

Assessing Clinical Efficacy of Acupuncture: What Has Been Learned from Systematic Reviews of Acupuncture?

J. Ezzo · L. Lao · B. M. Berman

7.1

Introduction

This chapter summarizes the evidence of the clinical efficacy of acupuncture. Treatment efficacy pertains to the differential effect observed from a treatment as compared to a placebo or another treatment using a rigorous methodological design. Efficacy is often assessed using a double blind controlled trial design [1]. While some chapters in this book focus primarily on animal studies, the question of clinical efficacy can only be addressed in human studies. Therefore, this chapter will focus on summarizing the evidence from human acupuncture trials. It is divided into two major sections. The first presents an overview of evidence-based medicine and explains why the systematic review is the most reliable and valid method for assessing efficacy. The second section summarizes the current evidence from systematic reviews on acupuncture for specific conditions and, when applicable, presents results of trials published subsequently to each systematic review.

7.2

The Role of Systematic Reviews in Evidence-Based Medicine

Although there are many experimental designs in research, the gold standards for assessing efficacy are large, well-designed, randomized controlled trials (RCTs) and systematic reviews (Table 1) [2], because these designs are less likely to mislead than other study designs [3]. To date, trials of acupuncture are notably small [4–6], and few fulfill the description of a large RCT. Therefore, the most comprehensive assessment of acupuncture efficacy will be found in the existing systematic reviews on the topic.

Until about the mid-1990s, the terms “systematic review” and “meta-analysis” were used interchangeably. That is why the review by ter Reit and colleagues [4] is termed a “meta-analysis.” More recently, these terms have been clarified: A systematic review is the whole process of conducting a comprehensive evidence synthesis from search strategy, evaluation of trial quality, and analysis, whereas a meta-analysis is a specific statistical technique of pooling data. By today’s definition the ter Reit chronic pain review is actually a systematic review without a meta-analysis. Examples of acupuncture systematic reviews which have used meta-analysis are the reviews of acupuncture for chronic pain by Patel and colleagues [30] and on acupuncture for smoking cessation by White and Rampes [31].

Table 1. Type and strength of efficacy evidence [2]

1. Strong evidence from at least one systematic review of multiple well-designed randomized controlled trials
2. Strong evidence from at least one properly designed randomized controlled trial of appropriate size
3. Evidence from well designed trials without randomization, single group pre-post, cohort, time series or matched case-controlled studies
4. Evidence from well designed nonexperimental studies from more than one center or research group
5. Opinions of respected authorities, based on clinical evidence, descriptive studies or reports of expert committees

Evidence-based medicine is the “conscientious, explicit, and judicious use of the current best evidence in making decisions about the care of individual patients” [2]. This requires that clinicians combine their best clinical wisdom and experience with the best systematically researched clinical evidence available. It is not, as some have suggested [7], a reliance on abstract clinical evidence to the exclusion of clinicians’ experiences and common sense. In essence, by applying their best clinical experience while keeping abreast of new research, clinicians have always striven to practice evidence-based medicine, which is why some have called it “old wine with a new label” [8].

What is new about evidence-based medicine is the systematic review method, which is an explicit methodology designed to summarize and synthesize evidence [3]. In the past, traditional literature reviews, commonly called narrative reviews, have been used to summarize evidence. One major difference between narrative reviews and systematic reviews is that systematic review methodology requires all trials which meet specified criteria to be included in the review, regardless of the trial results. By contrast, narrative reviews may exclude trials with findings that did not match the writer’s point of view or had simply been inadvertently haphazard and noncomprehensive and therefore prone to bias [9]. Experience has demonstrated that a treatment which appears promising according to a narrative review can appear less promising when the systematic review methodology is applied [10]. Because systematic reviews follow an explicit methodology designed to minimize bias, they are becoming increasingly preferred over narrative reviews. For example, policy makers are growing increasingly partial to them as a way of summarizing evidence [11]. And, in the midst of the information explosion, clinicians rank reviews as their most preferred source of new information [12] while ranking individual RCTs close to the bottom of the list. Consumers, too, can use reviews to guide health decisions [13].

Finally, because systematic reviews comprehensively summarize research already completed, they are valuable tools for guiding the design of proposed clinical research. Systematic reviews can prevent the inadvertent duplication of work already done, underscore where knowledge gaps exist, and generate hypotheses for future research. For these reasons, trialists advocate conducting a systematic review and meta-analysis in the planning stages of a trial [14, 15], and funding agencies are increasingly inquiring on grant applications whether a systematic review has been conducted on the topic for which funding is sought [16].

7.2.1

Validity of the Study Design: How Systematic Reviews Limit Bias

It is important to understand the systematic review process in order to appreciate how this type of data analysis can limit bias. McQuay and Moore [2] describe the following as key steps of any systematic review: finding, appraising, and combining the evidence. Bias, a systematic error which can make a treatment appear less or more effective than it really is, can occur in any stage of this process. Although the examples presented here pertain to acupuncture, the potential biases discussed below apply to all systematic reviews for any treatment of any condition.

7.2.1.1

Finding the Evidence

One of the most striking ways that systematic reviews differ from narrative summaries is the comprehensiveness of the search for relevant trials. A trials search should be both comprehensive and replicable. MEDLINE is the most obvious place to begin searching for relevant acupuncture trials, but a search should not stop there. MEDLINE sensitivity, which is the proportion of relevant RCTs found by MEDLINE divided by the total of known trials on the topic, is estimated to be only about 58% for acupuncture [17]. Obviously, this suggests that a universe of acupuncture trials exist that are not listed in MEDLINE. Additional sources can include other electronic databases [18], articles in journals not indexed by any database, conference abstracts that never become fully published articles, and trials which are completed but remain in researchers’ file drawer or diskette and never reach print at all [19].

Publication bias is “the tendency of investigators, reviewers, and editors to submit or accept manuscripts for publication differentially, based on the direction or strength of the findings” [20]. Therefore, articles which are the most easily accessible because they are published may not be representative of the results of all the trials done. To the extent that publication bias exists, systematic reviews that omit unpublished studies risk overestimating treatment effects. Bias has been demonstrated both in what investigators submit to journals [21] and in what editors accept [22].

The recent findings by Egger and colleagues [23] raise the question of whether trials published in English are representative of the available trials. They observed that authors publishing in both German and English tend to report nonsignificant findings in German and significant findings in English. This suggests that one must search not only beyond the published literature, but also beyond the English language.

7.2.1.2

Appraising the Evidence

It may seem obvious that evidence is only as good as the trials upon which it is based, but the concept of assessing the methodological quality of the trials included in a systematic review is a relatively recent practice. A validated, user friendly scale (Jadad scale) for assessing methodological quality did not exist until 1996 (Table 2) [24]. Therefore, many reviews exist which form conclusions about the efficacy of a treat-

Table 2. Validated Jadad scale for assessing trial quality [24]

1. Was the study described as randomized?
2. Was the randomization scheme described and appropriate?
3. Was the study described as double blind?
4. Was the method of double blinding appropriate?
5. Was there a description of drop outs and withdrawals ?

ment without taking into consideration the quality of the individual trials. The Jadad scale asks two questions pertaining to double blinding (defined as blinding both the patients and the outcomes assessor). Because it is not possible to blind the patients in most acupuncture trials, some reviewers have used a slightly different scale [6] which asks two separate questions: (1) Were the patients blinded? and (2) Was the outcomes assessor blinded? However, even if double blinding did not occur, it is still possible for a trial to earn a high quality score on the Jadad scale if the trial quality was satisfactory in other areas.

When trial quality is taken into account, results show that all RCTs are not created equal and, in general, low quality RCTs can bias trial results by exaggerating the positive effects of a treatment [25, 26]. Relative to acupuncture in particular, low quality RCTs have been associated with results favoring acupuncture for chronic pain [4, 27], asthma [5], and tobacco addiction [28].

Previously, many also assumed that RCTs not published in English were likely to be inferior methodologically compared to those in English, and therefore non-English language studies were often omitted from systematic reviews. Recently, Moher and colleagues [29] have demonstrated that the quality of studies in French, German, Spanish, and Italian is comparable with those published in English and, therefore, excluding these trials from a systematic review cannot be justified.

7.2.1.3

Combining the Evidence

Meta-analysis, the statistical pooling of results, is a preferred way to combine data from several trials when the data permit [2]. When the data are not conducive to pooling, an alternative method known as best evidence synthesis [32], which gives more weight to the high quality trials and less weight to low quality trials, is preferred over a simple one-trial, one-vote method. The latter method is discouraged because low quality trials should not be given the same weight as high quality trials. The reviews on acupuncture for low back pain by Tulder and colleagues [33] and for chronic pain by Berman and colleagues [27], demonstrate this approach.

As has already been noted, bias can occur if high quality studies are combined with low quality studies. One way to assess whether low quality studies may be biasing results is to perform a sensitivity analysis. This method combines the results of both high and low quality trials and compares these results to those of the high quality trials alone. If the combined low and high results appear more optimistic than the high quality results alone, then bias may be at work.

Bias can also occur if the reviewers inadvertently count the same trial twice (duplicate bias). Duplicate bias occurs when two articles appear sufficiently different so that the reviewers fail to realize that they actually represent the same trial. If the duplicate trial has a positive result, it biases results towards the positive side, whereas, if the

trial has a negative or nonsignificant result, it biases results the other way. Duplicate bias is no small problem in systematic reviews and can exaggerate a treatment effect by as much as 20 % [34].

7.2.2

Validity of the Acupuncture Procedure:

Assessing the Quality of the Acupuncture Treatment

So far, most of the discussion about trial validity has concerned how the methodological quality of trial design can bias trial outcome. However, inadequate acupuncture treatment can also jeopardize trial validity and bias outcome. Basing conclusions about acupuncture efficacy on suboptimal or inadequate acupuncture procedures is analogous to pharmaceutical trials formulating conclusions about the efficacy of drugs based on inadequate dose. While some conditions such as nausea and vomiting have a straightforward, commonly accepted acupuncture procedure (i.e., the stimulation of P6), other conditions such as chronic pain have tremendously different treatment approaches. Therefore, the question of acupuncture treatment adequacy is more urgent for some conditions than for others. Although it is an issue to be addressed and defined in primary studies (e.g., cohort studies or RCTs), systematic reviews must have a method for assessing whether treatments meet basic criteria. Three methods have been used to assess the adequacy of acupuncture procedures retrospectively.

One way that reviewers have retrospectively assessed acupuncture treatment adequacy was first proposed by Linde et al. [6]. An acupuncturist was presented with the inclusion criteria and acupuncture methods in the papers, blinded to trial results, and asked to rate whether the acupuncture treatment was adequate to treat the condition under investigation based on five aspects: (1) points selected, (2) total number of treatments, (3) number of times per week the patient was treated, (4) duration of each session, and (5) whether or not “de qi” was elicited.

A second way of assessing acupuncture treatment quality retrospectively is to specify a minimally acceptable treatment. This approach, used by Molsberger and Bowing [35], defined a minimally adequate acupuncture treatment as consisting of at least ten total treatments of at least 15 minutes each and a description of the points used. Only 16 of 88 studies on musculoskeletal and/or neurological conditions met the minimal criteria and, of those, only two were controlled trials. What was learned from this method is that most studies do not meet the minimal criteria of an acceptable acupuncture procedure.

A third, less efficacious approach to quantifying acupuncture procedure is to identify those aspects of the acupuncture procedure that are associated with positive outcomes. Although this cannot answer the question of whether the treatment is adequate, it can provide important information for guiding future clinical trials. Patel and colleagues [30] used this approach with their pooled data. They noted that individualized treatments significantly favored acupuncture, whereas formulaic approaches, in which all the patients received the same treatment, fared less positively. They proposed that individualized acupuncture treatments might produce better outcomes than cookbook approaches but that these findings could be confounded by the methodological quality of the trials.

This third approach was also used in a recent acupuncture and chronic pain review [27]. The specific aspects of the acupuncture treatment which were hypothesized as associated with positive outcomes were based on findings reported in a doctoral thesis [36]. The thesis analyzed the acupuncture procedures in textbooks from China, Japan, and Korea seeking recommendations which could be generalized for treating certain pain conditions. No generalized recommendation could be made on the specific acupuncture points, because they varied so much from textbook to textbook. On the other hand, generalizations could be made pertaining to the recommended number of needles used and the number of total treatments administered: 6 points per treatment were probably adequate but 10 were better, and six total treatments were probably adequate but ten were better. When these hypotheses were tested in the chronic pain review [27], no association was observed between the number of points used and positive outcome, but a statistically significant association ($p < 0.05$) was found between the total number of treatments given and the outcome, even when controlling for methodological quality of the trials. Interestingly, virtually no trial administering fewer than six acupuncture treatments achieved a positive outcome.

Although these significant findings merely show an association and not a causal relationship between six or more treatments and positive outcome, like Patel's observation about individualized treatments, they may be able to guide future studies. Given the paucity of information as to what constitutes an adequate acupuncture treatment, it is advisable to compare the effectiveness of various acupuncture protocols in a small pilot study to ascertain the preferred procedure before conducting a larger RCT. One common problem in assessing an acupuncture procedure retrospectively is that very few trials present sufficient information about the acupuncture procedure to assess its adequacy. For example, data on the various aspects of the acupuncture treatment were extracted for one chronic pain review [27] but found to be insufficient for assessment. The problem of missing information was also encountered by ter Reit and colleagues [4] when they attempted to use a proxy measure for treatment adequacy by assessing whether the papers reported on the extent of training and experience of the acupuncturist. What has been learned from this method is that reporting on the details of the acupuncture procedure needs to be improved in trials.

7.2.3

Summary

In conclusion, summarizing evidence can be likened to Aesop's story of the blind men and the elephant:

Each of five blind men is asked to touch an elephant and describe what the animal looks like. One blind man feels the tail and claims the animal is long and thin like a rope. Another, feeling the elephant's leg, insists that the animal is round, thick, and solid like a tree trunk. And so on.

As children hearing this story, we were amused at the obvious foolishness of each man forming an opinion of the elephant based on one small study. Now, as researchers and clinicians, we are challenged to transcend the mistaken assumption that one small study represents the whole reality. Small RCTs, which are typical of current acu-

Table 3. Eight reasons to use systematic reviews [9]

1. Large quantities of information must be reduced to palatable pieces for digestion
2. Various decision makers need to integrate the critical pieces of available biomedical information
3. It is an efficient scientific technique
4. The generalisability of scientific findings can be established
5. The consistency of relationships can be assessed
6. Inconsistencies and conflicts in data can be examined
7. When meta-analysis is performed, they increase power and precision
8. They offer an improved reflection of reality over traditional reviews

puncture trials, seldom answer efficacy questions definitively. A trial may lack sufficient power to detect a difference that exists in reality (known as a type II or false negative error), it may be biased, or the results may be applicable only to a certain group of people or patients but not to all. There are many advantages to using the systematic reviews method (Table 3). In short, this offers the most comprehensive way to describe "the elephant."

7.3

A Summary of the Results of Systematic Reviews of Acupuncture

This section summarizes the results of systematic acupuncture reviews. The first part of this section will summarize the systematic reviews which have been done for painful conditions. The remaining part will summarize the results of systematic reviews for other conditions, such as acupuncture for nausea and vomiting or for asthma (Table 4).

7.3.1

Acupuncture for Painful Conditions

7.3.1.1

Is Acupuncture Effective for Chronic Pain?

Patel and colleagues [30] published the first systematic review on chronic pain, observing that few of the 14 acupuncture trials included had statistically significant results. However, the pooled results favored acupuncture. Furthermore, pooled subgroup analyses by site of pain significantly favored acupuncture for low back pain and head/neck pain. However, statistical pooling produced another interesting finding: when trials with some degree of blinding (of patients and/or outcomes assessors) were compared to trials with no blinding, the unblinded trials significantly favored acupuncture, whereas the blinded trials did not. One interpretation for this is that lack of blinding may have biased the results to favor acupuncture. Other methodological deficiencies were also noted, and the more methodologically rigorous trials tended to yield less favorable results. Because of this, the authors stated that, although the pooled results favoring acupuncture were statistically significant, the various sources of bias prevented this finding from being a definitive conclusion.

Ter Reit and colleagues [4] published a more extensive systematic review on chronic pain the following year, including an extensive search strategy and a way of

