

Assessing Clinical Efficacy of Acupuncture: Considerations for Designing Future Acupuncture Trials

L. Lao · J. Ezzo · B. M. Berman · R. Hammerschlag

11.1

Introduction

The recent interest in acupuncture has led to an unprecedented amount of research funding in this area. This chapter presents a systematic, stepwise approach to research. It proposes investigating the efficacy of acupuncture by adapting the U. S. Food and Drug Administration (FDA) approach to clinical trials (Table 1). Two examples, one for chronic osteoarthritic (OA) pain [1, 2] and one for acute postextraction dental pain [3–5] are used to demonstrate how this method can be applied to acupuncture research. The final section discusses ways to integrate Western medicine more closely with traditional Chinese medicine (TCM) in research.

11.2

Phase I

According to the FDA description, phase I trials are small, typically do not have a control group, and are designed to establish basic parameters of dose and safety (Table 1). The phase I OA trial [2], for example, consisted of 12 patients with no control group. Although acupuncture has already been recognized as generally safe, establishing an adequate dose (i.e., a specific acupuncture protocol) in a phase I trial is a major challenge. This has seldom been done in past trials. Consequently, a number of randomized controlled trials (RCTs) of acupuncture have administered acupuncture treatments not based on preliminary data. The following key items should be considered in a phase I acupuncture trial: dose, primary and secondary outcomes, follow-up time, and safety.

Table 1. Features and types of clinical trials according to FDA nomenclature

Type	Key features
Phase I	Initial human studies; provides preliminary information on safety, appropriate dose, and chemical action; typically does not include a control group
Phase II	Larger than phase I but still generally small numbers of patients; provides preliminary information on efficacy; provides additional information on dose and safety; usually includes a control group; may randomize
Phase III	Large sample size based on phase II results; assesses dose effects; evaluates term safety and efficacy; always controlled and randomized; treatment must demonstrate favorable risk/benefit ratio

11.2.1

Establishing an Adequate Dose

Establishing an adequate treatment/dose of acupuncture is one of the fundamental first steps of a trial. In the past, many acupuncture treatments used in trials have been shown to be inadequate according to minimal criteria [6 and see Chapt. 10]. The fundamental decisions a researcher must make when establishing an adequate acupuncture treatment pertain to:

1. Type of acupuncture treatment (e.g., TCM, Japanese, five-element)
2. Formulaic vs. individualized treatments
3. Qualifications/experience of the acupuncturist
4. Point selection
5. Depth and techniques of needle manipulation, i.e., manual (reinforced or reducing technique) or electrical stimulation (frequency and intensity)
6. Whether de qi will be pursued
7. Duration of each treatment
8. Number of treatments per week and spacing between treatments
9. Total number of treatments.

11.2.2

Formulaic vs. Individualized Acupuncture

The ongoing debate about whether to use formulaic or individualized acupuncture revolves around two points of view. Those who argue for individualized treatment suggest that this approach reflects clinical reality whereas formulaic approaches do not. Those who argue for formulaic approaches suggest that individualized approaches may compromise methodological rigor and that a standardized treatment is less likely to create unintended confounders or noise in the study. There is no golden key that fits all the locks. Both methods are useful if the design is appropriate. Part of the decision depends on the condition being treated. Homogeneous populations are more conducive to formulaic acupuncture than are heterogeneous populations. For example, formulaic acupuncture was used in the postextraction dental pain study [3–5]. All the patients shared the same Western medicinal diagnosis (postoperative acute pain) and the same traditional Chinese medicinal diagnosis (excess heat condition with qi and blood stagnation), making this group an exceptionally homogeneous patient population. Heterogeneous populations in whom multiple signs and symptoms are treated, e.g., quality of life, pain, function, and mental status, may require individualized treatments.

Deciding on formulaic vs. individualized acupuncture can also be an integrated rather than either/or approach. For example, both can be considered in a stepwise approach, progressing from the more rigorously controlled (formulaic) to the more practical (individualized). If results are positive using a formulaic approach, then research can progress into individualized treatments based on TCM diagnoses. Another integrated approach is to select comprehensive formula points which cover multiple TCM diagnoses. For example, the formulaic treatment used in the osteoarthritis study [1] covered most of the TCM points that would have been used, had individual syndromes been treated. Yet another option is to include at least one point

from every TCM diagnosis. An additional compromise between formulaic and individualized protocols is to stratify patients by TCM subgroup diagnosis and treat patients at subgroup-specific acupoints (see discussion of double screen design later in this chapter).

11.2.3

Training/Experience of the Acupuncturist

How skilled and experienced the acupuncturist should be in clinical trials is another common question. Proficiency requirements of the acupuncturist may differ, based on whether formulaic or individualized treatments are used. Obviously, when individualized treatments are given, the acupuncturist must be very experienced in both diagnosing and treating the condition under investigation in order to make treatment decisions based on TCM diagnoses. However, when formulaic treatments are given, it may suffice for the acupuncturist to demonstrate proficiency and replicability of the needling technique being used.

11.2.4

Selection of Points

Birch has observed [7] that there is no consensus among acupuncture textbooks on the points to use for a given condition. The absence of standardized treatments makes it important to base the selected points both on the literature and personal clinical observations of effectiveness.

11.2.5

Total Number of Treatments

The sufficient total number of treatments to administer in a trial, especially for chronic conditions, is largely unknown and seldom specified in the literature. Phase I presents a low-cost opportunity to establish both total number of treatments and the frequency of treatments for the investigated condition. In establishing these parameters, outcomes need to be measured frequently to determine when improvement begins and whether additional improvements occur with more treatments. For example, in the osteoarthritis study, both clinical experience and prior RCTs in the literature suggested that treatments twice weekly for three weeks were probably not adequate [8–10]. Phase I results showed substantial improvements at 4 weeks and continued improvements at 8 weeks. Therefore, treatment twice weekly for 8 weeks was selected as the treatment modality for phase II.

11.2.6

Selecting Outcomes

Phase I provides an opportunity to select appropriate outcomes to be tested. The outcomes should be measurable according to validated instruments. A validated instrument means that it has been demonstrated through prior research that it is measuring what it says it is measuring. Examples of well-validated general health assessment

tools include the Sickness Impact Profile (SIP) [11] and the SF-36 [12]. Because these measurement tools have been validated using the whole instrument, it is a good idea when using these instruments to use them intact; using pieces of an instrument can jeopardize the validity of a trial. Instruments can be generic (measuring overall health status) or disease-specific (measuring clinically important changes pertinent to that specific disease) [13]. Frequently, studies include both. It is helpful to look at the recent literature pertaining to the disease under investigation in order to decide which instruments to use. Excellent, user-friendly guides on measurement selection exist [14].

Another important decision to be made is which will be the primary and which the secondary outcomes. Primary outcomes, designated by investigators as the most important outcomes, should be limited to one or two, clearly defined, and used for calculating sample size in subsequent trial phases. Secondary outcomes can represent other symptoms of interest, quality of life, sleep patterns, appetite, bowel movements, and mental status. However, it is important to remember that because sample size is not based on secondary outcomes, significant changes between groups might not be observed in all of these secondary outcomes. For example, in the dental pain study [5], the primary outcome was pain-free time after surgery, and sample size calculation was based on that. A secondary outcome was pain relief after the first acupuncture treatment. Results showed a significant difference between groups in the primary but not the secondary outcome. One possibility for this finding is that, in reality, there is no significant difference in pain relief in real vs. placebo acupuncture once the pain has reached moderate intensity. However, another possibility is that the sample size may have been too small to detect a difference between groups in the secondary outcome. Similar situations have occurred in other acupuncture trials [15].

11.2.7

Choosing a Follow-up Time

Selecting the appropriate follow-up time depends on the condition being tested and can be investigated in the early phases of a trial because, especially in chronic conditions, this may influence the final treatment protocol to be used in subsequent phases. For acute conditions such as dental pain [3–5], a week may be adequate to determine whether the treatment is effective. By contrast, for chronic conditions such as OA, a follow-up of 3, 6, or 12 months may be more appropriate [1, 2]. Long-term follow-up is important to determine long-term efficacy with acupuncture. If efficacy is not sustained, investigators can examine ways to sustain it.

Follow-up data can provide useful information for designing future trial phases. For example, long-term follow-up data from two previous OA studies guided the formulation of the acupuncture protocol in the large, randomized OA trial. Firstly, long-term follow-up data by Christensen and colleagues [9] in a group of OA patients showed that improvements could be sustained for a year using monthly acupuncture after treatments twice a week had stopped. However, it was not evident from this study whether improvement would be sustained without booster treatments. Then, long-term follow-up data from Berman and colleagues [1] showed that, without booster treatments, pain scores returned to pretreatment baselines by 24 weeks. As a result of these follow-up data, the decision was made in a large OA trial to give

monthly booster acupuncture treatments after patients had completed the 8-week treatment regimen.

11.2.8

Measuring Safety

Unlike phase I trials for a new drug, which must closely examine safety, acupuncture is already considered safe. However, it is important as part of any clinical assessment to document adverse events. Confusion arises when a paper mentions that no adverse effects were observed. Frequently, it is not clear whether adverse effects did not occur or simply were not measured. Therefore, it is advisable to have an explicit and systematic methodology such as a symptoms checklist for documenting side effects. Phase I provides the opportunity to refine the checklist based on the condition being investigated.

11.2.9

Comment

The major advantage of a phase I trial is that it enables investigators to generate preliminary information on dose and safety in a cost-effective way. If no benefit is observed during phase I, it may be because the selected acupuncture protocol (dose) is not appropriate and/or because the condition under investigation is not responsive to acupuncture. In such a circumstance, the researcher might alter the acupuncture protocol or inclusion/exclusion criteria and conduct another phase I trial. If patients do not respond, then the researcher knows early on that the research question may not be worth pursuing in a larger, more costly, randomized controlled trial.

However, if benefit is observed in phase I, it cannot be assumed to be due to the acupuncture protocol. This is because phase I trials have no control group and are not efficacy trials. There are three explanations for why patients may experience benefit during an intervention: (1) treatment effects, (2) nonspecific effects, and (3) disease remissions [16]. To be able to attribute improvements to treatment effects, the investigator must first rule out disease remissions and nonspecific effects. “Nonspecific effects” is frequently used to describe all of the nontreatment effects which result from the milieu in which the treatment occurs, including the physical environment, the caring staff, and beliefs and expectations of the patients. Confusion arises in acupuncture trials because the term “nonspecific effects” is often used to describe physiological effects due to needling stimulation, such as diffuse noxious inhibitory control (DNIC). For clarity, this term will be used to describe nontreatment-related effects, and “nonspecific physiological effects of needling” will be used to describe the effects of acupuncture that result from needle stimulation but are not acupuncture point-specific. Ruling out improvements due to nonspecific effects and spontaneous remissions requires the use of a control group, which is commonly introduced in phase II.

11.3

Phase II

Phase II trials build on the information gained in phase I. Phase II trials generally have a control group and generate additional information on dosing and safety as well as preliminary information on efficacy (Table 1). Comparison groups can concern different doses of the same treatment, different treatment, or placebo groups. Randomization is frequently done in phase II, although the numbers are generally small.

11.3.1

Selecting an Appropriate Control Group

Unlike pharmaceutical research, in which a placebo is easy to design and double blinding is easily achieved, it is difficult to design an appropriate control group for acupuncture, because acupuncture is an invasive, physical modality. No single control group can answer all the research questions. Therefore, many types of control groups, each with its own advantages and disadvantages, have been used, depending on the specific question being asked (Table 2) [17]. In choosing an appropriate control group, it is important to consider what current treatments are available for the condition being investigated. If a condition already has an effective standard medical treatment, then it makes no ethical or financial sense to examine sham vs. real acupuncture. It is more logical to examine acupuncture effectiveness compared to standard care. In addition to overall effectiveness, it can also be noted whether acupuncture has a quicker onset, lasts longer, or has a milder side effect profile than standard care. It is also logical to examine whether additional benefit to standard care can be derived by adding acupuncture treatments.

11.3.2

Waiting Lists

Waiting lists can track the natural history of a disease and assess for spontaneous remissions. Therefore, the value of a waiting list over not using a control group is that waiting lists provide a way to estimate remissions due to disease variation rather than treatment effectiveness. For example, in the OA trial [1], a waiting list control group was used. Both patient groups were instructed to continue using whatever medications they were already using. Acupuncture was administered immediately to one group and delayed in the other group. There were several advantages to the choice of this control group. It controlled for spontaneous remissions, provided clinically useful information on acupuncture as an adjunct to treatments already being used and offered an ethical design for the control group to receive the experimental acupuncture treatment. Obviously, the limitation to this group was that it could not control for placebo effects.

The importance of controlling for disease remissions using a waiting list group was underscored in an acupuncture trial in which 30% in the waiting list group experienced remissions in pain [18].

Table 2. Control groups in acupuncture trials

Type of control	Questions asked	Advantages	Disadvantages
Waiting list (delayed treatment)	Is acupuncture more effective than no treatment?	Controls for disease remissions; all patients receive treatment	Does not control for placebo effects
Nonacupuncture inert controls (sham TENS, sugar pills)	Is acupuncture more effective than a placebo?	Controls for some placebo effects	Does not resemble acupuncture; cannot blind patients
Placebo acupuncture (noninserted needle)	Is acupuncture more effective than placebo?	Resembles real acupuncture; patients can be blinded; eliminates the possibility of non-specific needling effects	Difficult to implement in long-term studies; may be effective only for acupuncture-naïve patients; does not test for specificity of acupoints
Sham acupuncture (inserted needle)	Is real acupuncture more effective than sham acupuncture? Does real acupuncture have point-specific effects on the condition under investigation?	Resembles real acupuncture; patients can be blinded	Likely to produce non-specific physiological effects of needling
Combined controls (e.g., placebo acupuncture plus sham acupuncture)	What is the magnitude of placebo acupuncture (no treatment vs. placebo)? What is the magnitude of nonspecific effects (placebo vs sham)?	The two treatments resemble each other, so patients can be blinded; can minimize nonspecific needling effects of sham	If placebo acupuncture is used, it may be difficult to implement in long-term studies
Positive controls (standard medical care)	Is acupuncture equivalent/ superior to standard medical care or is it, when combined with standard care, more effective than standard care alone?	Compares the effectiveness of acupuncture as replacement or adjunctive care; has practical value, since cost effectiveness, adverse effects, and efficacy can be compared	Cannot blind patients or practitioners; may risk a type II error (where the two treatments are believed to be equivalent but, in reality, one is better)

11.3.3

Nonacupuncture Inert Controls

A placebo is a physiologically inert intervention. Sham transcutaneous electrical nerve stimulation (TENS) has been used in acupuncture trials to estimate placebo effects because it provides a control intervention that is undoubtedly physiologically inert and has been demonstrated to be a credible treatment [19]. Unlike patients in the waiting list control, patients in the sham TENS group receive amounts of practitioner time and attention similar to those for patients in the acupuncture group. Therefore, the sham TENS group controls for some nonspecific effects, such as those resulting from patient-practitioner relationship. However, the placebo effect of acupuncture cannot be measured, because sham TENS does not resemble acupuncture. Patients cannot be blinded in a sham TENS vs. acupuncture trial. In other words, patients know whether they received acupuncture or not.

Sugar pills, another type of nonacupuncture inert control, have also been used [20]. However, sugar pills bear even less resemblance to acupuncture and cannot control for the nonspecific effects resulting from the patient-practitioner relationship, because the practitioner time and attention is not comparable in patients receiving acupuncture relative to those receiving sugar pills.

11.3.4

Placebo Acupuncture

Placebo acupuncture is a true control for the acupuncture treatment because it is not only physiologically inert but also resembles real acupuncture. In placebo acupuncture, the blunt end of the needle [21], acupuncture needles, or objects resembling needles such as toothpicks [22], needling guiding tubes [3–5], or specifically designed noninsertion needling devices [23, 24] create the appearance of insertion but are not actually inserted.

In order to enhance the success of patient blinding in placebo acupuncture, some investigators [3–5] exclude patients who have previously received real acupuncture. Patient blinding might also be improved by creating the expectation that acupuncture is painless, such as by showing placebo patients that the diameter (≤ 0.2 mm) of the commonly used acupuncture needle is much thinner than hollow hypodermic needles. Placebo acupuncture may be especially useful in short-term, acute intervention trials. A possible disadvantage is that in conditions requiring long-term treatment, maintaining patient blinding may be difficult. This is because curiosity of the patients to discover the type of treatment they have been receiving may lead to unblinding (e.g., patients may talk to each other, read about acupuncture, or go to another acupuncturist). A few investigators [3–5, 22, 32] have checked the blinding credibility after the placebo intervention by administering a questionnaire asking patients which treatment they believe they have received. This procedure is important in order to validate whether blinding has succeeded.

11.3.5

Sham Acupuncture

11.3.5.1

Sham Acupuncture for Measuring Nonspecific Effects of Needling

Unlike placebo acupuncture that involves noninvasive procedures, sham acupuncture involves actual needling but at sites inappropriate to the condition being examined [8,11, 25–28]. Like placebo acupuncture, the advantage in sham acupuncture is its resemblance to real acupuncture so that patients can be blinded to treatment group assignment. However, since sham needling mimics both the nonspecific effects and the nonspecific physiological effects of needling, the difference between real and sham groups may be slight, especially in pain trials. Such nonspecific physiological effects of needling may include local alteration in circulation and immune function [29] and triggering of neural pathways such as those resulting in diffuse noxious inhibitory control (DNIC) of pain [30].

In order to minimize the nonspecific physiological effects of needling, the sham procedure may use distal points, minimal needling depth, and minimal needle stimulation rather than local points, standard depth, and needle stimulation [31]. Avoiding de qi in the sham group is also important. Because of the magnitude of these nonspecific needling effects, careful consideration is needed to calculate an adequate sample size. As in placebo acupuncture, it is important to confirm patient blinding by providing a questionnaire asking patients which treatment they believe they have received [32].

In addition to producing nonspecific needling effects, another disadvantage of sham acupuncture is that there is no general agreement as to how it should be designed [17]. For example, sham acupuncture may be applied by needling nonpoints adjacent to real points [26], nonpoints distal to real points [27], or real acupuncture points not specific for the condition to be treated [33]. Needling techniques in sham acupuncture can be applied at various depths, e.g., superficial or standard, with various types of stimulation, e.g., manual or electrostimulation, using various needle manipulation techniques, e.g., reinforcing, reducing, and changes in needling angles and directions, and retaining needles for various times, e.g., for similar or shorter duration than in treatment groups. These variations in the way sham acupuncture is applied make it difficult even to compare the results of two sham acupuncture trials investigating the same condition. For example, in an OA RCT, Takeda and Wessel [10] noticed that de qi was sometimes inadvertently elicited in the sham group and sometimes not elicited in the real acupuncture group. Although results showed no significant differences between groups, de qi was a predictor for significant improvement in both the *Western Ontario and McMaster Universities Osteoarthritis Pain Index* and the pressure threshold scores in both groups. Clearly, comparing the results of this sham acupuncture trial may not be comparable with the results of another trial which avoids de qi in the sham group.

11.3.5.2

Sham Acupuncture for Measuring Acupuncture Specificity

Sham acupuncture is also the appropriate control for examining the validity of TCM claims for acupuncture point specificity. It can be used to test point specificity by comparing the physiological effects of needling a precisely located real acupoint to an adjacent nonpoint according to standard TCM textbooks. It can also test specificity by comparing the effects of needling a group of real points prescribed for a condition according to TCM theory to a group of inappropriate points. And it is useful for testing specific physiological effects of various needling techniques such as reinforced or reducing technique, or techniques that elicit de qi.

When sham acupuncture is devised to investigate acupuncture specificity, different design issues need to be considered than when sham is devised to measure placebo effects. In the latter case, the goal is to minimize nonspecific physiological effects of needling and, therefore, several features of the sham procedure may differ from procedures with the real acupuncture group, including points selected, depth of insertion, and amount of stimulation applied to the needle. In contrast, when sham acupuncture is used in the former case as a control to examine acupuncture specificity, i.e., to test whether stimulation of a real point has greater physiological effects

than a nonpoint, then all the features of the sham treatment should be identical to the real treatment except for the feature under investigation, i.e., point locations.

11.3.6

Combined Controls

Combining various controls within the same study is a creative way to minimize the limitations of each type of control. It is apparent that a major challenge in acupuncture research is the ability successfully to blind patients to their treatment group assignment, because when the comparison group's treatment does not resemble acupuncture, blinding is impossible. One innovative approach is to combine various placebo interventions within the same study to make the treatments in the two groups appear comparable. For example, in comparing medication to acupuncture for migraine, Hesse et al. [21] gave real medication plus placebo acupuncture to one group and real acupuncture plus placebo medication to the other. This innovative approach enabled them to blind patients to treatment group assignment.

Another challenge in acupuncture research is to develop placebos that resemble acupuncture while minimizing nonspecific needling effects. A rarely used, innovative approach is to combine both sham acupuncture (inserted needles) and placebo acupuncture (noninserted needles) in the same design. For example, in a recently funded dental study, the real treatment consisted of real acupuncture at points St.6 and 7, SJ.17, and LI.4. The placebo treatment consisted of placebo acupuncture (noninserted needles) administered to these same points. However, in order to improve blinding in the placebo group, a sham acupuncture (inserted needle) point was added to the leg 1 inch posterior to Liv.8, thereby ensuring that the placebo group experiences the sensation of needle insertion. Then, to make the two groups comparable in appearance, placebo acupuncture (noninserted needle) was added to the same leg point in the real acupuncture group. Combined controls such as these have two advantages: (1) blinding in both groups is enhanced because the two treatments are comparable in appearance, and (2) blinding in the real acupuncture group is further enhanced because the placebo leg point may make the real treatment appear less real.

11.3.7

Positive Controls

The question asked in the type of research design that uses a positive (active) control group [34] is different from those asked with any of the above control groups: is acupuncture at least as effective as standard medical care such as medication, physiotherapy, or occlusal splint? This design is useful for conditions in which side effects of medications are problematic, since it allows for side by side comparisons of adverse outcomes [21, 35]. Acupuncture can also be added to standard care to determine if it can increase the effectiveness of other treatment [36] or even reduce the required dose [37]. Designs which assess acupuncture compared or added to standard care also provide an opportunity for cost effectiveness comparisons [36]. When planning a trial using positive controls, it is important to ensure adequate power, because otherwise nonsignificant results are difficult to interpret. It cannot be assumed that acupuncture has performed as well as standard care unless a type II error can be ruled

out. A type II error occurs when there actually is a difference between the two treatments but the sample size is too small to detect it. Such results are not significant, but a larger sample size would have found a significant difference between the two treatments.

11.3.8

Selecting the Study Design: Parallel or Crossover

A final consideration for phase II trials is which type of study design to select. Several acupuncture trials have used crossover designs, mostly for the sake of convenience, because crossover designs increase statistical power without increasing patient enrollment. Three assumptions must apply for crossover designs to be an appropriate experimental design: (1) the disease must be stable throughout the study, (2) there is no reason to believe that the treatments interact with each other, and (3) the effects of the proposed intervention cease soon after discontinuing the intervention[38].

The latter point calls into question the appropriateness of crossover designs for acupuncture trials. Evidence of long-term effects of acupuncture have been observed in controlled clinical trials in which significant pain relief achieved after 6 weeks of treatment for migraine [26] or 3 months of treatment for dysmenorrhea [27] was maintained at 1-year follow-up. Not surprisingly, carryover effects (treatment effects which persist long after treatment ceases) have been documented in crossover trials of acupuncture [39]. The presence of carryover effects in acupuncture studies suggests that crossover designs are an inappropriate design choice, except in trials with very long washout periods.

Furthermore, in crossover studies comparing real versus sham acupuncture, it is questionable whether patients can maintain blindness to treatment groups once they have experienced both real and sham acupuncture treatments. For these reasons, parallel rather than crossover designs are preferred for acupuncture efficacy trials.

11.3.9

Summary and Example

There are several advantages to conducting a phase II trial before plunging into a large, definitive phase III randomized controlled trial (RCT): it (1) allows for the hypothesis to be tested on a small, cost effective scale, (2) provides the data needed to do the power calculation for a larger trial, (3) assesses the feasibility of the research setting, including staffing availability, and (4) provides a forum where unforeseen problems can arise and be resolved so that a solid study protocol can be developed for the larger phase III randomized trial.

For example, phase II of the dental pain trial [4] provided an excellent opportunity to refine the timing of activities within the trial. Prior to phase II, the treatment protocol had been based on a previous trial in the literature which administered acupuncture when patients experienced moderate pain [25]. However, during phase II, the investigators observed that acupuncture given pre-emptively was a more effective analgesic, so the protocol was modified accordingly.

Another example was that phase II provided an opportunity to assess whether the timing of administering the blinding questionnaire was appropriate. Prior to phase II,

it was determined that patients would be asked before they left the clinic which treatment they believed they had received. The investigators learned during phase II that, by that time, patients' guesses were based on their level of pain relief and were fairly accurate. However, when the question was asked immediately after the first treatment, prior to the experience of breakthrough pain, guesses were based on characteristics of the treatment, and the placebo acupuncture procedure was found to be credible.

11.4

Phase III

Phase III trials are the definitive efficacy trials. They are large and often conducted at multiple locations. In the OA study, the phase III trial consists of a three-arm parallel design ($n = 570$) comparing real and sham acupuncture vs. an attention control group. The systematic reviews discussed in chapter 7 provide excellent commentaries on the methodological quality of acupuncture efficacy trials. This section proposes guidelines for strengthening acupuncture efficacy trials based on those commentaries. Obviously, many of these guidelines apply to both phase II and III trials.

11.4.1

Select Biologically Meaningful Inclusion/Exclusion Criteria

One must be explicit about the biological rationale for the inclusion/exclusion criteria. For example, there are advantages and disadvantages to including heterogeneous study populations. An advantage of heterogeneous study populations is that the study results may be generalized to a broader group of people. Furthermore, as has been mentioned previously, it is unlikely that formulaic acupuncture can ever be devised for a very heterogeneous study population. The disadvantage is that heterogeneity may dilute treatment effects. In either case, the most important consideration is that the inclusion criteria make biological sense based on knowledge of the disease and of acupuncture.

11.4.2

Ensure Adequate Power

The most universal observation made in systematic reviews of acupuncture trials is that, with few exceptions, the numbers of patients were notably small. This makes acupuncture trials difficult to interpret because they can be prone to type II errors. Type II errors occur when a study obtains a nonsignificant result although a significant difference between groups exists in reality [40]. Unfortunately, there is no way to tell whether nonsignificant results reflect reality or result from type II errors due to inadequate statistical power. To reduce the possibility of type II errors, it is important to base sample size estimates a priori on a statistical formula [40], estimate expected differences from studies which used a control group comparable to the one intended for the planned study and, when possible, base sample size estimates on pilot data. It is best to recruit the skills of a biostatistician when addressing these statistical issues throughout the trial.

Trials comparing real vs. sham acupuncture for the treatment of pain may be especially prone to type II errors because they have larger sample size requirements than trials comparing real acupuncture to inert placebos for the treatment of pain. This is because sham (invasive) needling induces nonspecific physiological effects that include partial pain modulation [17]. A systematic review of acupuncture [41] for chronic pain demonstrated that the proportions of patients responding to sham acupuncture were significantly higher than the proportions responding to inert placebos.

11.4.3

Minimize Selection Bias Through Adequate Randomization

One of the most important potential biases can result from the way that patients are allocated to treatment groups. Known as selection bias, this is a systematic difference in the baseline characteristics of comparison groups [42]. Randomization offers the best way to balance known and unknown prognostic factors equally between groups. There are two distinctly important features in group allocation, both of which must be present in order to reduce the chances of selection bias: the first is that a true randomization procedure needs to be implemented (i.e., using a table of random numbers or a computer-generated randomizing scheme). The second, to be discussed in Section 11.4.4, is that the allocation procedure needs to be tamperproof.

Small studies are prone to selection bias because they are unlikely to achieve balance on important prognostic factors. If imbalances in group allocation result in less severely impaired patients being allocated to the treatment group and more severely impaired patients to the control group, then selection bias can lead to an artificially high response rate in the treatment group. Small trials are often thought to be problematic only because insufficient power may lead to findings of no treatment effect. However, they can truly be problematic due to selection bias and actually overestimate treatment effects by as much as 30% [43]. The importance of baseline comparability has led Linde et al. [44] to include a question in their quality checklist which asks whether groups were comparable at baseline. It is always a good idea to report baseline comparisons of the two groups in any of the study publications. The difficulty of how to interpret results of a trial when groups are not comparable in baseline characteristics was recently evidenced in an RCT for headache [22]. The results of this methodologically rigorous pilot study, which scored a perfect score on the Jadad scale, were deemed impossible to interpret [45] due to a baseline difference in the weekly headache index. More severe headache index scores clustered in the real acupuncture group, thus giving placebo acupuncture an unfair advantage.

One way to assist the balance between groups is to stratify before randomizing. Generally, one or two factors known to be associated with the outcome of interest, such as CD4 counts in an AIDS study, are selected for the strata. Randomization is then done within each stratum.

11.4.4

Minimize Selection Bias Through Allocation Concealment

Allocation concealment is a term used to describe the method by which foreknowledge of treatment assignment is prevented and the allocation procedure is kept tamperproof [42]. Each patient's eligibility for the trial should be determined before he is

randomly given a group assignment. Once assignment has been made, both the treatment group assignment and the eligibility decision should be unalterable. Obviously, allocation concealment is more likely to succeed if the person responsible for allocation is not the same one responsible for recruiting subjects or assessing eligibility.

Lack of adequate allocation concealment is more strongly associated with bias than how the randomization sequence is generated [42]. Lack of adequate allocation concealment can exaggerate treatment effects by as much as 41% [46]. Obviously, transparent allocation systems such as alternate assignment, open lists, allocation by date of birth, or pulling numbers from a hat are not sufficient. Systems which sufficiently conceal allocation can include calling a central office for the next allocation assignment or an in-house computer file which yields the allocation only after the study identification number of the eligible person is entered.

11.4.5

Minimize Performance Bias by Blinding Patients and Practitioners

Performance bias is a systematic difference in the way patients or practitioners behave as a result of knowing the treatment group assignment [42]. Performance bias in unblinded patients, for example, can result in patients' reporting more symptoms.

Performance bias in unblinded practitioners can occur if the practitioner inadvertently treats the patients in treatment and control groups differently, such as being more caring or encouraging with those in the treatment group. Because practitioners can seldom be blinded in acupuncture trials, care must be taken to minimize practitioner performance bias. For example, the protocol might require that practitioners minimize verbal communication with the patient or use only standardized answers. It has also been suggested that videotaping treatment sessions could provide a way of detecting differences in practitioner behaviors so that they may be corrected while the trial is in progress [47].

Clearly, the optimal way to prevent performance bias is to blind those providing and receiving care to treatment group assignment (called double blinding). If a trial is not double blind, effects can be overestimated [43]. Double blinding is seldom possible in acupuncture trials because blinding those who are receiving the treatment is not possible if the two treatments do not resemble each other, as in sham TENS vs. acupuncture. Therefore, investigators must anticipate how performance bias might be influencing results and find ways to minimize it.

11.4.6

Minimize Detection Bias by Blinding the Outcome Assessors

Even when patients and practitioners cannot be blinded, the outcome assessor always can and should be blinded. Avoid detection bias by blinding outcome assessors. Detection bias is a systematic difference in the way outcomes are assessed due to lack of unblinded outcome assessors [42]. Put another way, unblinded outcome assessors may see what they want to see. This is particularly problematic in outcomes such as x-rays or global estimates of improvement that require a level of subjective judgment. It is less problematic in outcomes that require no subjective judgment, such as death. In acupuncture trials, unblinded outcome assessors who believe that acupuncture is

beneficial may tend to find more global improvements or less disease progression in the real acupuncture group. It is important to note that some investigators use the term "double blinding" to mean that the outcome assessor is also blinded [48]. Therefore, it is important to ascertain exactly who is being blinded when "double blinding" is used.

11.4.7

Control for Cointerventions

Cointerventions, those treatments other than the experimental treatment that the patient may use while enrolled in the trial, must be accounted for in order to ascertain if benefit can be attributed to the experimental treatment [49]. Cointerventions can include medications, massage, physiotherapy, meditation practice, or lifestyle changes such as diet and exercise.

For example, one scenario might be that the control group does not experience benefit from the control treatment so that they turn to cointervention for relief. If cointerventions are not accounted for, then the two groups could appear to have equally effective treatments. This potential problem often leads trialists to ask patients to refrain from using cointerventions during the study period. When this is neither practical nor ethical, providing patients a way to document the use of cointerventions, such as recording in a daily diary, will provide a way to control for cointerventions in the statistical analysis.

11.4.8

Avoid Attrition Bias by Performing Intention-to-Treat Analysis

Even if selection bias has been successfully prevented during the allocation stage of a trial, it can still occur during the follow-up trial if only completers are counted. Selection bias during the follow-up period due to dropouts and withdrawals is called attrition bias. Attrition bias, therefore, results from a systematic difference in withdrawals [42]. Intention-to-treat analysis, which treats dropouts as nonresponders, is the preferred statistical approach to preventing attrition bias. Examples of intention-to-treat analysis are a relatively recent occurrence in the acupuncture literature [1, 50].

The rationale behind intention-to-treat analysis is to preserve the randomization scheme, thus preserving the balance in prognostic factors achieved during allocation to treatment group. Therefore, intention-to-treat analysis counts everyone who was randomized, including those who dropped out before ever receiving treatment, those who stopped coming for treatment, and those who were lost to follow-up. Although treating all dropouts and withdrawals as nonresponders may downplay the true effect of a treatment, it will not bias results.

11.4.9

Avoid Reporting Bias by Reporting on all Prespecified Outcomes

Reporting bias, the systematic difference in how outcomes are reported, can occur during writeup [42]. Because some consider positive findings more interesting than nonsignificant findings, investigators might be inclined to report only those out-

