Systematic review of systematic reviews of acupuncture published 1996–2005

CJ Derry, S Derry, HJ McQuay and RA Moore

ABSTRACT – Systematic reviews of acupuncture have tended to support its use, but few applied rigorous inclusion criteria. We tested the credibility of conclusions of systematic reviews of acupuncture published since 1996 by applying rigorous inclusion criteria. Reinterpretation used randomised and double blind trials with valid outcomes or design, and with information available from at least four trials or from 200 patients. Qualified support for acupuncture was originally reported in 12 out of 35 systematic reviews, and strong support was found in another six. Applying stricter inclusion criteria, however, showed that none of the 35 reviews supported acupuncture, predominantly because there were too few patients in the randomised, double blind studies. Six reviews with more than 200 patients in randomised, double blind studies had good evidence of no benefit. Systematic reviews of acupuncture have overstated effectiveness by including studies likely to be biased. They provide no robust evidence that acupuncture works for any indication.

KEY WORDS: acupuncture, bias, systematic review

Introduction

The place of acupuncture and other complementary therapies in mainstream healthcare is controversial. Systematic reviews and meta-analyses have claimed that these therapies can be distinguished from controls, and many people are turning to alternative therapies, including acupuncture. One in five European adults have chronic pain, and 13% of them use or have used acupuncture. A Department of Health publication for primary care clinicians in 2000 claimed that there was reasonable evidence of efficacy for acupuncture in chronic lower back pain.

Unfortunately many systematic reviews and meta-analyses of complementary therapies like acupuncture have included clinical trials whose design is not randomised or blinded. Reanalysis of a systematic review of homeopathy found no convincing evidence for efficacy in any condition.

Trials of acupuncture are frequently small, use outcomes of little practical relevance, or are invalid for some reason (short duration trials in chronic conditions, for example). These factors all affect the credibility of the findings. Our aim was to examine systematic reviews of acupuncture published in the last decade, and to compare the authors’ original conclusions with our own after applying criteria of quality, validity and size.

The aim was not to review the efficacy of acupuncture, but rather the efficacy of systematic reviews to accurately assess the evidence for acupuncture, using criteria for inclusion of trials that are known to minimise the possibility of bias, and which are used commonly in systematic reviews of most other medical interventions.

Methods

We searched PubMed, the Allied and Complementary Medicine Database (AMED), and the Cochrane Library for systematic reviews of acupuncture for any...
condition in humans, published from January 1996 to August 2005, using the terms 'acupuncture' and 'systematic OR meta-analysis'. We also looked for relevant reviews in our own in-house databases and reference lists of retrieved articles.

We accepted reviews published in English that examined the efficacy of traditional Chinese or mechanical acupuncture, electro-acupuncture, laser acupuncture or acupressure, electrical nerve stimulation but not transcutaneous (TENS) or dry needling (using empty hypodermic needles or acupuncture needles at trigger points for myofascial pain). Where one review clearly updated a previous review, only the most recent publication was used. If more than one review covered the same trials for the same outcome and indication, the most recent was taken. Reviews of adverse effects of acupuncture were not included.9

Two reviewers extracted the following information from each review:

- number of studies included
- number of patients included
- whether there was assessment of quality for included studies
- whether exclusions due to poor quality were made, or a sensitivity analysis presented
- main findings, including whether a pooled analysis was done
- original authors’ conclusions on efficacy
- original authors’ conclusions on strength of evidence (based on their assessment of whether there was no evidence of benefit, qualified support, or strong support)
- authors’ affiliation to complementary medicine departments.

In addition, we made our own assessment of the strength of the evidence presented in each review. We specified a priori criteria for quality, validity and size to remove studies that were susceptible to bias, and might leave uncertainty in the results. To be sure that any beneficial or harmful effect of acupuncture is not an artefact of biased design, several factors are important. Trials need to:

- be randomised and double blind (both patient and assessor blind to intervention), with appropriate controls. Trials that are not properly randomised and double blind have been shown to overestimate treatment effects. For trials of acupuncture, the practitioner cannot be blinded to the treatment groups. The most appropriate design uses sham acupuncture and outcome assessors blinded to treatment group. During ‘sham acupuncture’, needles are inserted at non-specific acupuncture points and usually penetrate the skin only a few millimetres, or needles are used that only indent the skin. Sham acupuncture involves all the various aspects of acupuncture, including practitioner’s time and attention, and is not equivalent to an inert placebo. Waiting list controls, where patients remain on a waiting list, with no treatment and no practitioner’s time or attention, cannot be blinded to their treatment group; they have no treatment of any sort.
- study patients with recognised clinical conditions (not for example experimental pain), have groups comparable at baseline, and use relevant outcomes (eg patient pain scoring, number of headache-free days, long-term outcomes for chronic conditions).
- study sufficient numbers of patients to minimise effects of chance. We know that small studies, or large studies with small numbers of events, can be affected by the random play of chance and mislead results.9,14 How much information is needed to be sure of a result depends on how sure one wants to be, and how large is the effect. For large effects (50% absolute risk increase, for instance) we need about 400 patients or 200 events;11 where the absolute risk increase is small, at below 20%, the number of patients needed in trials rises to the thousands. To set some sensible lower limit for numbers of patients involved in trials of sufficient quality and validity to make it worth trying to calculate statistical significance, we arbitrarily specified four trials and/or 200 patients as a minimum.

Relative benefit (or risk) was calculated with 95% confidence intervals using a random effects model,15 with no statistically significant difference between treatments assumed when the 95% confidence intervals included unity.

Results

We found 55 systematic reviews satisfying our inclusion criteria, of which 20 were excluded because a more recent review covered the same topic and included the earlier studies (n=17), or because they were not in English (n=3). Full details of the reviews, authors’ and reviewers’ conclusions, and references for included and excluded studies are presented in two supplementary files available from the authors upon request. The 35 included studies16-49 examined the use of acupuncture in various painful conditions (n=18), stroke (n=2), nausea and vomiting (n=2), and other conditions including insomnia, smoking cessation, weight loss, and asthma (n=11).

All but four of the 35 reviews made a formal assessment of methodological quality of included studies. Most of the reviews (22/35) claimed to use only randomised studies, and most had included trials that were not both patient and assessor blind. No review excluded trials from analysis because of low quality, though a small number performed sensitivity analysis according to study quality.

Most reviews (24/35) had information on fewer than 1,000 patients (Fig 1). However, the number of patients contributing to analysis of efficacy was often substantially smaller than the total number of patients in all the trials included in a review, because many reviews included studies with no relevant efficacy results.

Most reviews commented upon the relevance of the chosen outcomes and the validity of trials. Some were unclear about what effect they were reporting, and some reported inappropriate outcomes, especially short-term outcomes for chronic conditions. Waiting list controls were sometimes used.

Of the 35 reviews, 17 concluded that there was either no evidence of benefit, or evidence of no benefit. Twelve had a qualified
conclusion of some benefit for acupuncture, with authors commenting on issues around small trial size or poor methodology, or the need for further research, but still using words, usually in the abstract or conclusion, supporting the use of acupuncture. For instance, a Cochrane review of acupuncture for idiopathic headache maintained, ‘the existing evidence supports the value of acupuncture for the treatment of idiopathic headaches. However, the quality and amount of evidence are not fully convincing’. A second commented that acupuncture ‘may be beneficial to reduce symptomatic knee pain’ though ‘reviewers concluded that the poor quality of the trials, including the small sample size preclude [sic] its recommendation’.

Of the 35 reviews, six had an authors’ strong conclusion of benefit (Table 1) meaning that authors made comments like ‘acupuncture effectively relieves chronic low back pain’, supports the use of P6 acupoint stimulation in patients without antiemetic prophylaxis, or ‘there is strong evidence suggesting that acupuncture is effective in the short term for lateral epicondyle pain’. Of the 18 reviews with qualified or strong support for acupuncture, ten came from departments connected with complementary therapy. Five of the six studies with strong support were from departments connected with complementary therapy. Of the 17 reviews showing no benefit, eight were affiliated to departments of complementary medicine. Cochrane reviews were less likely to support acupuncture, though one gave strong support and three qualified support (Table 1).

Our assessment was that none of these 35 systematic reviews could demonstrate robust evidence of effectiveness for acupuncture when strict criteria of quality, validity, and size were used to judge the evidence. Using criteria known to reduce the possibility of bias, commonly used in systematic reviews assessing medical interventions, most acupuncture reviews had trivial amounts of good quality evidence. Only six had more than 200 patients in randomised, double blind trials, and in these acupuncture was not significantly better than control.

The authors of six reviews (not the same six) made strong claims of benefit that were not upheld in our evaluation of quality, size and validity, and it is useful to examine these six in more detail:

- Ernst and White included nine randomised trials in a review of acupuncture for back pain, and provided short-term outcomes. Five of these trials were not blind, and had a statistically significant benefit (relative benefit 1.8; 95% confidence interval 1.3 to 2.4). The four blind studies (n=173) showed no significant benefit (relative benefit 1.2; 0.9 to 1.5). All of the benefit of acupuncture reported in the review derived from the non-blind studies.

- Ezzo et al included seven randomised trials of acupuncture for knee osteoarthritis, of which three were high quality trials using sham acupuncture (n=174). Of these, only one trial (103 patients) showed consistent benefit for all short-term pain outcomes, with no benefit for function outcomes. Two trials reported longer-term (3 months) findings but with conflicting results.

- Trinh et al included six studies to conclude that there was strong evidence to support acupuncture for lateral epicondyle pain. Of these six trials, one was not properly randomised. Of the remaining five trials, two were not double blind. Of the remaining three (n=175), one had results only immediately after treatment. That left two randomised, double blind trials, reporting valid outcomes at two or three months after treatment. Both compared real acupuncture with sham acupuncture, with a pooled relative benefit of 1.2 (0.96 to 1.6), indicating no benefit.

- Lee and Ernst included six studies of patients undergoing endoscopy, only two of which were properly randomised and blinded (n=120). There was no meta-analysis because of different outcome measures. Some measures of discomfort were reduced in the acupuncture groups, and in one trial additional sedative use was also reduced (n=10), while in the other trial pain in some areas was reduced (n=45).

- Mannheimer included 33 randomised trials of acupuncture in low back pain, only four of which (n=343) had sham

Table 1. Support by original authors by type of review and by affiliation to department of complementary therapy.

<table>
<thead>
<tr>
<th>Authors’ support for acupuncture</th>
<th>None</th>
<th>Qualified</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>All studies (n=35)</td>
<td>17</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Cochrane (n=12)</td>
<td>8</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Affiliated (n=18)</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Not-affiliated (n=17)</td>
<td>9</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
Different pain outcomes were pooled, and in four trials short-term outcomes (about three weeks) were significantly improved for true, compared with sham, acupuncture. Longer-term outcomes were not significantly different in the two trials reporting them.

- The largest number of patients studied was for the use of P6 acupoint stimulation for preventing postoperative nausea and vomiting. The original analysis used randomisation as the only quality criterion. We performed a sensitivity analysis to investigate the effect of criteria of quality, validity and size on the strength of evidence for the different outcomes (Table 2). The use of increasingly stringent criteria reduced or eliminated statistical significance of benefit for nausea, vomiting, and antiemetic consumption.

**Discussion**

The 35 systematic reviews of acupuncture published since 1995 represent what should be the highest level of evidence available. Unfortunately, most of the reviews were based on a few small trials of inadequate design and statistical power. Many reviews included studies with designs known to be associated with bias and overestimation of treatment effects, notably trials that were not randomised, not blind, or neither randomised nor blind. Pooled analysis of trials with flawed design does not resolve, but rather accentuates, these problems.

In no case did reanalysis using only trials that fulfilled minimal criteria (randomisation, blinding, size and validity) support a strong conclusion of benefit from acupuncture. Perceived benefits of acupuncture were derived from low quality trials likely to be biased. The best statistical claim was for relief of back pain after three months, and then with limited information and from pooling different pain outcomes. The example with most information was acupressure for postoperative nausea and vomiting (Table 2). Sequential elimination of non-blind trials, small trials, and trials insensitive because of low event rates excluded most trials, leaving a conclusion of no effect for acupressure in the remainder, irrespective of how statistical significance was calculated.

**Table 2. Post-operative nausea and vomiting, with application of increasingly more stringent criteria for quality, size and validity.**

<table>
<thead>
<tr>
<th>Sensitivity analysis</th>
<th>Nausea</th>
<th>Vomiting</th>
<th>Antiemetic consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trials/patients</td>
<td>Relative risk (95% CI)</td>
<td>Trials/patients</td>
</tr>
<tr>
<td>Randomised</td>
<td>16/1,826</td>
<td>0.73 (0.57 to 0.93)</td>
<td>20/2,187</td>
</tr>
<tr>
<td>Randomised and blind</td>
<td>10/1,150</td>
<td>0.78 (0.58 to 1.05)</td>
<td>12/1,328</td>
</tr>
<tr>
<td>Randomised, blind, over 100 pts</td>
<td>5/885</td>
<td>0.82 (0.58 to 1.17)</td>
<td>6/988</td>
</tr>
<tr>
<td>Randomised, blind, over 100 pts,</td>
<td>5/885</td>
<td>0.82 (0.58 to 1.17)</td>
<td>4/567</td>
</tr>
<tr>
<td>control event rate ≥20%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Control event rate is the percentage of patients who had an emetic event (nausea, vomiting, or use of antiemetic medication) in the control group receiving placebo. Low control event rates limit the sensitivity of trials in antiemesis.

**Key Points**

- Systematic reviews have tended to support the use of acupuncture
- They have tended to use trials with known sources of bias
- Excluding trials with known sources of bias, no systematic review had evidence of efficacy
- There is no robust evidence from systematic reviews that acupuncture works in any indication

Cochrane reviews were less likely to support acupuncture than other reviews. Reviewers without affiliation to a department of complementary medicine were less likely to support acupuncture than those that had such affiliation. The sample in this review of reviews is probably too small to draw any conclusions from this.

Several larger well-reported randomised trials of acupuncture have been published since these reviews. Those in fibromyalgia, chemotherapy-induced nausea and vomiting, breech presentation, tension headache and migraine, have all been negative compared with sham acupuncture controls. One in osteoarthritis of the knee, had statistical improvement over sham acupuncture at three months, but not later. The large trials and this review of reviews come to the same general conclusion; that over a whole range of conditions and outcomes acupuncture cannot yet be shown to be effective.

This negative view of acupuncture after a decade of primary and secondary research may not be wholly justified. For instance, a commentary on the trial of acupuncture for knee arthritis suggested that longer term physical functioning differences might be a more important outcome than pain, because that and other trials showed effects (but not pain relief) lasting well beyond the use of acupuncture, compared with sham acupuncture. It may be that we have to look at different outcomes. Again, trials that have included both waiting list and sham acupuncture control groups tend consistently to show benefit for acupuncture and sham acupuncture over waiting list, suggesting that needles convey only a part of the benefits.
Future studies might usefully investigate which aspects of the ‘acupuncture experience’ could give rise to this observed benefit. The possibility of some small but clinically useful benefit cannot be excluded on the basis of the evidence to hand.

It has been argued\(^{14}\) that most published research findings are false, and that confirmatory meta-analyses of good quality randomised trials provide the least likelihood of being wrong. This review of reviews supports this, at least for acupuncture, where the effort seems to have been to find statistical benefits, irrespective of quality. The trouble is that statistical significance does not necessarily translate into clinical benefit.

The lack of evidence makes for problems for those providing acupuncture services, and for regulators. It is also a problem for purchasers of healthcare. Private individuals can please themselves. Public or private bodies that have previously purchased acupuncture may have to reconsider. For acupuncture and other alternative therapies, there has been a climate of permissive endorsement made on the basis of perceived low risk rather than evidence of efficacy. At what point will this view change in the face of mounting evidence of lack of evidence of efficacy?

Acknowledgement

The study was supported by the Oxford Pain Relief Trust, which had no involvement in any aspect of the original idea or the design, execution or writing of the review, nor in the decision to submit the paper for publication.

References


