Problems caused by heterogeneity in meta-analysis: a case study of acupuncture trials

Stephanie L Prady, Jane Burch, Simon Crouch, Hugh MacPherson

ABSTRACT

Objectives To illustrate the pitfalls of using meta-analysis to combine estimates of effect in trials that are highly varied and have a high potential for bias.

Methods We used a random-effects meta-analysis to pool the results of 51 sham-controlled acupuncture trials of chronic pain published in English before 2008 and explored the heterogeneity using meta-regression. We repeated the process on a subset of these trials that used a visually credible non-penetrating sham device as control (N = 12).

Results In both analyses there were high levels of heterogeneity and many studies were at risk from potential bias. The heterogeneity was not explained by meta-regression.

Conclusions Trials of interventions that have high potential for bias, such as many in the acupuncture literature, do not meet the assumptions of the statistical procedure that underlie random-effects meta-analysis. Even in the absence of bias, heterogeneity in meta-analyses is not accounted for by the CIs around the pooled estimate.

INTRODUCTION

Trials of complex (multicomponent) interventions like acupuncture have factors that may influence the underlying true effect of the treatment relative to trials with different designs. Such components include population, acupuncture ‘dose’, acupuncturists, condition severity, comparator intervention and analysis methods. Many systematic reviews of acupuncture contain meta-analyses providing a pooled estimate of effect. In a meta-analysis, variability between trials not due to sampling variation is termed heterogeneity (table 1).

Heterogeneity can complicate the interpretation of a meta-analysis because the observed differences in effect between studies could also be due to trial level methodological or clinical variation in addition to the intervention effect. If a review team decides that meta-analysis is appropriate for a particular set of trials, there are two main types of statistical methods that can be used: fixed-effect and random-effects (table 1). Fixed-effect modelling is used where there is no heterogeneity or when there is thought to be a common underlying effect to the treatment. However, the absence of heterogeneity is difficult to substantiate and there is often considerable variation between acupuncture trials, meaning that it may be hard to justify the assumption of a common effect in the presence of heterogeneity (table 1). Random-effects methods can be used where there is observed or expected heterogeneity. Although a complex range of factors should be considered by reviewers as they decide whether to report the results of fixed- or random-effects modelling for a particular set of trials, it is common to see statements in systematic reviews such as “We used a random-effects model for our meta-analysis to account for the observed heterogeneity”. Such statements imply that problems caused by heterogeneity are solved by using a random-effects model. There are, however, several assumptions that need to be met for a random-effects model to be valid, and in this paper we discuss the limitations of the random-effects model in the context of the difficulty in meeting these assumptions.

The random-effects method assumes that, due to variation in design, some studies may produce large effects and others small effects; crucially, there is no assumption of a common treatment effect across studies. Statistically, study effect is adjusted for the between-study variability in the sample, meaning that, in the
A method of meta-analysis that uses within-study variances and takes no account of between-study variation. It is used where there is expected or observed homogeneity and there is an assumed common treatment effect.6 The assumption that there is no heterogeneity can be hard to establish as the results of testing for heterogeneity can depend on the number of studies in the analysis.15 Such testing is discouraged.11

In random-effects analysis16 there is no assumption of a common treatment effect, and the standard error of effect estimate for each study is adjusted for the between-study heterogeneity.19 The overall pooled estimate is an estimate of the average effect rather than of a single effect common across studies. A random-effects model is usually recommended when between-study heterogeneity is expected.6

Limits applied to the upper and lower distribution of the observed effect, indicating 95% confidence that the effect lies between those limits.25

The regression of a factor (usually called a covariate) thought to be causing heterogeneity on a random-effects meta-analysis to explore its effect on the between-study variance.3 Can help indicate whether a particular factor (clinical or methodological) is associated with treatment effects. A minimum of 10 trials are recommended for each covariate; therefore exploring heterogeneity with four factors requires 40 trials in the meta-analysis.

A key feature of a random-effects meta-analysis is that the distribution of between-study variance should be random—that is, the studies should be a random sample of the relevant distribution of treatment effects, and these study estimates should be free from bias.4 Any bias that has influenced the effect of a particular study will also be classified as between-study variation and bias that occurs systematically may potentially amplify the pooled effect.6 7

Random-effects models also give more weight to smaller studies compared with a fixed-effect calculation6 (table 1), increasing the impact of the small studies on the pooled estimate. If the results of the small studies are systematically biased, as some methodologists have observed,8 the model assumption of unbiasedness is not met and the random-effects modelling will amplify the effect of the bias.7

Clearly, as the methodology of random-effects assumes unbiasedness of estimates, the identification of studies affected by bias is critical. Assessment of bias is usually by classification of characteristics thought to influence the outcomes of a trial such as the security of the random allocation from subversion, quantity and characteristics of study drop-outs, analysis of the data in the groups to which individuals were originally allocated regardless of what they actually received, and blinding of the person collecting the outcome data to the group allocation. However, although there is consensus that bias leads to effects that are more likely to be in favour of the intervention, there is uncertainty over how much such factors actually influence the outcomes of a trial. Methodological studies have shown conflicting results in this regard, and the level of influence may vary by trial or by groups of trials.9–11 Systematic reviewers take different approaches to whether they pool data based on the results of their bias assessment. Some might decide not to conduct a meta-analysis at all, or only pool studies that have a ‘low risk of bias’, or pool only those meeting some predefined criteria such as reporting that randomisation was adequately concealed. Others might pool all studies regardless of bias assessment or conduct a sensitivity analysis on these results. Including trials that are affected by bias violates the assumption underlying the statistical procedure of random-effects, rendering it an inappropriate method. Further complications arise when bias in a trial remains undetected.

A method that can be used to explore the effect of bias—or other factors thought to be common to a subset of trials and causing heterogeneity on the pooled estimate—is meta-regression (table 1). Meta-regression can only be carried out on a random-effects meta-analysis. Here, heterogeneity caused by a particular component can be quantified and its effect on the pooled estimate observed. In practice, exploration of the causes of heterogeneity in a meta-analysis are limited by requirements of a sample numbering at least 10 trials per covariate examined and uncertainty over which factors in a complex intervention are likely to explain the variation.

To summarise, trials of complex interventions such as acupuncture are likely to give rise to heterogeneity in meta-analysis. A common method used in such cases is random-effects meta-analysis, which provides
an average of estimates across the trials. For this method to be valid, the underlying effects of the trials must be drawn from a normal distribution and trials must be free from systematic effects of bias. Interpretation of range of the average effect is recommended through the use of predictive intervals, which more accurately represents this variability.

The goal of this paper is to illustrate some of the problems facing systematic reviewers of acupuncture trials when trying to meet the assumptions of a random-effects meta-analysis by interpreting the results of a heterogeneous meta-analysis of acupuncture versus sham and attempting to examine the underlying causes of heterogeneity.

**METHODS**

**Sample inclusion and exclusion criteria**

The search strategy, methods and definitions are those used in a larger project and have been published in detail elsewhere. For this illustrative analysis we use randomised controlled trials (RCTs) for the treatment of any medical or psychological condition in adults by acupuncture published by November 2007 in English. We did not contact authors for clarification of data.

Full papers were independently screened by two people for having one or more arms that used a sham acupuncture control, which we defined as a device used to mimic an acupuncture needle and believed by the investigator to be inferior to acupuncture. We included only RCTs of musculoskeletal or neuropathic ‘chronic’ (≥2 weeks in duration) pain. We pooled trials reporting continuous pain outcomes as these were more numerous than dichotomous outcomes. Extensive study data pertaining to the description of the trial, quality criteria, methods and outcomes were extracted by one author and checked by another, resolving discrepancies by discussion.

**Meta-analysis**

A random-effects meta-analysis of the first pain outcome taken after the end of treatment was conducted using the metan module of Stata V12 (StataCorp, Texas, USA). Because the outcomes were in a range of different units, Hedge’s adjusted g was applied to calculate the standardised mean difference (SMD) for each trial. CIs for the pooled estimate were calculated using standard methods. Heterogeneity was quantified using I² and the CI calculated using a non-central χ² approach using the heterogi module of Stata.

**Meta-regression**

We specified five factors (covariates) we thought might be causing heterogeneity a priori and explored their association with effect size and heterogeneity through meta-regression. Two were methodological factors (pain location, sham needle type), two were quality factors (allocation concealment, outcome assessor blinding) and one could be considered both a methodological and a quality factor (size of the trial, defined as having provided a sample size estimate and recruiting that target sample size).

We used the iterative restricted maximum likelihood estimator for I² (t²) and a conservative estimator for the variances of the effect estimates as implemented in the metareg module in Stata. To avoid problems in model estimation caused by small sample sizes, a minimum of 10 trials were required in the meta-analysis for each covariate in the model.

The proportion of heterogeneity explained by the covariate $R^2_{\text{ADJ}}$—that is, the relative reduction in between-study variance—was presented along with the two-tailed $p$ value ($\alpha=0.05$). The proportion of residual heterogeneity after adjustment with the covariate not due to sampling variation was presented as $I^2_{\text{RES}}$. As an interpretative note, it is possible for $R^2_{\text{ADJ}}$ to be negative; this simply indicates that the proportion explained is no greater than what would be expected by chance.

**RESULTS**

Fifty-one of 85 sham-controlled RCTs identified met the inclusion criteria. Because we do not aim to provide a systematic review of effectiveness, we have anonymised the trials and have not provided details of included or excluded studies. However, our comprehensive search strategy leads us to believe that the 85 trials represent the vast majority of English language trials with our selected characteristics published prior to 2008.

**Description of the full sample**

Most trials (77%) had an outcome assessor that was blinded but 75% did not report adequate allocation concealment and only 29% reported recruiting an adequate sample (see online supplementary table A1). Real needles were the most often used sham control (35/51, 69%), with a wide variety of depths and locations reported (see online supplementary table A2).

**Meta-analysis of the full sample**

Due to the level of potential bias we would not recommend pooling these trials as the unbiasedness assumptions for random-effects meta-analysis are not met. However, we continue for illustrative purposes. There is a large variation in study effects (see online supplementary figure A1). The average effect of acupuncture is $-0.53$ (95% CI $-0.70$ to $-0.36$). Interpreting this estimate alone, we might conclude that acupuncture is superior to sham. However, the CIs do not account for the between-study variance and there is substantial heterogeneity (81%, 95% CI 76% to 85%).
**Meta-regression of the full sample**

We now apply our meta-regression using the five pre-selected covariates. None of the tested covariates explain a significant proportion of the heterogeneity (all p values are >0.05; see online supplementary table A3). The most explained was just 6.5% (adequacy of sample size), with studies inadequately powered showing an average effect of −0.72 compared with −0.23 for those adequately powered. The high $I^2_{RES}$ indicates that the heterogeneity remains largely unexplained, and there will be many other differences in the design, conduct and analysis that could explain some of the heterogeneity.

Our illustrative research question “Is acupuncture more effective than sham for treatment of pain lasting ≥2 weeks?” is really too broad to be clinically useful, and most meta-analyses would focus on a question that has tighter inclusion criteria. Table 2 lists some plausible meta-analysis questions that could be asked in systematic reviews of effectiveness and the number of studies from our sample meeting the criteria for each question.

We can see that the number of trials available to analyse very quickly reduces once we apply additional selection criteria. Even in these selected samples there is still high heterogeneity (table 2). As random-effects meta-analyses are not recommended for fewer than four studies, this underscores the difficulty of having a numerous sample of acupuncture trials on which to conduct meta-analysis. For example, the 12 trials included for Question 1 (table 2) would reduce again once we selected out trials we considered were given an inadequate ‘dose’ of acupuncture, those with a flexible protocol of treatment or those that used electroacupuncture.

For this worked example we now conduct a meta-analysis of Question 1—that is, trials that used a visually credible non-penetrating sham device.

**Table 2** Potential meta-analysis questions, numbers included and heterogeneity statistics

<table>
<thead>
<tr>
<th>Questions</th>
<th>No of trials meeting criteria</th>
<th>Heterogeneity $I^2$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Trials that used a visually credible non-penetrating sham device</td>
<td>12</td>
<td>74% (49% to 84%)</td>
</tr>
<tr>
<td>(2) Trials of back, neck or hip pain</td>
<td>15</td>
<td>85% (77% to 90%)</td>
</tr>
<tr>
<td>(3) Trials of back, neck, hip or limb pain using superficially inserted needles</td>
<td>10</td>
<td>18% (0% to 61%)</td>
</tr>
<tr>
<td>(4) Adequately concealed trials of back, neck or hip pain</td>
<td>2</td>
<td>Sample too small</td>
</tr>
</tbody>
</table>

All questions select from the pool of 51 RCTs of sham-controlled musculoskeletal or neuropathic pain in adults that reported a pain outcome on a continuous scale after the end of treatment; $I^2$ measure of heterogeneity on meta-analysis (range 0 (low) to 100 (high)).

**Description of selected sample**

Table 3 reveals at least two characteristics of concern for the validity of a random-effects meta-analysis; half the studies did not report adequate allocation concealment and half of them may be inadequately powered. Both of these characteristics have been associated with potential systematic bias. If these factors are systematically associated with study effect, then the assumptions underlying the use of random-effects meta-analysis do not hold and inferences made from the applications of such a model in these circumstances cannot be trusted. The other characteristics indicate several other possible causes of heterogeneity, and the reader should bear in mind that only a small selection of the methodological and clinical variation that could potentially contribute to heterogeneity are presented.

**Meta-analysis of selected sample**

On meta-analysis, the pooled estimate is −0.19 and the CI is −0.47 to 0.08; as these limits cross zero, we conclude that there is little evidence of an effect of acupuncture over sham in this sample of trials (see online supplementary figure A2). However, most of the trials that do not report adequate allocation concealment (except study 37) have a point estimate that favours acupuncture, as do studies with smaller weights in the analysis (except study 6). This raises our suspicions that there may be an underlying systematic effect associated with larger effects seen for smaller trials and those with inadequate allocation concealment. These would invalidate the assumptions underlying a random-effects model. Both size and allocation concealment would seem to be good candidates for further investigation;

**Table 3** Study characteristics of trials that used a visually credible non-penetrating sham device

<table>
<thead>
<tr>
<th>Allocation concealment</th>
<th>Reported concealed</th>
<th>6 (50)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not reported concealed</td>
<td>6 (50)</td>
</tr>
<tr>
<td>Outcome measure</td>
<td>Visual analogue scale (VAS)</td>
<td>9 (75)</td>
</tr>
<tr>
<td></td>
<td>Non VAS</td>
<td>2 (17)</td>
</tr>
<tr>
<td></td>
<td>Condition-specific measure</td>
<td>1 (8)</td>
</tr>
<tr>
<td>Adequacy of sample size*</td>
<td>Adequately powered</td>
<td>6 (50)</td>
</tr>
<tr>
<td></td>
<td>Inadequately powered</td>
<td>6 (50)</td>
</tr>
<tr>
<td>Location of pain</td>
<td>Back, neck or hip</td>
<td>1 (8)</td>
</tr>
<tr>
<td></td>
<td>Limbs</td>
<td>5 (42)</td>
</tr>
<tr>
<td></td>
<td>Face or head</td>
<td>4 (33)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>2 (17)</td>
</tr>
<tr>
<td>Sham point location†</td>
<td>Active location</td>
<td>9 (75)</td>
</tr>
<tr>
<td></td>
<td>Non-active location</td>
<td>1 (8)</td>
</tr>
<tr>
<td></td>
<td>Unclear</td>
<td>2 (17)</td>
</tr>
</tbody>
</table>

Numbers are N (%); percentages may not total 100 due to rounding.

*Trials were classified as adequately powered if they reported a sample size calculation and recruited this sample, all others as inadequately powered.
†As judged by the trial investigators, either away from the verum location or points considered to be irrelevant for the condition.
however, we are faced with only limited options. (1) Remove these trials from the analysis, leaving only larger trials reporting concealed allocation, but how should we define ‘small’ and ‘large’ trials? (2) Re-specify the meta-analysis to weight trials in the analysis by these criteria, knowing this method may introduce yet more bias. (3) Conduct a meta-regression to examine whether trial size and allocation concealment account for the heterogeneity, but with a sample size of 12, we should use just one covariate.

Meta-regression of selected sample
We tested the covariate of reporting allocation concealment in a meta-regression (table 4). The negative value of $R^2_{ADJ}$ indicates that this covariate does not explain any more heterogeneity than would be expected by chance. We conclude that there are other factors at work that account for the observed differences in effects between the trials, but we cannot explore them because their ratio to the number of trials is too low. We cannot rule out that the sample is biased, but we cannot further explore the cause of heterogeneity without data dredging or excluding studies because their estimates ‘look different’ to other trials with similar characteristics (here study 6 and 34).

<table>
<thead>
<tr>
<th>Covariate</th>
<th>No of studies (% weight)*</th>
<th>$R^2_{ADJ}$, $I^2_{RES}$, $p$ value</th>
<th>SMD (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No covariates</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>Concealed</td>
<td>6 (59)</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-0.19$ ($-0.47$ to $0.08$)</td>
<td></td>
</tr>
<tr>
<td>Not reported concealed</td>
<td>6 (41)</td>
<td>$-11.0, 76.1, p=0.52$</td>
<td>$-0.26$ ($-0.53$ to $0.00$)</td>
</tr>
</tbody>
</table>

*Percentage weight of that subgroup in the meta-analysis.
Reference, the category with which the other categories are compared; $R^2_{ADJ}$ proportion of heterogeneity explained by the covariate; $I^2_{RES}$ proportion of residual heterogeneity after adjustment with the covariate; SMD, standardised mean difference.

As in other areas, systematic reviewers wishing to estimate any underlying effect by pooling acupuncture trials are hamstrung in their efforts by the complexity of the intervention and the relatively small number of trials available. Pooling just a few trials rules out exploring the causes of heterogeneity. Pre-specifying which trials are to be pooled is preferable, but random-effects meta-analysis is not advised for fewer than four trials and assessment of heterogeneity is impossible in a small sample. To further add to the difficulty of assessing whether a trial or group of trials is ‘biased’, various dimensions of bias appear to have different effects in individual trials and across different systematic reviews.

CONCLUSIONS
Authors of systematic reviews of acupuncture should apply caution before using meta-analysis; in particular, the potential effect of bias should be seriously considered. The exploration of heterogeneity using meta-regression is unlikely to explain the variation present. Caution should be applied when interpreting CIs because they do not reflect the variation between trials.

**Summary points**
- Acupuncture studies are often heterogeneous, so systematic reviewers often choose the random effects model for meta-analysis.
- Random effects model rests on certain assumptions, particularly lack of systematic bias.
- In a sample of typical acupuncture study reports, we found these assumptions were not met.

**Contributors** SLP designed and analysed the study, which was advised upon by JB, SC and HM. JB also advised on and contributed to the screening and general data extraction. The authors thank Gillian Worthy, Alison Longridge, Laura Vanderbloemen and Ann Hopton who contributed to the screening and general data extraction.

**Competing interests** None.
Funding The Foundation for Traditional Chinese Medicine financially contributed towards the project.

Patient consent Obtained.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data are available on request.

REFERENCES
Problems caused by heterogeneity in meta-analysis: a case study of acupuncture trials
Stephanie L Prady, Jane Burch, Simon Crouch and Hugh MacPherson

Acupunct Med  published online October 18, 2013

Updated information and services can be found at:
http://aim.bmj.com/content/early/2013/10/18/acupmed-2013-010364

These include:

Supplementary Material
Supplementary material can be found at:
http://aim.bmj.com/content/suppl/2013/10/18/acupmed-2013-010364.DC1

References
This article cites 20 articles, 5 of which you can access for free at:
http://aim.bmj.com/content/early/2013/10/18/acupmed-2013-010364#ref-list-1

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://www.bmj.com/company/products-services/rights-and-licensing/

To order reprints go to:
http://journals.bmj.com/content/subscribers

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/