

# The play of chance

John Collins, M.D.<sup>a,b</sup>

<sup>a</sup> Departments of Obstetrics and Gynecology, McMaster University, Hamilton, Ontario, and <sup>b</sup> Department of and Obstetrics and Gynecology, Dalhousie University, Halifax, Nova Scotia, Canada

Most randomized controlled trials are small relative to the clinical question they address, and chance causes more variability in the results of small trials. Thus, when small studies herald new treatment interventions, clinicians might wish to wait until the body of evidence is weighty and consistent enough to be convincing. (*Fertil Steril*® 2006;85:1364–7. ©2006 by American Society for Reproductive Medicine.)

Conscientious clinicians are always on the lookout for ways to improve their practice and benefit patients. Thus, they read articles about apparently effective new treatments with great interest, especially when the studies are well-designed, randomized controlled trials (RCTs). It is good to have RCT evidence because the evaluation of emerging technologies is often limited to studies with less valid and less convincing designs. The question for clinicians is whether to adopt the new treatment immediately or wait for further evidence. The uptake of new interventions in medical practice depends on many factors, but inherent features of the studies can help one decide whether to be among the first to adopt a new intervention.

The quality of treatment studies can be judged on several levels. The Consolidated Standards of Reporting Trials (CONSORT) statement helps investigators to describe their trials; numerous ratings of quality attempt to grade the validity of studies; and evidence-based medicine (EBM) provides guidelines for study validity, importance of the results, and relevance to patients.

The CONSORT statement lists 22 criteria that should be followed by investigators in submitting reports from randomized controlled trials for publication. The goal of the CONSORT group was to provide transparency so that readers could make informed judgments about the design, conduct, analysis, and results of trials (1). Although the CONSORT checklist can improve the reporting of RCTs and enable readers to understand a trial's conduct and to assess the validity of its results, merely ensuring that the items on the checklist are included in the report does not guarantee that the methods actually reported are adequate. Thus, guidelines such as CONSORT do not help to assess the quality of RCTs, although they might contribute to better designs if investigators consider such details when they plan their studies.

Another method of judging the validity of RCTs is the use of a quality score. Quality scores fail clinicians for two

reasons: most quality scores focus on the methods, leaving readers to judge the results and relevance (2); and there is variability among the available scoring systems. A 2002 analysis for the Agency for Healthcare Research and Quality found that among 49 scoring systems to rate the strength of scientific evidence in RCTs, only 8 systems could be used without modification, and their usefulness depended on the topic under study. Quality scores with 3 items seemed to function as well as those with 20 or more items (3).

A meta-analysis of 17 trials that compared low-molecular-weight heparin (LMWH) with standard heparin for prevention of postoperative thrombosis pinpoints the variability in quality scores (4). The investigators scored each of the 17 trials on each of 25 different published quality scoring scales and reported 2 meta-analyses for each scale: one for the high-scoring trials and one for the low-scoring trials. In the 25 pairs of meta-analyses, with 12 of the quality scales LMWH and heparin were equivalent for high-scoring and low-scoring trials; LMWH was better with high-scoring trials using 6 scales; and standard heparin was better with high-scoring trials using the remaining 7 scales. The investigators concluded that relevant methodological details of the individual trials should be identified and assessed individually.

Evidence-based medicine recognizes the clinician's dilemma by acknowledging that evidence by itself is never sufficient for a clinical decision. It also helps the clinician by providing accessible guidelines on the validity of a study's design, whether the effect is large enough to matter in practice, and how to determine whether valid, important results are relevant to one's patients (5). Evidence-based medicine is founded on clinical practice, easy to learn, and does not require statistical expertise. Despite these strengths, EBM applies to individual studies and might not be sufficient to make decisions on all of the trials (the body of evidence) about a new treatment.

Two studies from the leading edge in this month's issue of *Fertility and Sterility* illustrate the difficulty with making decisions about new treatments while evidence is accumulating (6, 7). The studies are randomized controlled trials on the use of acupuncture early in the luteal phase of assisted reproductive technology (ART) cycles; specifically, at ET

Received, October 26, 2005; revised and accepted October 26, 2005.  
Reprint requests: John Collins, M.D., 400 Mader's Cove Road, RR 1,  
Mahone Bay, Nova Scotia B0J 2E0, Canada (FAX: 902-624-0115;  
E-mail: collinsj@auracom.com).

and also 2 or 3 days later. The investigators' rationales are broadly similar, although they differ in detail, and one control group received placebo acupuncture. Compared with controls, the pregnancy rate was higher in the acupuncture group in one study and in one of two acupuncture groups in the other study.

Clinicians are growing accustomed to assessing studies by EBM criteria. The validity criteria for treatment studies include randomization, allocation concealment, balance between groups, intent to treat analysis, completeness of follow-up, and blinding (5). If the results are valid, whether the treatment effect is important depends on whether it is large enough to matter to patients. Importance is usually determined by the absolute difference or rate difference, and by the number needed to treat, which is the inverse of the rate difference. In each case, the estimate should be precise enough to be sensible. If the results are both valid and important, then are they relevant to one's practice? Are one's patients similar to those in the study, or are they so different from the study patients that they would not have been eligible?

With respect to validity, Westergaard et al. (6) were able to randomize 300 of 1,000 eligible cycles, lost 27 patients after randomization, and make no statement about allocation concealment. There were no losses to follow-up (the usual case in ART cycles), and the word "blind" does not appear in the article. The analysis was not by intention to treat because it ignored the 27 postrandomization exclusions (8). Dieterle et al. (7) randomized 225 patients from an unstated total of eligible cycles and do not mention any postrandomization losses, suggesting that randomization took place only after it was known that an ET would occur. Postrandomization losses are smaller if consent is obtained early in the ART cycle but randomization is delayed until just before the intervention (9). The allocation sequence was concealed to the physician performing the ET, and there were no losses to follow-up. The design included placebo acupuncture, but the word "blind" does not appear in descriptions of the trial. Because the analysis involved all reported patients, it was by intention to treat.

Empiric evidence indicates that reporting concealment of the allocation sequence is a key element of trial validity, although trial methodology has advanced since that 1995 report (10). In both trials, sealed envelopes were used to manage the allocation, and in one trial the allocation was concealed from the physician performing the ET (7); but neither report explicitly described how the allocation sequence was concealed from clinical personnel and investigators until interventions were assigned. Perhaps the investigators will clarify the uncertainty in their responses to this and other commentaries. What was the method used to generate the random allocation sequence? Were there any restrictions, such as blocking or stratification? What was the method used to ensure that the allocation sequence was concealed from all clinical and trial personnel?

Opinions differ regarding the relative importance of the various validity criteria, and despite the good intentions of these investigators, neither trial is flawless. They are, however, effectiveness trials conducted under the day-to-day pressure of normal clinical practice, with limited personnel and funding, a setting that typically involves concessions to perfection. On balance, their validity is probably near the average among trials with efficacy and effectiveness aims. The investigators should be justly proud: they have conceived, designed, managed, and reported clinical trials of a non-trivial intervention under difficult conditions.

What about importance? Both trials had clinical pregnancy rate as their primary outcome for sample size determination. Clinical pregnancy rate is no more than a surrogate for live birth but can be multiplied by 0.85 to approximate the live birth rate (Table 1) (11). Significance depends on the number of events, so with fewer estimated live births the lower 95% confidence limits for the differences come nearer to or cross the zero line, which is the marker for no difference. The lowest number needed to treat is seven (100/15.3), indicating that for every seven ART cycles with two perimplantation administrations of acupuncture, there might be one more live birth than in seven ART cycles receiving acupuncture that was designed not to influence fertility.

**TABLE 1**

**Estimated live birth rates and rate differences in trials of early luteal-phase acupuncture in assisted reproductive technology cycles.**

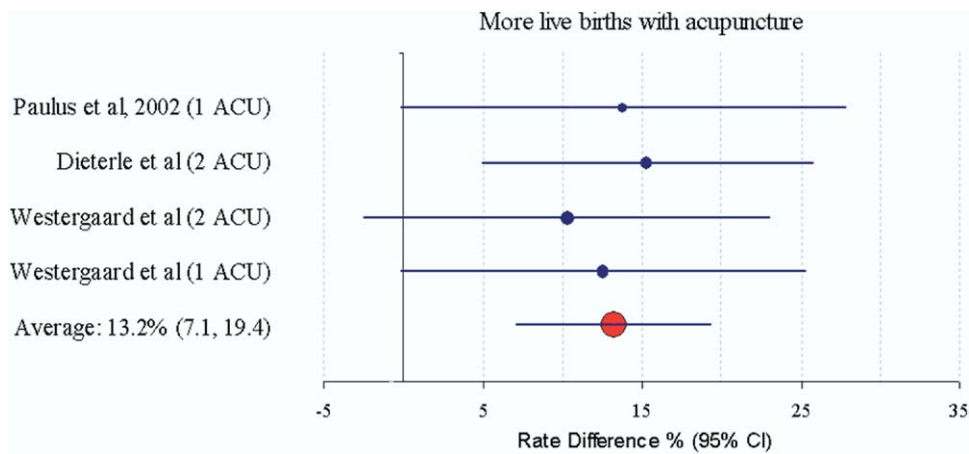
Intervention	Estimated live birth rates (%)		Rate difference (%) (95% CI)
	Experimental	Control	
ACU at ET only (6)	33.1	20.5	12.6 (−0.1, 25.3)
ACU at ET and +2 d (6)	30.8	20.5	10.3 (−2.4, 23.0)
ACU at ET and +3 d (7)	28.6	13.3	15.3 (4.9, 25.7)

Note: CI = confidence interval; ACU = acupuncture.

Collins. *The play of chance. Fertil Steril* 2006.

**FIGURE 1**

Randomized controlled trials of the effect of early luteal-phase acupuncture (ACU) (6, 7, 16). X axis: differences in estimated live birth rates (percent) with 95% confidence intervals (CI). Bubble point size is proportional to study weight (inverse of the variance). Heterogeneity  $Q = 0.374$ , (3 degrees of freedom),  $P = .94$ .



Collins. *The play of chance*. *Fertil Steril* 2006.

With respect to relevance, if the findings are both valid and important, then the results of both acupuncture trials in this issue of *Fertility and Sterility* would be relevant to most ART clinics, because there were few exclusion and inclusion criteria (6, 7).

After this EBM assessment, many clinicians would still feel that they are in a quandary about the evidence. This is the time to take a step back from the individual reports and consider the entire body of evidence on acupuncture and ART success. The scientific quality of a body of evidence depends on three factors: quality, quantity, and consistency (3). Quality is the aggregate of the quality ratings for the individual studies, ratings that reveal the extent to which bias was minimized. As noted, the quality ratings of the trials of acupuncture in the early luteal phase are probably near average. Quantity is the magnitude of the effect, the number of studies, and the sample size or power of the individual studies. Consistency for a given topic is the extent to which similar findings are reported using similar and different study designs.

It might be premature to consider consistency, given that only seven trials of acupuncture are known. There have been four trials of acupuncture for analgesia at the time of oocyte retrieval; despite the promise of the first one, three subsequent trials showed no effect on pregnancy rate (12–15). Acupuncture in the early luteal phase has been evaluated in three trials: an earlier Danish (16) trial and the present Danish and German trials (6, 7). One inconsistency in the effect on the primary outcome of clinical pregnancy demands attention, however: in one trial a single acupuncture treatment was effective, but with an added treatment the benefit was not significant, in contrast to the other trial, in

which the paired acupuncture treatments were successful (6). Therefore, it might be worthwhile to examine the quantity factor for the body of evidence on luteal-phase acupuncture.

As noted above, judgments about the quantity of a body of evidence consider the magnitude of the effect, the number of studies, and the sample size or power of the individual studies. The magnitude of the rate difference effect on live birth estimates is shown in Figure 1. The overall rate difference is 13%, and the number needed to treat is 8 (95% CI 5–14). This is a modest effect: it would take eight cycles of acupuncture to achieve a single additional live birth compared with controls.

The next point in judging quantity is the number of studies: does four studies constitute a persuasive body of scientific evidence? It might be a sufficient number for some clinicians, but there were more than 12 trials of antenatal corticosteroids for the mother before preterm delivery before that very effective treatment to prevent respiratory distress in the newborn was widely adopted (17).

The critical quantity issue is the sample size or power of the individual studies. This is critical because, although each trial is a scrupulous effort to estimate truth, each sample size estimation is an exercise in compromise between what is needed and what is possible. As a result, many—if not most—clinical trials have marginal power and are susceptible to  $\beta$  errors (missing a true difference). It is less obvious, but a small trial also is susceptible to  $\alpha$  errors (finding a significant difference just by chance) because one event more or less in either group could make a dramatic difference in the observed  $P$  values. For example, the clinical pregnancy rates per transfer in the placebo groups of the

luteal acupuncture trials are all a few events lower than the average clinical pregnancy rates per transfer generated from European registers for 1999 (16%–24% compared with 28%) (18).

To envision the effect of trial size, consider a target with truth at the bull's eye. The results of large trials would cluster closely around truth—their size allows them to make a precise estimate of truth. The results of small trials, affected more by the play of chance, will be somewhat scattered. If the methods are valid, there should be little bias, and the scatter would have truth at its center. The trouble is that looking at the results of just a few small studies does not reveal where the center lies.

Is it fair to categorize the luteal acupuncture trials as small? The previously published trial enrolled 80 patients per group and does not report how this sample size was determined or the associated significance and power assumptions (16). The power analysis in the present Danish study indicated that evaluation of an 11% increase over a 25% baseline pregnancy rate would require 100 patients in each of three groups; the significance and power assumptions are not mentioned (6). The German trial plan assumed a clinical pregnancy rate of 20% in the control group, a minimal detectable difference of clinical pregnancies between study group and control group of 15% at a power of 80% ( $\beta$  20%), and a type I error ( $\alpha$ ) of 5% (7). That plan assumed a one-sided test situation; either the intervention or the control could be superior, however, which makes this  $\alpha$  equivalent to 10%.

What is the minimum sample size needed to evaluate whether early luteal-phase acupuncture can improve live birth rates in ART cycles? With standard assumptions for power (80%) and significance (two-tailed  $\alpha$  of 5%) and an uncorrected chi-square test, to distinguish between a standard live birth rate of 20% per cycle initiated and the 13% higher summary rate with acupuncture in Figure 1, 180 patients per group would be needed (19). If all else holds but the standard live birth rate is 25%, the number would be 199 per group. As clinicians might guess, the sample size varies inversely with the expected difference. For planning purposes, it would be wise to obtain consent from an additional 10% of patients to cover the losses between the day of consent and the day of randomization and ET.

So, what should ART clinicians do about incorporating acupuncture into their practices? For all we know, there might be a true benefit from acupuncture, and delay in acceptance would not benefit patients. Thus, some clinicians who are exceptionally receptive to new information will begin to explore the logistics and cost. Fortunately, adverse effects seem to be minimal. For most clinicians, however, the right course remains uncertain: the trials might represent an accidental collection of  $\alpha$  errors. More studies are needed, and there is nothing new about that observation. Given the difficulties of mounting trials in clinical practice, future trials

might be no larger than the current ones. With more than 300,000 ART cycles performed every year, however, surely more studies involving several hundred cycles should not be a problem.

## REFERENCES

1. Moher D, Schultz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191–4.
2. Greenland S. Quality scores are useless and potentially misleading: reply to “Re: A critical look at some popular analytic methods.” *Am J Epidemiol* 1994;140:300–2.
3. West S, King V, Carey TSKN, McKoy N, Sutton SF, Lux L. Systems to rate the strength of scientific evidence [report 47]. Rockville, MD: Agency for Healthcare Research and Quality, 2002.
4. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054–60.
5. The Evidence-Based Medicine Working Group. Users' guides to the medical literature: a manual for evidence-based clinical practice. Chicago: AMA Press, 2002:5.
6. Westergaard LG, Mao Q, Kroglund M, Sandrini S, Lenz S, Grinsted J. Acupuncture on the day of embryo transfer significantly improves the reproductive outcome in infertile women: a prospective, randomized trial. *Fertil Steril* 2006;85:1341–6.
7. Dieterle S, Ying G, Hatzmann W, Neuer A. Effect of acupuncture on the outcome of in vitro fertilization and intracytoplasmic sperm injection: a randomized, prospective, controlled clinical study. *Fertil Steril* 2006;85:1347–51.
8. Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. *BMJ* 2002;325:652–4.
9. Arce JC, Nyboe AA, Collins J. Resolving methodological and clinical issues in the design of efficacy trials in assisted reproductive technologies: a mini-review. *Hum Reprod* 2005;20:1757–71.
10. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–12.
11. Grimes D, Schulz KF. Surrogate end points in clinical research: hazardous to your health. *Obstet Gynecol* 2005;105:1114–8.
12. Stener-Victorin E, Waldenström U, Nilsson L, Wikland M, Janson PO. A prospective randomized study of electro-acupuncture versus alfentanil as anaesthesia during oocyte aspiration in in-vitro fertilization. *Hum Reprod* 1999;14:2480–4.
13. Stener-Victorin E, Waldenström U, Wikland M, Nilsson L, Hagglund L, Lundeberg T. Electro-acupuncture as a peroperative analgesic method and its effects on implantation rate and neuropeptide Y concentrations in follicular fluid. *Hum Reprod* 2003;18:1454–60.
14. Humaidan P, Stener-Victorin E. Pain relief during oocyte retrieval with a new short duration electro-acupuncture technique—an alternative to conventional analgesic methods. *Hum Reprod* 2004;19:1367–72.
15. Gejervall AL, Stener-Victorin E, Moller A, Janson PO, Werner C, Bergh C. Electro-acupuncture versus conventional analgesia: a comparison of pain levels during oocyte aspiration and patients' experiences of well-being after surgery. *Hum Reprod* 2005;20:728–35.
16. Paulus WE, Zhang M, Strehler E, El-Danasouri I, Sterzik K. Influence of acupuncture on the pregnancy rate in patients who undergo assisted reproduction therapy. *Fertil Steril* 2002;77:721–4.
17. Crowley P, Chalmers I, Keirse MJ. The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. *Br J Obstet Gynaecol* 1990;97:11–25.
18. Nygren KG, Andersen AN. Assisted reproductive technology in Europe, 1999. Results generated from European registers by ESHRE. *Hum Reprod* 2002;17:3260–74.
19. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. *Control Clin Trials* 1990;11:116–28.