THE JOURNAL OF ALTERNATIVE AND COMPLEMENTARY MEDICINE
Volume 7, Number 3, 2001, pp. 213–218
Mary Ann Liebert, Inc.

EDITORIAL

The Efficacy Paradox in Randomized Controlled Trials of CAM and Elsewhere: Beware of the Placebo Trap

Ever since its formal inauguration in 1945, the double-blind, randomized controlled trial (RCT) has become the "Holy Grail" of medical methodology (Kaptchuk, 1998a). This is understandable because the RCT has an advantage that other methods are lacking: it—ideally—precludes bias. By randomizing patients to groups and by concealing the allocation the RCT precludes selection bias. By blinding doctors and patients this type of trial prevents observation bias. By blinding outcome assessors and statisticians the RCT prevents reporting bias. Easy, do-it-yourself kits of methodology assessment such as the Jadad-score (Jadad, et al., 1996) have incorporated these virtues of RCTs and take it for granted that only studies that include all of these elements can be methodologically sound. This, recursively, leads to the situation that authors of studies that naturally want to do "good" studies seek to fulfill all the criteria posed by quality codings such as the Jadad score. Thus, a standard is created that is then called the "gold standard" without anybody reflecting how it came to be a gold standard in the first place. This standard was instilled by a part of the medical community that equated effective therapy with the use of pharmacologically active and powerful drugs (Kaptchuk, 1998b). And, meanwhile, this global and universal hypnotherapeutic intervention has succeeded: At least the public, but also a large part of the scientific medical community, is convinced that treating is always better than not treating and that treating, if it is to be effective, is treating with specific pharmacologic agents. Thus, an array of implicit presuppositions (Collingwood, 1940) has taken hold of researchers in the medical community, which, by virtue of its unreflected nature, make the gold standard a "golden calf" (Kaptchuk, 2001). The most important of these presuppositions is also the most dangerous one. It is the presupposition that effectiveness can only be granted if there is efficacy, and efficacy is identical to specific efficacy against placebo. In what follows, I am going to show that this equation is ill-founded because it makes presuppositions that are—maybe—true for pharmacologic interventions but—very likely—false for complex interventions that are not meant to act directly but indirectly via stimulation of the autoregulative functions of the body.

To prevent misunderstandings: I am not arguing that the RCT is a bad or even wrong methodology. On the contrary, the RCT is a superb instrument, like the lancet of the surgeon. But you would not want to use the lancet for slicing bread or chiseling marble.

Modern pharmacology was founded on the basis of the prime paradigm of modern medicine: Virchow's cellular pathology. This stated that disease is identical to pathologic changes on the cellular level. Thus, understanding the physiologic and cellular nature of the organism in states of health and disease enabled modern medicine also to tailor powerful interventions in order to prevent or counteract these pathologic changes. This research paradigm has provided us with magnificent insights into the working of the body's physiology and is far from having reached an end. Pharmacologic interventions try to act on the supposed or known pathology at the cellular level. Aspirin inhibits the synthesis of prostaglandines, thus intervening at the inflammation cycle and preventing inflammation and pain associated with this cycle. While we have come to know the

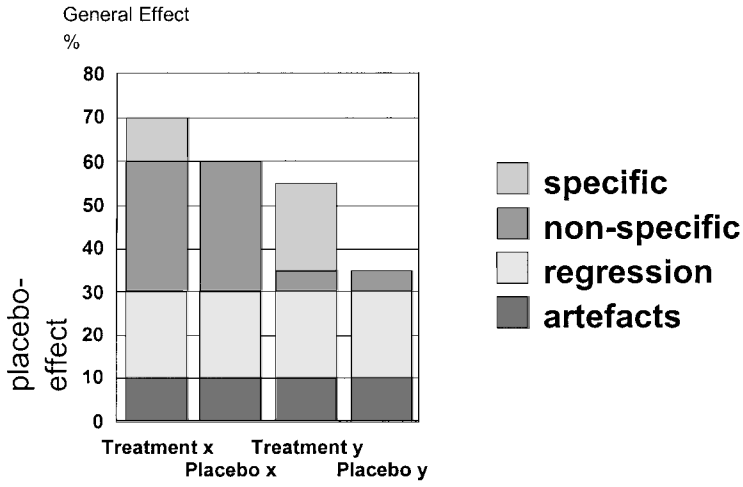

FIG. 1. Thought experiment illustrating the efficacy paradox. See explanation in text.

precise mechanisms of aspirin only recently, it has been in use as a powerful therapeutic agent for decades before the precise knowledge of its mechanism was discovered, solely on the basis of careful observation and experience (Schonauer, 1994). Ideally, this should happen the other way around: We know about the pathologic physiologic processes that we want to inhibit and we look for a substance that can do this. Hence, specific efficacy is a prerequisite of effectiveness for pharmacologic interventions. It is the very aim of pharmacologic interventions to find a specific inhibitor, enhancer, or modifier for a specific, known pathway to be modified. What we tend to forget is that this is only one of many options for therapeutic interventions. The modern medical paradigm has adopted this mechanistic cellular–pathologic model and has fared well with this decision, at least concerning acute diseases and interventions. Conventional medicine has fared less well when it comes to chronic diseases. Most of these chronic diseases have to do with a lack of regulation within highly complex interconnected systems, such as the immune system or the interactions between the immune system and the neuroendocrine systems, or even between the psyche and these systems. And with complex diseases such as rheumatoid arthritis, asthma, fibromyalgia, chronic fatigue syndrome, to name but a few, it is too simplistic to look at one effector and try to find one specific intervention for it.

Enter complex interventions, such as the ones used in complementary and alternative medicine (CAM). These interventions rarely just address one effector mechanistically. They try mostly to stimulate or regulate the organism. Thus, their way of achieving efficacy is nonspecific, although they use very specific theories and interventions. The important point to note here is that the specificity stipulated within a model—such as the specificity of homeopathic remedies—need not necessarily be a specificity on the level of physiologic processes. It may be just a clever "trick" to bring about quite nonspecific changes. These CAM methods spur the organism to do something by itself to find its original balance. It is possible that these methods also have specific parts but, if so, this occurs in a different sense than modern pharmacologic interventions. And it is very likely that a large part of their effectiveness is covered by most of their nonspecific effects. This leads to what I call the efficacy paradox, which is illustrated in Figure 1. This is a thought experiment, i.e., an idealized situation, which can easily be converted into practical reality.

Imagine the following situation as depicted in the figure: Let there be two treatments x and y for the same condition, say chronic pain. Let there be two placebo controlled RCTs with comparable patient populations. In every one of these trials we will have measurement artifacts caused by unreliability of measures; let

them be equal in all groups. In every one of these trials, we will also have regression to the mean as a statistical artefact and as a result of the natural course of the disease studied; some patients will improve regardless of the treatment applied. Then there will be nonspecific treatment effects: Patients expect to get better when treated, especially in a trial. Hope will work against the general demoralization caused by disease. The attention of doctors and nurses within the context of a trial and perhaps the special attention paid to patients within the context of a particular CAM intervention such as homeopathy, healing, or acupuncture, will also contribute to the nonspecific part of improvement. Let us not forget that a treatment that can help patients to understand their suffering by providing an explanation, a common explanatory myth, is a therapeutic factor, too (Frank, 1989). And then there will be specific factors of treatment. Let us assume that treatment *y* is specifically effective. Its specific efficacy will be 20%, which, in a trial that is adequately powered, will be significant. Thus, everybody will conclude: Treatment *y* is an effective treatment for chronic pain. Treatment *x* only has 10% specific efficacy and let us assume that studies of treatment *x* are generally underpowered to find this effect. Everybody will conclude: Treatment *x* is an ineffective treatment for chronic pain. What usually is overlooked is the fact that the nonspecific treatment effects of treatment *x* are much larger. In the thought experiment, I have chosen them to be 30% for treatment *x*. For treatment *y*, they would only be 5%. In such a case treatment *x*, although overall much more powerful with 70% of patients potentially benefitting from it by virtue of its strong nonspecific effects, would be neglected in favor of treatment *y*, with 55% of patients benefitting from it, because *y* has a stronger specific treatment effect.

I maintain that this situation is frequently true for CAM therapies. Studies are often underpowered, e.g., for acupuncture, and thus potential specific effects are overlooked. The conclusion of reviewers and the educated public then is the verdict "inconclusive evidence" (Ezzo, et al., 2001), and the political consequence, as just happened in Germany, is the decision to not include acupuncture in the scheme for public reimbursement, because the evidence for specific efficacy is inconclusive (Bundesausschuss Ärzte und Krankenkassen, 2001). However, nobody pays attention to the fact that perhaps the magnitude of nonspecific effects makes a treatment effective and not the specific effects. An even more complicated situation can arise when the circumstances of a trial, such as blinding and changing the natural flow of patient–doctor interaction and treatment sequences, change the context of a treatment dramatically and thus alter the potential nonspecific effects in a detrimental way. This can happen in blinded trials of homeopathy, in which insecurity arises from the blinding of doctors, and also in trials of acupuncture, when blinding procedures make it necessary that the doctor who is taking the case and making the assessment is different from the person who is administering the treatment. In all such cases, trials may alter the context of a treatment and thus diminish potent nonspecific factors and thereby underestimate effectiveness.

That the thought experiment is not just "cooked up" but simply an idealization of reality can be seen in a recent example of research in CAM. Abbot and colleagues conducted an intelligent trial of healing (Abbot, et al., 2001). Sixty patients who were suffering from chronic pain were randomized to either receive real healing or sham healing. Sham healing was performed by stage players who were otherwise not acquainted with healing. They studied healers and mimicked their conduct during a typical healing session such that, for the patients, it was not discernible which was true and which was sham healing. The second part of the study was a study of distant healing, in which healers were either present behind a one-way mirror or not present there. The second study was clearly negative. We therefore will not deal with this part any more but focus on the first one. Patients were seen twice for baseline measurement, an applaudable attempt to minimize error variance in measurement. Main outcome criteria were the Pain Rating Intensity Total Score of the McGill Pain questionnaire, a validated and frequently used measure, and the general score of the Measure Your Medical Outcome Profile (MYMOP), an individual scaling procedure for individual prob-

lems. Other measures, which we leave aside, were the Medical Outcomes Study Short Form 36, depression and anxiety measures, Visual Analogue Scale of pain, and specific MYMOP measures. We focus only on two major points. The overall result was negative. There was no specific difference, the authors claim, between healing and sham healing, and thus healing is not effective for chronic pain. The authors have provided a power analysis, which stipulated a specific effect of D = 0.8. The effect size d is a standardized mean difference, standardized by the standard deviation. It is frequently used, particularly if outcome measures are continuous, as in pain measurement. We know of no treatment, pharmacologic or otherwise, for patients with chronic pain, which has a specific effect of that size. After all, one of the definitions of chronic pain is that it is resistant to known effective treatment. Recent meta-analyses of treatments for chronic pain have found "large effects" of the size of D = 0.48 to D = 0.6. Nonsteroidal anti-inflammatory drugs for chronic pain reach effect sizes of d = 0.68, but often have smaller effects (Ward and Lorig, 1996). Patient education trainings for patients with rheumatoid arthritis reduce pain by D = 0.2 (Ward and Lorig, 1996) and are considered to be worthwhile today. Behavioral interventions have effects sizes of D = 0.5, which is considered by the authors of the meta-analysis to be "large" (Morley, et al., 1999) and antidepressant given to patients in pain have the same effects of D = 0.48 (Fishbain, et al., 1998). Thus, an effect size of D = 0.8 is quite unrealistic from the outset. Remember that it is still specific effectiveness we are talking about. Now if we look at the actual data and leave aside some statistical technicalities, we see the following: The within-group effect sizes for the primary outcome measures are quite large. These are measures, which are based on pre–post differences per group and reflect the improvement patients have experienced in each particular group, measurement error, regression to the mean, nonspecific effects, and specific effects, all included. This effect amounts to D = 1.12 in the healing group and to D = 0.83 in the sham-healed group for the main outcome measure and is somewhat smaller with D = 0.62 for the secondary outcome measure in the healed group and D = 0.34 in the sham group. The difference between the groups that reflects the specific effect is D = 0.29/0.28, which is small but consistent among primary and secondary outcome measures. (Note that the specific effect is not just the arithmetic difference of the within-group effect sizes, because the standard deviations are different.) The same is true for the secondary outcome measure. We should also mention that other measures, such as the scales used, produce a paradoxical picture because sometimes the control group is better than the treated group. I take this to be caused by the instability of measurement, with only 25 subjects per group, because those scales usually only provide stable measurements with larger numbers. Thus, the bottom-line of this RCT of healing really is: There is a huge effect for patients treated with healing, but it is mainly a nonspecific effect. The specific effect, although present, is small and was not detectable with a study powered to detect an effect of D = 0.8. Thus, the verdict of the trial is "healing is ineffective." This is a precise illustration of the efficacy paradox described above. If we convert an effect size of D = 1.12 in what is known as Rosenthal's *r*, we have something like *r* = 0.8. If we use Rosenthal's binomial effect size display (Rosenthal, 1991), we can say that, with such an effect size, approximately 90% of the patients treated would improve. Let us be conservative and turn the estimate down to 60%, which was approximately the size we found in our own trial (Wiesendanger, et al., 2001). My question is: Would a patient with chronic pain who is told that a treatment—specific or not, sham or not—will have a 60% chance of improving him or her decline treatment if told that "however, we must warn you that the treatment is not proven to be effective in the strict scientific sense?" What we have here is a real-world example of the efficacy paradox: A treatment is not proven as efficacious in the strict sense of the word, because a trial was unsuccessful in proving specific efficacy. However, the treatment is immensely effective in relieving patients overall, probably even more effective than proven efficacious treatments are, precisely because the nonspecific parts of the treatment are large, probably larger than in other types of treatment.

This is so because we remain hypnotized about placebo. Placebo is the "bad guy" in pharmacologic research. Legal requirements demand that a newly introduced drug is better than nonspecific elements of treatment, and justly so, because pharmacologic treatments also introduce risks, cost money, and produce waste products. Because of this situation, we are used to jumping to the conclusion that placebo effects are bad for everybody. But this is not so. Placebo effects exemplify all those processes that are not at the direct command of doctors and researchers. These effects exemplify all those elements of a treatment that lead to improvements without doctors or patients knowing how and why. These effects are the reflection of self-healing tendencies of an organism, possibly because of expectancy effects (Kirsch, 1997), possibly because of complex activations of psychoneuroimmunologic processes via the tiny specific stimuli of a treatment, such as the needling of acupuncture, the rituals of homeopathy, or the good intentions of spiritual healing. To just dump all these effects in the huge waste bin of medical research and dub these effects as "nothing but placebo" is, at best, scientific stupidity and testifies to an unwillingness of using one's brain to differ in the face of mainstream opinion. To call all these effects "nothing but artifacts" (Kirsch, 1997) is not seeing one's actions. Ludwik Fleck, the Jewish-German discoverer of the typhoid vaccine, who advanced the thesis of paradigmatic changes in science one generation before Thomas Kuhn, once said: "A scientific fact is a collective decision to stop thinking." (Fleck, 1980) This is true for the notion of placebo: People just dump everything in the container they label "placebo." And every therapeutic intervention that will be accepted as effective needs to be "better than placebo." But what if a therapeutic intervention is better than anything else but not better than placebo because the intervention is an exemplification of placebo processes namely of the self-healing capacity of the organism? What if "placebo" has completely different meanings depending on the context it is used? What if placebo effects are different in trials and in normal everyday practice? What if the magnitude of placebo effects outperforms the magnitude of the so-called specific effects? And there is some evidence that this is the case (Kirsch and Sapirstein, 1999; Walach and Maidhof, 1999; Maidhof, et al., 2000).

The placebo trap is the idea that only specific effects over and above placebo effects are worth looking for, worth demonstrating, and worth achieving. This is only true for the small sector of research that has the aim of proving the efficacy of newly developed drugs. It is not true for all those complex interventions that do not rely on one mechanistic specific agent but intervene in a more complex fashion, stimulating organisms toward self-healing actions.

In order to beware of the placebo trap we need to do three things:

(1) Argue against all those would-be methodologic "popes" who want to make everybody believe that efficacy is identical with specific efficacy against placebo.
(2) Diversify research strategies to use multiple methods, such as randomized comparison trials of CAM therapies against standard care or against waiting lists. This will enable us to quantify general therapeutic effects. Other research options would be large outcomes studies or comparative cohort studies in natural settings to address selection process.
(3) Start emptying the placebo waste bin and disentangle what it contains. Perhaps, at the bottom, we will find what is at the base of every true healing process: the capacity of the organism to heal itself. Starting to ask the question: "What is self-healing after all?" will be the beginning of meaningful research and will point the way out of the placebo trap.

REFERENCES

Abbot NC, Harkness EF, Stevinson C, Marshall FP, Conn DA, Ernst E. Spiritual healing as a therapy for chronic pain: A randomized, clinical trial. *Pain* 2001;91:79–89.

Bundesausschuss Ärzte Krankenkassen. (2001). Akupunktur: Zusammenfassender Bericht des Arbeitsausschusses "Ärztliche Behandlung" des Bundesausschusses der Ärzte und Krankenkassen über die Beratungen der Jahre 1999 und 2000 zur Bewertung der Akupunktur gemäss 135 Abs. 1 SGB V.

Collingwood RG. An Essay on Metaphysics. Oxford: Clarendon Press, 1940; reprinted in 1998.

Ezzo J, Lao L, Berman BM. Assessing clinical efficacy of acupuncture: What has been learned from systematic reviews of acupuncture? In: Stux G, Hammerschlag R eds. Clinical Acupuncture: Scientific Basis. Heidelberg: Springer Verlag, 2001:113–130.

Fishbain DA, Cutler RB, Rosomoff HL, Rosomoff RS. Do antidepressants have an analgesic effect in psychogenic pain and somatoform pain disorder: A meta-analysis. Psychosomatic Med 1998;60:503–509.

Fleck L. Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv. Mit einer Einleitung herausg. v. L. Schäfer und T. Schnelle. Frankfurt: Suhrkamp, 1935; reprinted in 1980.

Frank JD. Non-specific aspects of treatment: The view of a psychotherapist. In: M. Shepherd M., & Sartorius N, eds. Non-Specific Aspects of Treatment. Bern: Huber, 1989:95–114.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds, DJM, Gavaghan DJ, McQuay H. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17:1–12.

Kaptchuk TJ. Intentional ignorance: A history of blind assessment and placebo controls in medicine. Bull Hist Med 1998a;72:389–433.

Kaptchuk TJ. Powerful placebo: The dark side of the randomised controlled trial. Lancet 1998b;351:1722–1725.

Kaptchuk TJ. The double-blind randomized controlled trial: Gold standard or golden calf? J Clin Epidemiol 2001; in press.

Kirsch I. Specifying nonspecifics: Psychological mechanisms of placebo effects. In: Harrington, A (ed.) The Placebo Effect: Interdisciplinary Explorations. Cambridge, MA: Harvard University Press, 1997:166–186.

Kirsch I, Sapirstein G. Listening to Prozac but hearing placebo: A meta-analysis of antidepressant medications. In: Kirsch I. ed. Expectancy, Experience, and Behavior. Washington, DC: American Psychological Association, 1999:303–320.

Maidhof C, Dehm C, Walach H. Placebo response rates in clinical trials: A meta-analysis (abstr.). Int J Psychology 2000;35:224.

Morley S, Eccleston C, Williams A. Systematic review and meta-analysis of randomized controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. Pain 1999;80:1–13.

Rosenthal R. Meta-Analytic Procedures for Social Research. Newbury Park, CA: Sage, 1991.

Schonauer K. Semiotic Foundation of Drug Therapy: The Placebo Problem in a New Perspective. Berlin, New York: Mouton de Gruyter, 1994.

Walach H, Maidhof C. Is the placebo effect dependent on time? In Kirsch I. ed. Expectancy, Experience, and Behavior. Washington, DC: American Psychological Association, 1999:321–332.

Ward MM, Lorig KR. Patient education interventions in osteoarthritis and rheumatoid arthritis: A meta-analytic comparison with nonsteroidal antiinflammatory drug treatment. Arthritis Care Res 1996;9:292–301.

Wiesendanger H, Werthmüller L, Reuter K, Walach H. Chronically ill patients treated by spiritual healing improve in quality of life: Results of a randomized waiting-list controlled study. J Altern Complement Med 2001;7:45–51.

Address reprint requests to:
Harald Walach, Ph.D.
Department of Environmental Medicine and Hospital Epidemiology
University Hospital Freiburg
Hugstetterstraße 55
79106 Freiburg
Germany

E-mail: walach@ukl.uni-freiburg.de