronic neck pain: results of a randomized controlled pilot study. Eur J Phys Rehabil Med 2018;54(6):890-9.

23. Monticone M, Ambrosini E, Rocca B, Cazzaniga D, Liquori V, Lovi A, et al. Multimodal exercises integrated with cognitive-behavioural therapy improve disability of patients with failed back surgery syndrome: a randomized controlled trial with one-year follow-up. Disabil Rehabil 2020;27:1-8.

24. Monticone M, Ambrosini E, Portoghese I, Rocca B. Multidisciplinary program based on early management of psychological factors reduces disability of patients with subacute low back pain. Results of a randomised controlled study with one year follow-up. Eur J Phys Rehabil Med 2021;57(6):959-67.

25. Moore A, Derry S, Eccleston C, Kalso E. Expect analgesic failure; pursue analgesic success. BMJ 2013;346:f2690. doi: 10.1136/bmj.f2690. PMID: 23645858.

26. Moore RA, Moore OA, Derry S, Peloso PM, Gammaitoni AR, Wang H. Responder analysis for pain relief and numbers needed to treat in a meta-analysis of etoricoxib osteoarthritis trials: bridging a gap between clinical trials and clinical practice. Ann Rheum Dis. 2010;69(2):374-9.

27. Nicholas MK, Asghari A, Blyth FM. What do the numbers mean? Normative data in chronic pain measures. Pain. 2008;134(1-2):158-7.

28. O'Connell N, Moore A, Stewart G, Fisher E, Erskine A, Hearn L, Eccleston C, Williams ACdeC. Investigating the veracity of a sample of published trials with divergent results in spinal pain. OSF Registries. https://osf.io/345vq

29. O'Keeffe M, Purtill H, Kennedy N, Conneely M, Hurley J, O'Sullivan P, Dankaerts W, O'Sullivan K. Comparative Effectiveness of Conservative Interventions for Nonspecific Chronic Spinal Pain:

Physical, Behavioral/ Psychologically Informed, or Combined? A Systematic Review and Meta-Analysis. J Pain 2016;17(7):755-74. doi: 10.1016/j.jpain.2016.01.473 .

30. Review Manager (RevMan) [Computer program]. Version 5.4. The Cochrane Collaboration, 2020.

31. Smith R. Time to assume that health research is fraudulent until proven otherwise? BMJ Opinion 2021. https://blogs.bmj.com/bmj/2021/07/05/time-to-assume-that-health-research-is-fraudulent-until-proved-otherwise/ accessed 28/7/21

32. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM Jr (1949) The American Soldier, Vol. 1 - Adjustment during Army Life. Princeton, Princeton University Press.

33. The jamovi project (2021). *jamovi* (Version 1.6.23) [Computer Software]. Retrieved from https://www.jamovi.org.

34. Williams ACC, Fisher E, Hearn L, Eccleston C. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev. 2020;8(8):CD007407

35. Wise J. Boldt: the great pretender. BMJ 2013;346:f1738.