Original citation:

Permanent WRAP URL:
http://wrap.warwick.ac.uk/81942

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) license and may be reused according to the conditions of the license. For more details see: http://creativecommons.org/licenses/by-nc/4.0/

A note on versions:
The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk
The power of low back pain trials: A systematic review of power, sample size, and reporting of sample size calculations over time, in trials published between 1980 and 2012

Robert Froud, PhD\textsuperscript{1,2*}, D\textsuperscript{é}van Rajendran, MSc\textsuperscript{1,3}, Shilpa Patel, CPsychol\textsuperscript{2}, Philip Bright, MSc\textsuperscript{3}, Tom Bjørkli, BSc\textsuperscript{1}, Sandra Eldridge, PhD\textsuperscript{4}, Rachelle Buchbinder, PhD\textsuperscript{5}, Martin Underwood, MD\textsuperscript{2}

*Corresponding author r.froud@warwick.ac.uk

1. Kristiania University College, Prinsens Gate 7-9, 0152, Oslo, Norway
2. Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
3. European School of Osteopathy, The Street, Boxley, Maidstone, Kent. ME14 3DZ, UK
4. Monash Department of Clinical Epidemiology, Cabrini Institute and Department of Epidemiology and Preventive Medicine, Monash University, Suite 41, Cabrini Medical Centre, 183 Wattletree Road, Malvern, 3144, Melbourne, Victoria, Australia
5. Barts and the London School of Medicine and Dentistry, Queen Mary University of London, 2 Newark Street, Whitechapel, London, E1 2AT, UK
1 Funding: Campus Kristiania, University of Warwick, Monash University, and Barts and the London Charity. RB is funded by an Australian National Health and Medical Research Council (NHMRC) Senior Principal Research Fellowship.
Abstract

Study design


Objectives

To explore what proportion of trials have been powered to detect different bands of effect size; whether there is evidence that sample size in LBP trials is increasing with time; what proportion of trial reports include a sample size calculation; and whether likelihood of reporting sample size calculations is increasing with time.

Summary of background data

Clinical trials should have a sample size sufficient to detect a minimally important difference for a given power and type I error rate. An underpowered trial is one within which probability of type II error is too high. Meta-analyses do not mitigate underpowered trials.

Methods

Reviewers independently abstracted data on sample size at point of analysis, whether a sample size calculation was reported, and year of publication.

Descriptive analyses were used to explore ability to detect effect sizes, and
regression analyses to explore the relationship between sample size, or reporting sample size calculations, and time.

Results

We included 383 trials. Only one-third were powered to detect a standardised mean difference of 0.5 or less. Average sample size was 153 people, which increased only slightly (~4 people/year) from 1980 to 2000, and declined slightly (~4.5 people/year) from 2005 to 2011 ($P<0.00005$). Sample size calculations were reported in only 41% of trials. The odds of reporting a sample size calculation (compared to not reporting one) increased until 2005 and then declined ($OR_{year}=1.06, OR_{year^2}=0.99; P<0.00005$).

Conclusions

Trialists should make LBP trial samples sizes large enough to detect realistic effect sizes, and report sample size calculations. It may be justifiable to power a trial to detect only large effects in the case of novel interventions. Funders, peer-reviewers, editors, and readers should be vigilant for and critical of underpowered trials.

Key words

Back pain; Sample size; Sample size calculations; Power; Reporting; Ethics; Effect sizes; Minimally important difference; Clinical trials; Systematic Review.
Key points

1. Only around one-third of trials are powered to detect an effect size of 0.5 or less. The vast majority (95%) do not have the power to detect realistic effect sizes.
2. Sample size in back pain trials has only increased by trivial magnitudes since 1980 and has been declining since 2005.
3. Sample size calculations are reported by only 41% of trials, and since 2005 the proportion reporting sample size calculations has fallen.
4. Except for cases in which novel interventions are being tested and a small sample size is clearly justified by design, it is imperative that back pain trial sample sizes are made larger, the required size is established *a priori* and is reported in publications.
Froud et al systematically reviewed non-specific low back pain trials published between 1980 and 2012, and show most are not powered to detect realistic effect sizes and that the majority do not report sample size calculations. They discuss the ethical implications, possible mitigating circumstances, and outline a strategy for improvement.
Introduction

Statistical tests in clinical trials generally test null hypotheses that differences between populations are zero.\(^1\) Customarily, type I error rate (or alpha) is set at 0.05, \(i.e.\) rejection of the null hypothesis if there is <5% probability it is true, given the observed result. Also customarily, the power of the study, \(i.e.\) the probability of detecting a difference if one truly exists in the population, is usually set at 80% or 90%.

The magnitude a study is powered to detect should be based on the population-difference that is perceived as worthwhile.\(^2\) Non-specific low back pain (LBP) trials utilise many different outcome measures, in different units, but outcomes can be standardised by dividing the between-group difference by the standard deviation to produce a standardised effect size (ES).\(^3\) Conventionally, ESs of 0.2 are considered small, 0.5 medium, and 0.8 large.\(^6\) Whether or not the magnitude can be considered worthwhile, at a population level, may depend on the cost of providing the treatment in question.\(^7\)\(^8\) Large, high-quality trials suggest that effective interventions may have a 'typical' ES of around 0.3.\(^3\)\(^9\)\(^10\)

If a trial is underpowered then there is a greater risk of type II error; \(i.e.\) failing to detect a difference when one exists in the population.\(^1\) Thus, assuming good internal validity, it is difficult to interpret outcomes in trials reporting null results, in cases where sample sizes are small. It has been suggested that in these cases meta-analyses of underpowered studies provide remedy.\(^11\) However, LBP populations may often be too heterogeneous to permit sensible pooling in fixed
effects models, and random-effects models make assumptions about population
distributions that may be unrealistic.\textsuperscript{12-15} Additionally, pooling of different
outcomes purporting to measure the same domains, with outcomes described in
terms of ES may be unsafe.\textsuperscript{16} Further, therapist-delivered interventions testing
the same basic intervention may be so dissimilar, and poorly specified, that it is
not possible to describe the intervention evaluated by the meta-analysis in a
manner that would allow replication.

There is a risk that underpowered trials have very limited value and contribute
little to the evidence base. There have been several calls to increase sample sizes
in trials over the past decades, and the importance of undertaking sample size
calculations has been emphasised.\textsuperscript{1,17-19} The impact that calls and checklists have
had on practice relating to power and reporting sample size calculations in back
pain trials is unclear.

Our aims were to explore: 1) what proportion of LBP trials have been powered to
detect bands of effect size ranging between small ($d=0.2$), 'typical' ($d=0.3$),
medium ($d=0.5$), or large ($d=0.8$); 2) whether there is evidence that sample size
in LBP trials is increasing with time; 3) what proportion of trial reports include a
sample size calculation, by year; and 4) whether the likelihood of reporting a
sample size calculation is increasing with time.
Materials and methods

Two reviewers (RF, SP) independently identified randomised controlled trials (RCTs), published from 1980 inclusive, of interventions for non-specific low back pain (nsLBP) from COST-B13’s European Guidelines for the Management of Low Back Pain (EGLBP) and from the systematic reviews reported in the EGLBP.20 The EGLBP search was extended in two phases. In 2007, two reviewers (RF, SP) searched from November 2002, when the COST-B13 search finished, to the end of December 2006, using the Cochrane library, PubMed, Embase, HTA, and Lilacs databases. In 2012, three reviewers (PB, DR, TB) searched from January 2007 to January 2012, using PubMed, Embase, and the Cochrane library, within which the majority of LBP trials are indexed. Supplemental Digital Content file 1 (PDF) shows a typical search string. Table 1 shows inclusion and exclusion criteria.
We combined material from the EGLBP and extended searches, removing duplicates, and short-listing by title and abstract. Full-texts were obtained if the titles and abstract alone contained insufficient information for assessment against Table 1 criteria.

All five reviewers abstracted data from the LBP trials on numbers analysed for the primary outcome for the pre-specified main comparison, details of whether or not a sample size calculation was reported, and the year of publication. An outcome was identified as ‘primary’ if (1) the outcome was nominated as such; if no outcome was nominated, or if multiple outcomes were nominated, we used (2) the outcome measure on which the sample size calculation was based; if this was not reported, we identified (3) the first outcome measure referred to in the abstract; and if this was not identified, we used (4) the first outcome mentioned in the paper. This approach has been taken in several other methodological reviews.\(^{21-23}\) We identified the primary end-point, or used the first follow-up time point in cases when this was not clear. Disagreements were resolved through discussion and, if necessary, with arbitration and a fourth reviewer (RF, SE, or MU). Reporting of sample size calculation was judged present if, as a minimum, it was reported how many participants were estimated as required to detect a given between-group difference, or if the latter was not reported, if the detectable between-group difference could be readily calculated from what was reported. The calculations needed to be \textit{a priori} (\textit{i.e.} \textit{post hoc} power calculations were not counted).
For all data abstraction we used expert validation of 20% of papers, as has been done in other methodological reviews.\textsuperscript{24} It was decided \textit{a priori} that in the case of agreement being less than 80%, full independent abstraction would be undertaken. Validation disagreements were settled by arbitration involving an independent reviewer.

For each trial, the average sample size by arm was calculated and compared to the threshold of participants needed to detect an ES of small (Cohen’s $d=0.2$), small-to-medium ($d=0.3$), medium ($d=0.5$), or large ($d=0.8$), under the assumption of comparisons to groups of approximately equal size. Trials were categorised according to their ability to detect within these bands.\textsuperscript{4}

The proportion of trials reporting sample size calculations was calculated by year. To explore our objective of whether sample size is increasing with time, we used linear regression with a log-transformation of sample size to control for the heteroscedasticity (\textit{i.e.} the changing variance) of sample size on year. To explore the likelihood of reporting a sample size calculation by time, we fitted a logit model, regressing reporting of sample size (binary) on year. Non-linear relationships between predictor and outcome variables were explored through fitting polynomial terms in predictor variables and log rank tests were used to test these terms for statistical significance ($\alpha=0.05$). Model fit was examined using residual plots, and in the case of the logit model, a combination of Pseudo $R^2$, AIC, and residual plots. Locally weighted least squares (Lowess) smoothers were added if they aided interpretations of trends over time. All analyses were performed in Stata, version 12 (Statacorp, Texas).
Results

From our combined searches, we sifted a total of 7,066 hits, removing 4,398 of these based on titles and abstracts and 1,931 duplicates, and assessed 737 papers at full-text level. We rejected 354 papers that did not fit our inclusion criteria. We included 383 trials (Figure 1). [Note: this makes for a large reference list. References can be included with the supplemental digital content tables if preferred.] Supplemental Digital Content file 2 (PDF) shows a table of the characteristics of included studies, and Supplemental Digital Content file 3 (PDF) shows a table of the characteristics of excluded studies.

Ability of trials to detect effect sizes

Across all trials, the average total sample size was 152.8 (IQR=126; range 12 to 2,594). Figure 2 shows histograms of average group size and logged average group size, with reference guidelines for ES detection ability. Table 2 shows the frequency of trials by ES detection category. The majority (67.1%) of trials were only powered to detect ES of greater than 0.5 (medium).

Sample size over time

A quadratic term in year was required to adequately model sample size over time (i.e. the relationship follows a quadratic curve rather than a line (P<0.0005)). Linear regression of sample size on year and year² explained 4.6% of the variance in sample size. For use in estimation, β_{year}=-0.0131702, β_{year²}=-0.0027846 (centered around the average in year, of 2002.141), with a constant of 31.1585; P<0.00005). For example, the predicted log sample size for the year 2000 (i.e. the year with the largest sample sizes) is: 31.1585+(2000*-
The power of LBP trials

0.0131702 + ((2000-2002.141)^2 - 0.0027846) = 4.82; i.e. a total sample size of around 124 people. Other years may be inserted into this equation, as required.

Figure 3 shows a plot of the model with 95% CIs fitted, and a Lowess smoother, which more clearly highlights the turning point and fall in sample size after 2005.

The reporting of sample size calculations in general

Across the whole time period of interest, the proportion of trials in which a sample size calculation was reported is 41% (95%CI 36.3 to 46.4).

The likelihood of reporting sample size calculations by time

From the logit model, the odds ratio for reporting a sample size calculation by year is 1.093 (1.06 to 1.13), P<0.00005, Pseudo R^2=6.7% (i.e. very strong evidence of an increase in odds of reporting sample size calculation of 9% per year). However, model fit is improved by adding quadratic term in year (i.e. the increase in odds of reporting sample size calculation curves over time), P=0.0001. In the quadratic model, β_{year}=0.057851 (OR=1.060, 95% CI 1.017 to 1.104), β_{year^2}=-0.0099829 (OR=0.990, 95% CI 0.984 to 0.996), P<0.00005, and Pseudo R^2=9.7%. For use in prediction (i.e. where \( Pr=e^{xb}/1+e^{xb} \) and xb is the linear predictor), the constant, when year is unadjusted and year^2 is centered on its average, is -115.806. Figure 4 shows the proportion of trials reporting sample size calculations, by year, and weighted by the number of trials per year, with fitted values for the probability of reporting a sample size calculation. It shows 2005 marks a turning point toward deterioration in terms of the proportion reporting sample size calculations.
Discussion

Main findings

Only around one-third of trials are powered to detect an ES of 0.5 or less. Only around 5% of trials are powered to detect ‘typical’ magnitudes of 0.3 or less. There is evidence of an increase of sample size over time from 1980, but the increase is trivial, and slips into decline from 2005. Sample size calculations are reported in only 41% of trials published between 1980 and 2012. The likelihood of reporting sample size calculation increases by several percent each year, but begins to decrease after 2005.

Implications

Large, high quality trials of interventions for back pain that have convincingly demonstrated effectiveness, suggest a typical ES for LBP interventions is around 0.3. However, between 1980 and 2012, 95% of published trials have not had sufficient power to detect such magnitudes. There is a culture of conducting underpowered trials in the field of nsLBP research that needs to be addressed – the distribution of sample size shown in Figure 2 needs to be right-shifted. Underpowered trials have limited utility and are ethically questionable. As explained at the introduction, meta-analytic techniques do not mitigate underpowered trials, as even in random effects models assumptions must be made about the between-study heterogeneity that may often be invalid; especially in cases where heterogeneity may be due to methodological reasons.
rather than clinical heterogeneity.\textsuperscript{15} That said, there might be a place for smaller trials. As many larger high-quality trials of interventions for low back pain demonstrate only small-to-moderate differences, a counter-argument can be made that continuing to do these will do little to progress the field. Smaller well-conducted trials of novel interventions that are only powered to detect larger effect sizes could indicate promising avenues for intervention development. However, in such cases there should be clear rationale for this and great care to describe the novel intervention for subsequent replication.

Only 41\% of trials even report sample size calculations. In 1992, Cohen pointed out that while there is no disagreement between methodologists on the utility of such calculations, progress over the last 25 years had been slow.\textsuperscript{1} It is regrettable that in the back pain field, in the quarter-century that followed his observation, progress continues to be slow. From 2005, it has even deteriorated.

As change must be affected at the level of individual trialists, the implications are that trialists, funders, peer-reviewers, and journal editors each have a role to play in helping to improve the situation.

\textit{Comparisons to existing research}

Castellini \textit{et al}, in 2016, reviewed RCTs of mechanical low back pain published after 1968.\textsuperscript{761} Among 222 included studies, they found that 36\% reported a sample size calculation, but only 16\% reported what they deemed complete calculations. They found that both reporting and completeness of reporting improved between 1968 and 2013, and show a relative decline in the proportion
reporting sample size calculations since 2005. This is consistent with our data.

Their point estimate that 36% report sample size calculations is slightly less than ours. We suggest the following possible reasons; 1) Castellini’s work dates back to publications in the late 1960s, which predates several calls for improvements in sample size (although only eight trials (3.6%) predate 1980); 2) our sample is based on 58% more trials; 3) Castellini’s search used the terms ‘rehabilitation’, so the population may have been slightly different to ours; and 4) we required a priori reporting of the number of participants required to detect a given between-group difference, as minimum criteria to judge reporting the calculation. While it is clear how Castellini judged completeness of reporting of sample size calculations, it is not clear whether minimal criteria were used to judge reporting of a basic sample size calculation.

Moher et al reviewed 383 trials, published between 1975 and 1990, exploring their power to detect 25% and 50% relative differences if null results were reported. They found that 25% reported null results, and of these only 16% and 36% had the power to detect these magnitudes of difference. They observed that these percentages did not increase over time. Whilst our results are not directly comparable, we found that sample size in general has increased over time, albeit by a trivial amount, until 2005. Moher also found that in trials reporting null results, 32% reported sample size calculations, and that this increased from 0% in 1975 to 43%, in 1990. Since we did not separately explore reporting in trials reporting null results, our results are not directly comparable, although we also observed an increase over time, at least until 2005. Given both Moher’s and our results, we note the implication that in 1990, the likelihood of
The power of LBP trials

reporting a sample size calculation seems greater if the trial reported a null result.

Strengths and limitations

This work provides a useful account of trends of sample sizes in back pain trials and reporting of sample size calculations. It comes from a comprehensive review of trials that were published between 1980 and 2012, during which time there were several calls to ensure such trials had appropriate sample sizes.\textsuperscript{1,17-19}

One limitation of our work is that we did not consider the extent of the reporting of sample size calculations or likelihood of reporting by positive or negative results. This work was part of a larger systematic review addressing several other methodological research questions and these aspects were not prioritized; however, we acknowledge Castillini’s and Moher’s useful contributions here.\textsuperscript{17,761}

Caution should be noted in using the quadratic models for future prediction, as trajectories of falls will be predicted to continue along the quadratic curve, which peaked in 2000 and 2005, respectively. Should any authors like to update the model we will be happy to share our data. We could have described our findings descriptively, but model fitting was useful insofar as it demonstrates the trend and the weight of evidence for the observations.
Conclusions

1. That the vast majority (95%) of trials are not powered to detect realistic effect sizes might be considered unethical, unless novel interventions are being tested and the small size is clearly justified. Increases in sample sizes have been trivial, and decline after 2005. The majority of authors still do not report sample size calculations.

2. Trialists should plan to make the size of LBP trials large enough to detect realistic effect sizes and report sample size calculations. It may be justifiable to power a trial only to detect large effect sizes in the case of testing a novel intervention. Funders, peer-reviewers, editors, and readers should be vigilant for and critical of underpowered trials, especially if these report null results.
Conflicts of Interest and Source of Funding: RF and MU are directors and shareholders of Clinvivo Ltd, a company that provides health outcome data collection services. These services were not used in this study. Høyskole Kristiania received funding from University of Warwick and Monash University to complete this research. Barts and The London Charity provided funding for the first stage of this research.
References


77. Codding C, Levinsky D, Hale ME, et al. Efficacy and safety evaluation of 12 weeks extended-release hydrocodone/acetaminophen treatment in patients...


The power of LBP trials


144. Hancock. letter. 2010.


The power of LBP trials


188. Kinals R, Kuwik W, D P. The comparison of the results of manual
therapy versus physiotherapy methods used in treatment of patients with low

experimental sensitization and pain in chronic back pain patients. Anesth Analg

therapy, physiotherapy, and treatment by the general practitioner for
nonspecific back and neck complaints. A randomized clinical trial. Spine

trial of manual therapy and physiotherapy for chronic back and neck complaints:

manipulative therapy and physiotherapy for persistent back and neck

treatments for low-back pain: results from a randomized controlled trial. Medical

194. Konrad K, Tatrai T, Hunka A, et al. Controlled trial of balneotherapy in

increases work days in patients with nonacute nonspecific low back pain: 1-year
results from a randomized controlled trial. Arch Phys Med Rehabil 2007;88:1089-
94.


The power of LBP trials


The power of LBP trials


254. Nath S, Nath CA, Pettersson K. Percutaneous lumbar zygapophysial (Facet) joint neurotomy using radiofrequency current, in the management of
The power of LBP trials


276. Podichetty VK, Varley ES. Re: Oleske D M, Lavender S A, Andersson G B, et al. Are back supports plus education more effective than education alone in...


8. Smeets RJEM, Vlaeyen JWS, Hidding A, et al. Active rehabilitation for chronic low back pain: Cognitive-behavioral, physical, or both? First direct post-

---

The power of LBP trials
The power of LBP trials

treatment results from a randomized controlled trial [ISRCTN22714229]. *BMC Musculoskelet Disord* 2006;7.


The power of LBP trials


The power of LBP trials


1 354. Tsukayama H, et al. Randomised Controlled Trial Comparing the
2 Effectiveness of Electroacupuncture and TENS for Low Back Pain: A Preliminary
4 355. Underwood M, Mistry D, Lall R, et al. Predicting response to a cognitive-
5 behavioral approach to treating low back pain: Secondary analysis of the BeST
7 356. Vasseljen O, Fladmark AM. Abdominal muscle contraction thickness and
8 function after specific and general exercises: a randomized controlled trial in
10 357. Vickers AJ. Statistical reanalysis of four recent randomized trials of
13 improves outcome in comparison with instrumented posterolateral fusion: long-
16 movement/(re)injury in chronic low back pain: further evidence on the
19 Prolotherapy injections, saline injections and exercises for chronic low-back
22 361. Warming S, Ebbehoj NE, Wiese N, et al. Little effect of transfer technique
23 instruction and physical fitness training in reducing low back pain among nurses:


The power of LBP trials


The power of LBP trials


Chan CW, Mok NW, Yeung EW. Aerobic exercise training in addition to conventional physiotherapy for chronic low back pain: a randomized controlled trial. *Arch Phys Med Rehabil* 2011;92:1681-5.


The power of LBP trials


Durmus D, Durmaz Y, Canturk F. Effects of therapeutic ultrasound and electrical stimulation program on pain, trunk muscle strength, disability, walking


The power of LBP trials


The power of LBP trials


The power of LBP trials


The power of LBP trials


The power of LBP trials


The power of LBP trials


The power of LBP trials


The power of LBP trials


601. Little P, Roberts L, Blowers H, et al. Should we give detailed advice and information booklets to patients with back pain? A randomized controlled


The power of LBP trials


667. Petersen T, Kryger P, Ekdahl C, et al. The effect of McKenzie therapy as compared with that of intensive strengthening training for the treatment of
patients with subacute or chronic low back pain: A randomized controlled trial.


2  on a non-injury model compared with individual symptom-based physical
3  training in chronic LBP. A pragmatic, randomised trial with a one-year follow-up.
4  *BMC Musculoskelet Disord* 2010;11:212.
5  705.  Soriano F, Rios R. Gallium Arsenide Laser Treatment of chronic low back
7  706.  Spinhoven P, Ter Kuile M, Kole-Snijders AM, et al. Catastrophizing and
8  internal pain control as mediators of outcome in the multidisciplinary treatment
12  2004;140:77-84.
14  prospective randomized trial: McKenzie method of treatment versus patient
17  multi-stage return to work program for workers on sick-leave due to low back
20  intervention in sub-acute low back pain: short-term results of a single-blind


The power of LBP trials


1 744. Weiner DK, Rudy TE, Glick RM, et al. Efficacy of Percutaneous Electrical
2 Nerve Stimulation for the Treatment of Chronic Low Back Pain in Older Adults.
5 acceptance and commitment therapy and cognitive-behavioral therapy for
8 acute low back pain patients: a randomized clinical trial with one-year follow-up
13 and efficacy of iyengar yoga therapy on chronic low back pain. *Spine (Phila Pa
14 1976)* 2009;34:2066-76.
16 manipulation study (ROMANS): Pragmatic trial for spinal pain in primary care.
18 750. Witt CM, Jena S, Selim D, et al. Pragmatic randomized trial evaluating the
19 clinical and economic effectiveness of acupuncture for chronic low back pain.
22 combined with group exercise for acute and subacute low back pain. *Spine.*
24 752. Wu JX, Wang B. Early intervention of aerobic exercise to rehabilitation of
The power of LBP trials


8. Zerbini C, Ozturk ZE, Grifka J, et al. Efficacy of etoricoxib 60 mg/day and diclofenac 150 mg/day in reduction of pain and disability in patients with

### Tables

#### Table 1: Inclusion and exclusion criteria, and the order of their evaluation

<table>
<thead>
<tr>
<th>Inclusion criterion</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCTs of nsLBP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order of evaluation</th>
<th>1</th>
<th>Non-English language reports</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>Studies that were not RCTs or presented insufficient information for us to determine whether randomisation was used to allocate participants</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Reports that self-identified as pilot/feasibility studies</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Cross-over designs (because of limited utility in the LBP field)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>RCTs with mixed samples (<em>e.g.</em> neck or thoracic pain in addition to LBP), samples of participants with radiating leg pain, or referred pain extending past the knee, or samples including LBP specific pathology (<em>e.g.</em> cancer, ankylosing spondylitis, or disc herniation) or pregnancy</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Trials using solely objective or psychological outcome measures</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Non-inferiority designs</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Follow-up studies with no new outcome measures, and multiple publications. In the case of multiple publications, we included the first published article and excluded subsequent publications</td>
</tr>
</tbody>
</table>

RCT=Randomised controlled trial; nsLBP=Non-specific low back pain
Table 2: Trials' ability to detect different categories of effect size

<table>
<thead>
<tr>
<th>Ability to detect category of ( d )</th>
<th>( n )-band (average arm guide) where power=80%, ( \alpha=0.05 )</th>
<th>Number of trials in category</th>
<th>Proportion of trials in category (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d&lt;0.2 )</td>
<td>( 392&lt;n )</td>
<td>3</td>
<td>0.8 (0.2 to 2.3)</td>
</tr>
<tr>
<td>( 0.2&lt;d&lt;0.3 )</td>
<td>( 174&lt;n&lt;393 )</td>
<td>17</td>
<td>4.4 (2.6 to 7.0)</td>
</tr>
<tr>
<td>( 0.3&lt;d&lt;0.5 )</td>
<td>( 62&lt;n&lt;175 )</td>
<td>106</td>
<td>27.7 (23.3 to 32.4)</td>
</tr>
<tr>
<td>( 0.5&lt;d&lt;0.8 )</td>
<td>( 23&lt;n&lt;63 )</td>
<td>175</td>
<td>45.7 (40.6 to 50.8)</td>
</tr>
<tr>
<td>( 0.8&lt;d )</td>
<td>( n&lt;24 )</td>
<td>82</td>
<td>21.4 (17.4 to 25.9)</td>
</tr>
</tbody>
</table>
**Figure legends**

**Figure 1** – The figure shows a flow chart of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions.

**Figure 2** – The figure shows histograms of average group size and logged average group size, with reference guidelines for effect size detection thresholds (assuming comparison to another group of equal size). Referenced effect sizes can be detected only by trials to the right of each line.

**Figure 3** – The figure shows fitted values with 95% CI, Lowess smoother and residuals of log sample size on year of publication.

**Figure 4** – The figure shows the proportion of trials reporting sample size calculations, by year, and weighted by the number of trials per year, with fitted values for the probability of reporting a sample size calculation.
Figure 1

1980 - 2002
- EGLBP: 320

2002 - 2006
- PubMed, EMBASE, PsycINFO, Cochrane, HTA, Lilacs: 1,776

2007 - 2012
- PubMed 1,207
- EMBASE 2,641
- Cochrane 1,122

Total hits = 7,066

Duplicates excluded = 1,931
Title & Abstract filter = 4,398

Non-English text = 4
Not RCT of nsLBP, or insufficient information to identify as RCT = 85
Pilot/preliminary/feasibility = 17
Cross-over design = 9
Mixed samples, radiations, or referred pain below the knee = 179
Purely objective or psychological outcome measures = 17
Non-inferiority designs = 4
Follow-up with no new outcomes or multiple publication = 39
Total Excluded = 354

Full text retrieved = 737

Articles included = 383
Click here to access/download

Supplemental Data File (doc., pdf., xls., etc.)
SDC1.pdf
Click here to access/download

**Supplemental Data File (doc., pdf., xls., etc.)**

SDC2.pdf