Journal of Andrology, Vol. 22, No. 6, November/December 2001

Is the Randomized Clinical Trial the Gold Standard of Research?

Andrology Lab Corner

STEPHEN D. SIMON

From the Department of Medical Research, Children's Mercy Hospital, Kansas City, Missouri.

There's a story I heard long ago about research gone bad. If it isn't a true story, it should be. Some researchers wanted to study how the concentration of how a particular water pollutant affected the mortality of guppies. The researchers had a big tank with 100 guppies. They wanted to allocate these guppies equally across 5 smaller tanks. The first 20 guppies that they caught went into the first tank, the next 20 went into the second tank, and so forth, until the last 20 guppies went into the last tank. Each tank got a different level of the pollutant.

To the researchers' surprise and horror, mortality was related not to concentration, but to the order in which the tanks were filled. The first tank, which had all the sickly, slow-moving, and easy-to-catch guppies, had the highest mortality rate. The last tank, which had the vigorous, fast-moving, and hard-to-catch guppies, had the lowest mortality rate.

The randomization in a typical randomized clinical trial (RCT) performed today would have prevented this research disaster. Randomization does not mean alternating between treatment and control (ABAB . . .); rather, it involves a random device (eg, a coin or a die) or a series of random numbers for assigning patients in a pattern that is inherently unpredictable.

Alternating assignment can cause serious problems. We can see this by examining how plants grow in a garden. A row of cabbages, for example, will often have a pattern of big cabbage, little cabbage, big cabbage, little cabbage, etc, especially when the cabbages are planted a bit too close together. Competition for resources causes this pattern. One cabbage may get a head start on its neighbor and extend its roots a bit farther. It steals some of the nearby water and soil nutrients and grows fast at the expense of its neighbor. So, an experiment that gave a treatment to the even-numbered cabbages and used the odd-numbered cabbages as controls would likely lead to a spurious finding. Colditz et al (1989) provide empirical evidence that alternating allocation leads to a systematic bias in the outcome measure.

A fascinating examination of randomization gone bad appears in Marks and Colwell (2000). This paper criticizes research into the staring effect, the belief that people can perceive when they are being stared at. Many people can recall incidents when they could tell that they were being stared at, even though the person staring at them was outside their field of vision. To test this rigorously, though, would require a random series of trials in which people would try to identify when they were being stared at and when they were not being stared at. This is an experiment simple enough to be run in a school classroom.

Rupert Sheldrake, a PhD trained biologist who is currently a fellow of the Institute of Noetic Sciences, has packaged and promoted such an experiment. Feedback is a critical part of this experiment; subjects are told whether their guess in one trial is correct before the next trial starts. With feedback, the issue of randomization becomes very important; an alternating ABAB sequence would easily be spotted. So, Dr Sheldrake provided some randomized sequences in the research package that he has promoted. Unfortunately, these sequences deviated significantly from perfect randomness. The sequences tend to oscillate too often. For example, the frequency of 3 consecutive staring trials or 3 consecutive nonstaring trials is far less than could be expected by chance.

A careful replication of this experiment (Colwell et al, 2000) showed three things. First, a trial without feedback with Dr Sheldrake's randomization sequences showed no staring effect. In contrast, a trial with feedback with those same sequences showed a significant staring effect. Finally, a trial with feedback but with a properly randomized sequence showed no staring effect. Thus, research supporting a staring effect probably was artifactually produced by the conscious or subconscious detection of patterns in the flawed randomization sequence.

Covariate Imbalance

In a study without randomization, look carefully for covariate imbalance. A covariate is a variable that is usually not of direct research interest, but one that needs attention because it may be related to one of the outcome measures.

Correspondence to: Dr Stephen D. Simon, Department of Medical Research, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, Missouri 64108.

Received for publication June 22, 2001; accepted for publication June 22, 2001.

Covariate imbalance is a tendency for that variable to be larger or smaller in one group compared with the other group(s). For example, in a study of reproductive function among military personnel (Schrader et al, 1998), those personnel involved with artillery (firing the 155mm howitzer) had much lower average smoking levels than the other two comparison groups. It is hardly surprising to see less smoking among those who work regularly with explosives, but this covariate imbalance requires some statistical adjustment.

Unfortunately, the adjustments for covariate imbalance are tricky, especially when the covariate is measured crudely. Chen (1999) examines the relationship between maternal smoking and Down syndrome. Maternal age is an important covariate, but when it is measured as a dichotomy (younger or older than 35 years), the adjustment produces a negative association between smoking and Down syndrome. Adjustment using the exact year of maternal age, on the other hand, produces a more accurate picture that shows no association between smoking and Down syndrome.

Furthermore, we can adjust only for those covariates that can be measured. Many factors, such as the patient's psychological state, initial severity of disease, and the presence of comorbid conditions are difficult or impossible to measure, but these factors can have an important effect on the outcome measures. Randomization ensures approximate balance for both measurable and unmeasurable covariates.

Even in a study with randomization, covariate imbalance may sometimes be encountered. Random numbers are not perfect at balancing things out, just like flipping a coin 20 times will not always guarantee exactly 10 heads and 10 tails. When assessing covariate imbalance in a randomized study, Altman (1985) emphasizes the need to examine the clinical relevance rather than the statistical significance of the covariate imbalance. Matching on important covariates can prevent this imbalance, but matching often has logistical difficulties. There is another approach, minimization, which also prevents the chance occurrence of covariate imbalance. Patients are not assigned randomly, but are preferentially assigned to one group or another to reduce any covariate imbalance that may have crept into the design (Treasure and MacRae, 1998).

There are indeed quantifiable advantages to using randomization (Chalmers et al, 1983; Schulz 1996). Studies with randomization show fewer problems with covariate imbalance and tend to have smaller effects than nonrandomized studies. There is, however, conflicting evidence. Two recent studies (Benson and Hartz, 2000 and Concato et al, 2000) show that a well-designed study without randomization can be comparable to studies with randomization. To confuse the issue further, there is some dispute about the significance of these two recent studies (Ioannidis et al, 2001). In my opinion, recent advances in the statistical design and analysis of nonrandomized studies allow the best of these studies to have a level of evidence that is comparable to a well-designed randomized study.

We have probably not heard the last in this debate. But one thing is clear: studies that use randomization remove covariate imbalance as a possible source of doubt and uncertainty about the validity of the research findings.

Blinding in a Randomized Study

Blinding involves the use of some type of placebo to hide information about treatment status from the patients and from those treating and evaluating the patient. The placebo could be a sugar pill when evaluating a new medicine. When the treatment involves a procedure, a placebo would represent some type of sham procedure that appears the same to the patient.

An example of this sham procedure was used by Bullock et al (1989), wherein a single, blind, randomized trial was used to examine acupuncture as a treatment for alcoholism. Blinding an acupuncture study is indeed challenging. This research used needles for both the treatment and control groups, but in the control group, the needles were placed in an incorrect location that was within 5 mm of the correct placement. This study was only a single blinded study, because the acupuncturists knew who was receiving the correct treatment. Although this study showed that acupuncture was effective, Sampson (1997) criticized the research design. There was ample opportunity for the control subjects to pick up clues about their status. The greater than 90% dropout rate among the controls was evidence that the study was not adequately blinded.

No study is perfectly blinded; the person who fills the prescription order for the active medication or sugar pill will know who is getting what. But the study would be considered adequately blinded as long as the patient and everyone evaluating the patient and everyone having substantive interactions with the patient are kept unaware of who got what.

Blinding provides two benefits. First, it ensures a high level of objectivity in measurements. A researcher who is blind to the treatment status cannot be accused of conscious or unconscious attempts to influence the evaluation of research subjects. This is particularly important when the outcome is subjective, such as a quality-of-life measurement or an assessment of symptom relief (Johnson and Dixon, 1997). Second, blinding removes the possibility of a placebo effect creating an artifactual response in the treatment group.

There is some recent evidence that the placebo effect is not all it is cracked up to be. A systematic review of 130 trials that compared a placebo arm with a no-treat-

ment arm showed mixed results (Hrobjartsson and Gotzsche, 2001). The authors divided the studies into those with an objective binary primary outcome, with a subjective binary outcome, with an objective continuous outcome, and with a subjective continuous outcome. Only in the last case did the placebo show a statistically significant effect. Even here, the effect was rather small, roughly one-third of a standard deviation (95% CI, −0.47 to −0.25).

It is indeed possible that many of the effects that we ascribe to the placebo effect may actually represent something else. For example, some diseases are cyclical and patients are more likely to join a clinical trial at a trough. Thus, any short-term improvement represents the natural course of the disease and not a placebo effect.

Even without a placebo effect, their use is still important. Placebos can minimize or prevent the disparities in dropout rates, patient reporting of beneficial and harmful events, and physician assessments of outcome (Hrobjartsson and Gotzsche, 2001).

Blinding is often not possible for a surgical trial. Patients undergoing an orchiectomy, for example, will notice that something is missing sooner or later. When blinding is possible, it can be controversial. A placebo/sham operation used in study of Parkinson disease (described in Freeman et al, 1999) was criticized on the grounds that it subjected the control group to an unnecessary risk (Macklin, 1999).

When blinding is not possible, the researchers should try to ensure that those evaluating the patients are unaware of treatment status (Johnson and Dixon, 1997). This does not prevent a placebo effect, but it does ensure that the outcome measures are evaluated without any conscious or subconscious bias.

Blinding is often overlooked as a way to add credibility to laboratory studies. Even though these studies usually have objective outcome measures, the use of blinding will prevent the possibility of or the suspicion of differential handling of control and treatment samples.

Generalizability

RCTs are often problematic because they exclude important segments of the population, in particular, the elderly and women (Gurwitz et al, 1992; Bugeja et al, 1997). Sometimes these exclusions are justified. Reproductive studies, in particular, would usually have ample justification for excluding elderly subjects or for focusing just on men. But when the exclusions are unjustified, however, we are limited in how far we can extrapolate the results.

RCTs often study narrowly targeted groups to improve the efficiency and precision of the research. When a study restricts its sample to those who visit a fertility clinic, for example, it is studying a highly select group of patients that differs in many important ways from the general public. The extent to which the findings of a study can be extrapolated or generalized is often referred to as external validity. Indeed, difficulty in extrapolating results from RCTs represent their biggest drawback.

In a further effort to reduce variation, an RCT may exclude ''troublesome'' subjects, such as those who cannot speak English, those who cannot read English, those who are taking other medications, those who are likely to move out of town before the study ends, those who are unlikely to comply with a complex intervention, or those with comorbid conditions. The precision gained by each exclusion comes at a price: a further limitation on generalizability. For example, RCTs will often screen out patients that demonstrate during a pretesting phase that they are incapable of complying with a complex intervention (eg, Adkinson et al, 1997). But what physician has the option of treating only compliant patients?

The flip side of the coin is a study that uses a narrowly defined sample for the sake of convenience. The classic example of this is the use of a student in a Psychology 101 class as the white rat of choice for all psychological studies. But there are other examples as well. Studying a disease in a hospitalized population may be easy, but when and whether a patient arrives at a hospital depends largely on many patient and physician preferences, and is strongly influenced by insurance status and socioeconomic levels. Thus a sample of hospitalized patients is unlikely to be representative of all patients who have a certain disease (Ellenberg, 1994). Thornley and Adams (1998) cite the overuse of institutionalized patients as a major limitation of much of the research into schizophrenia.

The setting of the research can also sometimes limit its usefulness. Kippax and Van de Ven (1998) note a commonly cited study that touted the success of a 100% condom use program in reducing the incidence of HIV. This program, however, was conducted in Thailand, a country with a highly institutionalized sex industry that makes it quite different from most other countries.

Ethical considerations require us to use only volunteers for RCTs, and this can also lead to problems with generalizability. People who volunteer for research studies often differ in important ways from those who do not volunteer. This is often called selection bias or volunteer bias.

A study that involves an extended stay in a clinic, for example, may be more likely to attract unemployed patients. A study that involves a free physical examination (see below for an example) is likely to attract people who are worried that they may be sick. A study that involves a new therapy may preferentially attract volunteers who are dissatisfied with the current intervention (Gotszche, 1990). One of the best examples of a study that had an unusual set of volunteers is by Wilson (1984). That study

used sublingual administration of boiled and unboiled urine as a treatment.

There are examples of volunteer bias that are specific to reproductive studies. A study that required patients to produce a semen sample would exclude patients with certain religious beliefs about masturbation. The free semen analysis provided by many reproductive studies can skew the sample. Such an incentive would be likely to attract men who were thinking about starting a family or who were already trying to start and were encountering some difficulty.

There are two good studies that provide quantitative evidence of how volunteers can differ from the general population. Gustavsson et al (1997) describe the possible relationship between serotonin turnover and disinhibitory and self-destructive behavior. The way to examine this relationship would be to obtain cerebrospinal fluid in healthy volunteers using a lumbar puncture. Many people will refuse to participate in such a study because of the pain and risk involved. Fortunately, in this study, volunteers were sought out from a panel of subjects who had recently completed several psychological tests. The 39 subjects who volunteered for a lumbar puncture scored higher on an impulsivity scale than the 48 subjects who refused to participate. The authors cite this difference as making interpretation of the research findings problematic.

As a second example, Chen et al (1997) screened for individuals with a genetic deficiency in CYP2D6 expression. In a group of 188 patients recruited by clinical contract laboratories, only 2 (1.1%) showed this deficiency. In a group of 142 subjects recruited from within and around the University of Kentucky, 9 (6.3%) showed the same deficiency. The latter percentage was close to other reported values. The sample from the contract laboratories included a large number of subjects who were part of an established database of patients who had volunteered for previous drug studies. Because individuals deficient in CYP2D6 expression are more likely to experience adverse drug effects, these individuals may be underrepresented in these databases. This implies that a group of healthy volunteers who participate in several drug studies may represent a group that has fewer problems with adverse drug reactions than the general population.

If volunteer bias is a concern, it is often possible to obtain some demographic information about those who refused to volunteer, as in the Gustavsson et al (1997) study. This information could be restricted to just a sample of the nonvolunteers and still provide useful information (Ellenberg, 1994).

Subversion of Randomization

Randomization implies that the researcher has complete control over treatment assignment; neither the patient nor the physician has any say in the matter. When the patient, physician, or both are able to influence treatment assignment, then no longer is there an RCT. Patients can influence their treatment assignments by withdrawing from the study or by not complying with the treatment (Jurs, 1971).

If possible, analyze the dropouts and noncompliant patients as if nothing had happened (ie, treat their data values in the same way as those who stayed on the study and complied with the assigned treatment). This approach, often called intention to treat analysis, seems counterintuitive and may bias the findings. Typically, the dropouts and noncompliant patients dilute the effect of the treatment in an intention to treat analysis (Ellenberg, 1994). Still, the bias is less troublesome than the bias caused by other types of analyses.

Excluding noncompliers will usually result in a serious bias in the results. Noncompliant patients tend to be sicker and have poorer self-care than compliers. In fact, patients who are noncompliant with a placebo have been shown to have worse outcomes than patients who comply with the placebo therapy (Freedman et al, 1998, page 14).

As an example of the problems with excluding noncompliant patients, consider a study that compares a surgical intervention with a nonsurgical intervention. In some cases, the patients may die prior to receiving surgery, which is an extreme example of noncompliance. Excluding these patients from the analysis will cause serious biases. The rapidly dying patients are being excluded from the surgical arm of the study but not from the nonsurgical arm.

Dropouts, of course, are harder to handle than noncompliant patients. A change in serum testosterone levels cannot be analyzed in a patient who is not around to provide the second blood sample. In some studies, a plausible assumption can be made that someone who has dropped out (ie, labeling as a smoker someone who stops showing up at smoking cessation program). In other situations, there may be an intermediate value that can be substituted for the long-term value (Lasky, 1962). Neither of these approaches is ideal. Researchers should take aggressive efforts to minimize the number of dropouts (Crider, 1971) and should try to establish that there is no covariate imbalance between dropouts and those who complete the study.

The doctors who recruit patients into RCTs can also sometimes subvert randomization (Schulz, 1996). They may consciously or subconsciously apply exclusion criteria differently if they know which treatment group a patient is going to be assigned to. The physician may also try to steer the patient into a different arm of the study by delaying entry of the subject into the RCT. These problems are one reason to avoid an alternating (ie, ABAB . . .) assignment. A good RCT should use concealed allocation, an approach in which information about treatment assign-

ment is hidden until after the physician has determined that a patient is eligible for a research study (Schulz, 1996). Concealed allocation can use a series of sealed envelopes or a centralized telephone randomization service, though the former was shown to be subverted by a group of physicians in a multicenter trial, leading to a large age discrepancy between the treatment and control groups (Kennedy and Grant, 1997).

The researchers themselves can subvert randomization through subgroup analysis. Subgroup analysis involves the selective analysis of a subgroup of patients (eg, showing that a fertility treatment is effective, but only for nonsmokers). Because the researchers, in most cases, did not randomize within the subgroup, these findings need to be interpreted with caution. Subgroup analyses should not be ignored, but they do need to be interpreted with care (Buyse, 1989; Freemantle, 2001). There are good examples of when a subgroup analysis found valuable information, such as when Byar and Green (1980) showed that estrogenic therapy was a beneficial therapy for prostate cancer in patients without cardiovascular disease, but that it was harmful as a treatment for patients with cardiovascular disease. On the other hand, a subgroup analysis led to the false and later overturned conclusion that streptokinase was an effective treatment for suspected myocardial infarction only if it was administered within 6 hours of the onset of pain (Oxman and Guyatt, 1992).

Randomized Trials for Complex Interventions

RCTs may not be the best approach for evaluating a complex intervention. Kippax and Van de Ven (1998) suggest that RCTs are not the best way to evaluate the effectiveness of HIV health promotion. They argue that the restrictive nature of clinical trials forces researchers to design interventions that are amenable to this type of research. This leads to an oversimplification of interventions that focus too much on the individual and ignores the larger societal context of the problem. In contrast, Oakley (1998) argues that RCTs have been used successfully to evaluate social interventions such as negative income tax programs for the poor and postrelease and job assistance schemes for ex-prisoners. Arguing a middle position are Campbell et al (2000), who state that a mixture of RCTs and other research designs are needed to fully assess complex interventions.

Some research problems cannot be adequately addressed by a randomized trial. For example, Butler et al (1998) use a qualitative research approach to elicit information about the attitudes of smokers to the advice that their physicians give them about quitting. It is hard to imagine how to obtain information about this topic in an RCT.

Interpretation Bias

Finally, even when the RCT is objective, our interpretations of it often are not. McCormack and Greenhalgh (2000) write that, ''interpretations of clinical trials are often neither objective nor value-free,'' and cite as an example a major diabetes study in which their interpretation is much more limited than that provided by the original authors (UKPDS, 1998). Given the complexity of the study (2 treatment arms compared with a control, and multiple end points including microvascular effects, macrovascular events, deaths related to diabetes, all causes of mortality, and a surrogate marker of disease), it is hardly surprising to find such disagreements. Also adding to the confusion is the sad fact that quite often, 2 RCTs of the same problem will lead to opposite conclusions (Horwitz 1987; Furukawa et al, 2000).

Summary

All research has flaws. Some flaws are so trivial that the research can still stand as the definitive study. Other flaws prevent a study from being definitive, but the study still provides useful guidance in the context of other research. Some flaws are so serious that the research provides no useful information at all. The tricky part is not finding flaws in the research but in deciding to what extent the flaws erode the credibility of the research.

In general, the use of RCTs can add substantial credibility to a research study. There are calls for greater use of RCTs in many areas, such as surgery (Baum, 1999) and psychiatry (Andrews, 1999). Of course, nonrandomized trials are an important complement to RCTs when the latter are ethically inappropriate or logistically impossible (Black, 1996).

Failure to use randomization or blinding, however, is not a fatal flaw. Furthermore, the artificial nature of RCTs will often restrict their applicability to overly simple interventions. When RCTs focus on narrow patient groups or exclude important segments of the population, there may be difficulty in generalizing their results. So it would be a mistake to label the RCT as a gold standard for all research. A silver standard may be a more appropriate label.

References

Adkinson NF Jr, Eggleston PA, Eney D, et al. A controlled trial of immunotherapy for asthma in allergic children. *New Engl J Med.* 1997; 336:324–331.

Altman DG. Comparability of randomised groups. *Statistician.* 1985;34: 125–136.

Andrews G. Randomised controlled trials in psychiatry: important but poorly accepted. *BMJ.* 1999;319:562–564.

Baum M. Reflections on randomised controlled trials in surgery. *Lancet.* 1999:355(suppl 1):SI6–SI8.

Benson K, Hartz AJ. A Comparison of observational studies and randomized, controlled trials. *New Engl J Med.* 2000;342:1878–1886.

Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ.* 1996;312:1215–1218.

Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. *BMJ.* 1997; 315:1059.

Bullock ML, Culliton PD, Olander RT. Controlled trial of acupuncture for severe recidivist alcoholism. *Lancet.* 1989;335:1435–1439.

Butler C, Pill R, Stott NC. Qualitative study of patients' perceptions of doctors' advice to quit smoking: implications for opportunistic health promotion. *BMJ.* 1998;316:1878–1881.

Buyse ME. Analysis of clinical trial outcomes: some comments on subgroup analyses. *Control Clin Trials.* 1989;10(suppl 4):187S–194S.

Byar DP, Green SB. The choice of treatment for cancer patients based on covariate information. *Bull Cancer.* 1980;67:477–490.

Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P. Framework for design and evaluation of complex interventions to improve health. *BMJ.* 2000;321:694–696.

Chalmers T, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trails. *N Engl J Med.* 1983;309:1358–1361.

Chen S, Kumar S, Chou WH, Barrett JS, Wedlund PJ. A genetic bias in clinical trials? Cytochrome P450–2D6 (CYP2D6) genotype in general vs selected healthy subject populations. *Br J Clin Pharmacol.* 1997; 44:303–304.

Chen CL, Gilbert TJ, Daling TJ. Maternal smoking and Down syndrome: the confounding effect of maternal age. *Am J Epidemiol.* 1999;149: 442–446.

Colditz G. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med.* 1989;8:441–454.

Colwell J, Schroder S, Sladen D. The ability to detect unseen staring: a literature review and empirical tests. *Br J Psychol.* 2000;91:71–85.

Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000;342:1887–1892.

Crider D. Tracking respondents in longitudinal surveys. *Public Opinion Q.* 1971;35:613–620.

Ellenberg JH. Selection bias in observational and experimental studies. *Stat Med.* 1994;13:557–567.

Freedman DR, Pisani R, Purves R. *Statistics.* 3rd ed. New York, NY: WW Norton; 1998.

Freeman TB, Vawter DE, Leaverton PE, Godbold JH, Hauser RA, Goetz CG, Olanow CW. Use of placebo surgery in controlled trials of a cellular-based therapy for Parkinson's disease. *New Engl J Med.* 1999; 341:989–992.

Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ.* 2001;322:989–991.

Furukawa T, Streiner DL, Hori S. Discrepancies among megatrials. *J Clin Epidemiol.* 2000;53:1193–1199.

Gotzsche P. Bias in double-blind trials. *Dan Med Bull.* 1990;37:329–336.

Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA.* 1992;268: 1417–1422.

Gustavsson JP, Asberg M, Schalling D. The healthy control subject in psychiatric research: impulsiveness and volunteer bias. *Acta Psychiatr Scand.* 1997;96:325–328.

Horwitz RI. Complexity and contradiction in clinical trial research. *Am J Med.* 1987;82:498–510.

Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New Engl J Med.* 2001;344:1594–1602.

Ioannidis JPA, Haidich A, Lau J. Any casualties in the clash of randomised and observational evidence? *BMJ.* 2001:322;879–880.

Johnson AG, Dixon JM. Removing bias in surgical trials. *BMJ.* 1997; 314:916.

Jurs S. The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *J Exp Educ.* 1971;40:62–66.

Kennedy A, Grant A. Subversion of allocation in a randomised controlled trial. *Control Clin Trials.* 1997;18(suppl 3):S77–S78.

Kippax S, Van de Ven P. An epidemic of orthodoxy? Design and methodology in the evaluation of the effectiveness of HIV health promotion. *Crit Public Health.* 1998;8:371–386.

Lasky J. The problem of sample attrition in controlled treatment trials. *J Nerv Ment Dis.* 1962;135:332–338.

Macklin R. The ethical problems with sham surgery in clinical research. *New Engl J Med.* 1999;341:992–995.

Marks DF, Colwell J. The psychic staring effect. An artifact of pseudo randomization. *Skeptical Inquirer.* 200;24:41–44,49.

McCormack J, Greenhalgh T. Seeing what you want to see in randomised controlled trials: versions and perversions of UKPDS data. *BMJ.* 2000;320:1720–1723.

Oakley A. Experimentation and social interventions: a forgotten but important history. *BMJ.* 1998;317:1239–1242.

Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med.* 1992;116:78–84.

Sampson W. Inconsistencies and errors in alternative medicine research. *Skeptical Inquirer.* 1997;21:35–38.

Schrader SM, Langford RE, Turner TW, et al. Reproductive function in relation to duty assignments among military personnel. *Reprod Toxicol.* 1998;12:465–468.

Schulz KF. Randomised trials, human nature, and reporting guidelines. *Lancet.* 1996;348:596–598.

Thornley B, Adams C. Content and quality of 2000 controlled trials in schizophrenia over 50 years. *BMJ.* 1998;317:1181–1184.

Treasure T, MacRae K. Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. *BMJ.* 1998;317:362–363.

UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet.* 1998;352:837–853.

Wilson CW. The protective effect of auto-immune buccal urine therapy (AIBUT) against the Raynaud phenomenon. *Med Hypotheses.* 1984; 13:99–107.

Andrology Lab Corner welcomes the submission of unsolicited manuscripts, requested reviews, and articles in a debate format. Manuscripts will be reviewed and edited by the Section Editor. All submissions should be sent to the ***Journal of Andrology*** Editorial Office. Letters to the editor in response to articles as well as suggested topics for future issues are encouraged.