ORIGINAL ARTICLE

Alternatives to Placebo-Controlled Trials

David L. Streiner

ABSTRACT: Until recently, the gold standard for assessing the efficacy and effectiveness of new medications has been the placebo-control randomized clinical trial (RCT). However, there are serious ethical concerns about placing patients on a placebo when effective treatments exist. Further, if a new agent is tested only against a placebo, there is no guarantee that it is more effective, or even as effective, as an existing agent. For these and other reasons, ethicists and regulatory bodies have said that, under these circumstances, new drugs should be tested against an active agent. There are three types of such trials: superiority, equivalence, and non-inferiority. In superiority trials, the goal is to establish that the new drug is better (i.e., more effective, or with a more benign side-effect profile) than the standard. Because such trials require much larger sample sizes than placebo-control studies, and are rarely required to bring a drug onto market, they are rarely done. In equivalence trials, the aim is to show that the new and standard agents have similar degrees of effectiveness or adverse events. Due to sample size requirements, most studies of new drugs are non-inferiority trials, in which it is sufficient to demonstrate that the new drug is not significantly worse than the existing ones. However, there are methodological concerns with equivalence and non-inferiority trials, including (a) an inability to determine if the drugs were equally good or equally bad; (b) poorly executed trials with low power can be mistaken for "proving" equivalence or non-inferiority; (c) the equivalence interval is arbitrary; (d) successive non-inferiority trials may lead to a gradual reduction in effectiveness; and (e) often larger trials are necessary. The paper also discusses "add on trials." It is recommended that, even when existing drugs exist, trials should consist of at least three arms, one of which is a placebo. This paper briefly considers the ethics of placebo, and conditions are stated under which such studies can be conducted.

RÉSUMÉ: Alternatives aux essais contrôlés par placebo. Jusqu'à tout récemment, l'étalon or pour évaluer l'efficacité potentielle et l'efficacité réelle de nouveaux médicaments était l'essai clinique randomisé contrôlé par placebo. Cependant, ceci soulève de sérieuses préoccupations éthiques quand il existe un traitement efficace. De plus, si un nouveau médicament est évalué seulement par rapport à un placebo, rien ne garantit qu'il est plus efficace ou aussi efficace qu'un médicament dont l'efficacité est déjà établie. Tenant compte de ce fait, mais aussi pour d'autres raisons, les éthiciens et les organismes de réglementation ont stipulé que les nouveaux médicaments devraient être comparés à un médicament dont l'efficacité est établie. Il existe trois types d'essais de comparaison : de supériorité, d'équivalence et de non-infériorité. Dans les essais de supériorité, le but est d'établir que le nouveau médicament est meilleur (c'est-à-dire plus efficace ou ayant un profil d'effets secondaires plus bénin) que le médicament standard. Comme ces essais requièrent des échantillons de taille beaucoup plus considérable que les études contrôlées par placebo et qu'elles sont rarement requises pour obtenir l'autorisation de mise en marché, elles sont rarement effectuées. Dans les essais d'équivalence, le but est de démontrer que le nouveau médicament et le médicament standard ont le même degré d'efficacité ou un profil d'effets secondaires similaire. À cause de la taille de l'échantillon requis, la plupart des études sur les nouveaux médicaments sont des essais de non-infériorité où il est suffisant de démontrer que le nouveau médicament n'est pas significativement moins bon que les médicaments existants. Cependant, les essais d'équivalence et de non-infériorité soulèvent des questions méthodologiques soit : a) l'incapacité de déterminer si les médicaments sont également bons ou également mauvais; b) les essais mal exécutés ayant une faible puissance peuvent errer en « prouvant » l'équivalence ou la non-infériorité; c) l'intervalle d'équivalence est arbitraire; d) des essais de non-infériorité successifs peuvent entraîner une réduction graduelle de l'efficacité réelle des études et e) il est souvent nécessaire d'utiliser des échantillons de plus grande taille. Cet article discute également des « essais complémentaires». Il est recommandé que les essais comportent au moins trois bras dont un bras placebo, même quand il existe des médicaments dont l'efficacité est reconnue. Cet article considère brièvement l'éthique de l'utilisation du placebo et précise les conditions à remplir pour procéder à ces essais.

Can. J. Neurol. Sci. 2007; 34: Suppl. 1 - S37-41

It is generally accepted that the strongest research design to determine the effectiveness and efficacy of an intervention is the randomized controlled trial (RCT),[1] in which patients who meet the inclusion criteria are assigned at random to receive either the new, investigational treatment or to one or more comparison groups. Until recently, the comparison group has almost always been a placebo. However, the use of placebos when proven therapies exist has come under increasing criticism; in particular for possibly violating the principles of the Declaration of Helsinki.[2] In this paper, I will first discuss why the placebo controlled trial is so common, and then discuss some alternatives to placebo controls, with their advantages and disadvantages. I conclude by discussing some ethical implications.

From the Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada.
RECEIVED OCTOBER 7, 2005. ACCEPTED IN FINAL FORM OCTOBER 10, 2006.
Reprint requests to: David L. Streiner, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, Ontario, M6A 2E1, Canada.

The Placebo-Control Trial

The advantages of the RCT are many, and include balancing of baseline risks, reducing the possible effects of confounding variables (most especially those we are unaware of), and satisfying the criteria for statistical tests.[3] There are few ethical problems in using placebo-control trials when there is no treatment of proven effectiveness for the condition being studied or when the current treatment has serious side effects (but see 4). We may suspect (or hope) that the treatment being investigated is better than placebo, but in the absence of any previous evidence, this conjecture remains unproven, so that the patients in the placebo arm of the trial are not being denied therapy that the experimental group is receiving. There are two major advantages to using a placebo control: (a) efficiency, and (b) clear evidence of effectiveness or the lack of it.

Efficiency means that we can achieve statistical significance with the smallest number of participants in the trial. The alternatives to the placebo controlled trial, which I will discuss shortly, very often require a larger sample size to show a statistically significant effect. This increases the cost, complexity, and length of the study.

Secondly, the results of a placebo-controlled trial are usually unequivocal – either the treatment was more effective than no treatment, or it was not. The answer is not as clear-cut when the comparison group is receiving an active agent.

The two major shortcomings of placebo controls when an effective treatment exists are (a) ethical (which I briefly mention at the end), and (b) ambiguity about the usefulness of the results. That is, while they may give an indication of the absolute benefit of the new treatment, it is impossible to say if that benefit is better than, equivalent to, or worse than the gains that the patients would show if they were on an already approved drug.

Alternatives to Placebo Control

The alternative to using a placebo in the control arm of the study is to use an active agent as the comparison condition. However, this presents the researcher with the issue of what the study question is. There are three alternatives: superiority, non-inferiority, or equivalence.[5]

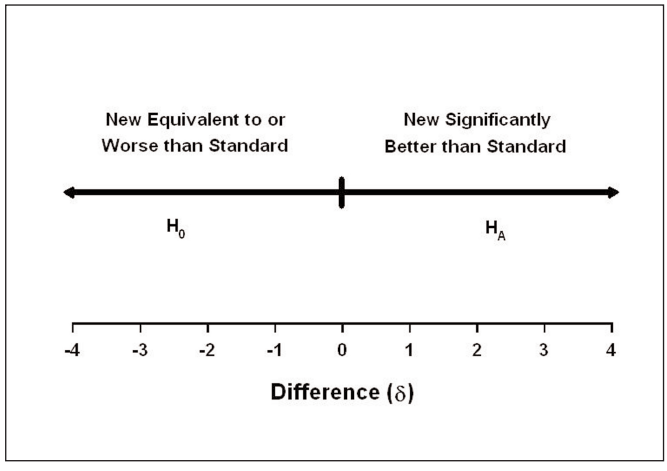

Figure 1: *The null and alternative hypotheses for a superiority trial.*

Superiority Trials

In a superiority trial, the aim – as the name implies – is to show that the new drug is better than the standard one. "Better" can mean either a greater therapeutic effect or, more commonly, fewer or better tolerated adverse events. To keep the discussion simpler, though, I will refer only to effectiveness or efficacy, because the same arguments apply for side effects; and I assume that higher scores are better than lower ones. In a superiority trial, the null hypothesis is that the difference between the means of the two groups, abbreviated as δ, is zero or negative (i.e., favouring the standard treatment); versus the one-tailed alternative hypothesis that the new drug is better (Figure 1). In other words:

$$H_0: \delta \leq 0$$
$$H_A: \delta > 0$$

An alternative way of writing these hypotheses is in terms of the means of the standard drug group (M_S) and the new drug group (M_N):

$$H_0: M_N \leq M_S$$
$$H_A: M_N > M_S$$

The rationale for a one-sided test of significance is that people (or at least drug manufacturers) are not interested in results that show that the new drug is equal to or inferior than the standard; only that it is better. However, many statisticians are uncomfortable with one-tailed tests,[6] because it means that, strictly speaking, "significant" results in the opposite direction must be dismissed as chance findings. Any attempt to "explain away" the contrary finding is an admission that the effect is real, and thus the rationale for the one-tailed hypothesis has been violated (i.e., it is due to chance). There have been many unfortunate examples where it seemed logical to posit that the results could go in only one direction, only to find that the new intervention was in fact harmful. In both the CAST[7] and the Finnish Trial,[8] for example, those in the intervention group – to reduce ventricular arrhythmias and lower cardiac risk factors, respectively – died at a rate between 1½ and 3½ times that of the control group.

The advantage of a one-tailed test of significance is that, for a given effect size, a smaller sample size is needed in order to achieve statistical significance. However, this gain in efficiency is usually more than offset by the more stringent requirement of needing to demonstrate superiority to an already-proven intervention. For these reasons, superiority trials are almost never seen in drug studies (although they are relatively common comparing drugs to behavioural therapies for anxiety and affective disorders)

Sample Size Calculation – Superiority

On placebo, patients decline 5.6 points on the Cognition scale of the ADAS (ADAS-Cog) over 12 months (SD = 7.3). If we want a drug to slow this decline to 2 points, and using a one-tailed test with $\alpha = 0.05$ and $\beta = 0.20$, we would need 51 subjects per group (versus 65

https://doi.org/10.1017/S0317167100005540 Published online by Cambridge University Press

Equivalence Trials

In an equivalence trial, we specify some interval, *I*, and say that the drugs are equivalent if the difference between them, δ, falls within that interval. The rationale for using an interval rather than testing for exact equivalence is predicated on two facts: first, given a sufficient sample size, any difference, no matter how small, can be shown to be statistically significant; and second, some differences, even though statistically significant, may be clinically trivial.[9] For example, if we had two groups with 390 subjects in each, a difference between them of 2 points on the Brief Psychiatric Rating Scale (BPRS) would be statistically significant at $p < .05$, even though a difference this small (i.e., about 1/5 of a standard deviation) would be considered to be meaningless.

The null hypothesis for the two-tailed test is that δ falls outside the interval *I* (that is, the new drug is much better than or much worse than the standard); and the alternative hypothesis is that the difference lies within the interval:

$$H_0: |\delta| \geq I$$
$$H_A: |\delta| < I$$

The alternative notation explicitly states the two null hypotheses (shown in Figure 2):

$$H_{01}: M_N \geq (M_S + I)$$
$$H_{02}: M_N \leq (M_S - I)$$
$$H_A: (M_S - I) < M_N < (M_S + I)$$

That is, if both null hypotheses are rejected (although only one need be tested; see 9), then by default the alternative – that the mean of the new drug falls within the interval – is supported.

Because equivalence trials are based on two-tailed tests, they require larger sample sizes than studies based on one-tailed tests – and may, under certain circumstances, call for much larger sample sizes than even traditional trials,[9] which increases the cost of a study. For these reasons, combined with the drug companies' satisfaction with simply showing that the new drug is no worse than the current standard, equivalence studies are very rarely conducted.

Sample Size Calculation – Equivalence
Setting *I* at 2 points on the ADAS-Cog, with $\alpha = 0.05$ (two one-sided tests), $\beta = 0.20$, and a pooled SD of 7.3, 229 subjects per group would be required.

Non-Inferiority Trials

In non-inferiority trials, the goal is to show that the new drug is not any worse than the existing standard therapy. The question is raised why such studies need be done, since there is a proven intervention. The goals can be either laudatory – the new drug may have a similar therapeutic effect but a better side effect profile – or crass – the drug company simply wants a share of a lucrative market. For whatever reason the study is done, the null hypothesis in a non-inferiority trial is that the new drug is worse than the standard one by at least some amount, *I*, against the alternative hypothesis that the superiority of the standard drug does not exceed this interval *I* (Figure 3); this is also a one-tailed test:

$$H_0: \delta \leq I$$
$$H_A: \delta > -I$$
or:
$$H_0: M_N < (M_S - I)$$
$$H_A: M_N \geq (M_S - I)$$

with, again, all of the caveats against one-tailed tests. Because the role of Type I and Type II errors are reversed in non-inferiority trials,[9,10] we usually set $\alpha = .20$ and $\beta = .05$.

Sample Size – Non-Inferiority
With an equivalence interval *I* of 2 points on the ADAS-Cog, $\alpha = 0.20$, $\beta = 0.05$, and a pooled SD of 7.3, 165 subjects per group would be needed.

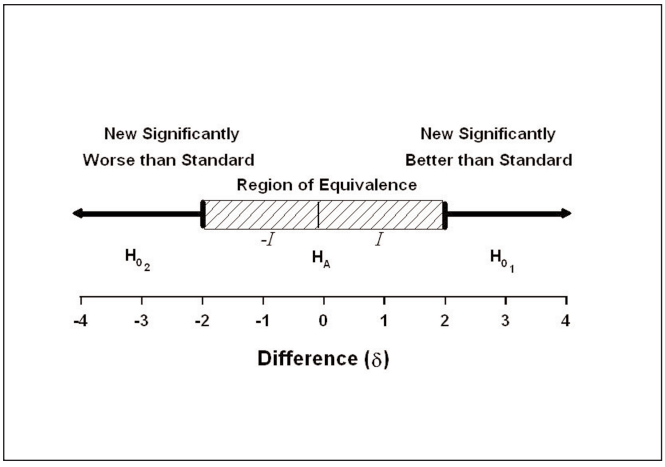

Figure 2: *The null and alternative hypotheses for an equivalence trial.*

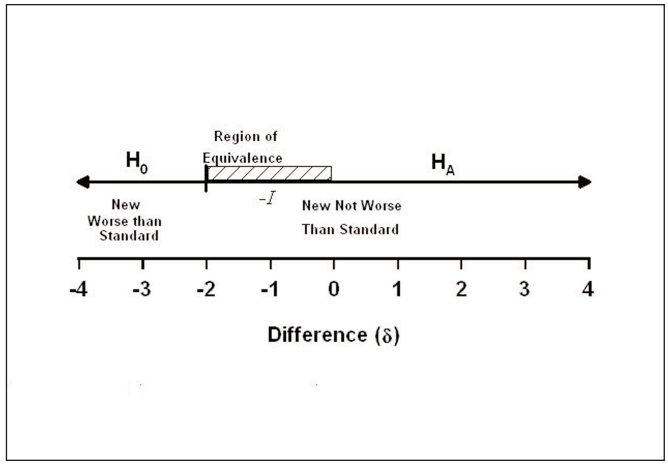

Figure 3: *The null and alternative hypotheses for a non-inferiority trial.*

https://doi.org/10.1017/S0317167100005540 Published online by Cambridge University Press

Problems with Equivalence and Non-Inferiority Trials

Despite their popularity, there are a number of problems with both equivalence and non-inferiority trials. First, the size of the non-inferiority region, *I*, is quite arbitrary; what may be a clinically unimportant difference to one person could be quite important to another. Compounding this difficulty, the sample size is highly dependent on*I*[9]; as it shrinks (that is, the new drug must be less and less inferior to the standard), the sample size increases quite rapidly. Thus, there is great incentive to make the interval as large as possible.

This in turn leads to the second problem. The new drug can be less effective than the standard, but still pass the test of non-inferiority. It is easy to imagine a series of three or four trials, in which the new drug of one becomes the standard of the next. If, in each case, the new drug is worse than the standard, but not significantly so, the effectiveness of the interventions will decline from one study to the next.

The third problem is related to the previous two, and can be much more serious. In a placebo-controlled trial, there are two alternative outcomes: the new drug is better than the placebo, or it is not. That is, the results are unequivocal (ignoring Types I and II errors). This is not the case when a new drug is compared against an active comparison. Again there are two alternatives: the new drug is significantly worse (because, as I've said, it's highly unusual to test for superiority), or it is equivalent. If it is worse, there is no problem interpreting the findings (assuming they ever see the light of day); the new drug is worse than the old one. But, if the results show equivalence or non-inferiority, there are two possible reasons – both are equally effective, or both were equally ineffective in this particular study, and there is no way to determine which is the case.

We may like to assume that if the comparison drug has been shown to be effective, then there is no doubt – both drugs in the trial must have had some positive effect. However, this assumes that all trials were well-conducted: that they have enrolled only participants who are likely to respond to the drug; that there are a sufficient number of people so that the study isn't under-powered; that it was carried out competently (e.g., few people missing appointments, being erroneously given the wrong drug; not dropping out of the trial; and so forth); that the outcome measures were both appropriate for the outcome of interest and administered in a reliable way; and on and on. Unfortunately, this is not always the case. For example, despite the fact that the efficacy of MAOIs for depression has been shown repeatedly, a large trial sponsored by the British Medical Research Council, and where one of the principal investigators (A. Bradford Hill) was the person regarded as the father of the RCT, failed to find any effect of phenelzine, due to the enrollment of the wrong types of patients, inadequate levels of the drug, too short a duration of treatment, and an inappropriate outcome measure.[11] Similarly, Peet et al[12] were unable to demonstrate any effect of either propranolol or chlorpromazine with schizophrenic patients, most likely because their sample size was woefully inadequate.[13] As mentioned above, the role of Type I and Type II errors are reversed in equivalence and non-inferiority trials. Consequently, it is possible to show equivalence merely by running a poorly designed, badly executed, and low-powered study.[10]

The lesson is that even trials that are led by an experienced researcher, which use a proven drug, and are published in prestigious journals may be faulted on one or more grounds. Consequently, we cannot assume that a study that demonstrates no difference between a new drug and a standard has shown the equal efficacy of both. It is also possible that the drugs were equally ineffective in this particular trial, and absent a placebo group, we often cannot determine which situation applies.

Other Considerations

In recent years, there has been a move to replace placebo-controlled trials with add-on designs; that is, comparing treatment as usual (TAU) with one drug versus TAU plus a second drug. This leads to problems in sample size, but as we shall shortly see, makes the design of studies more straightforward. Because the TAU group is expected to improve more (or decline less) than a placebo group, it may be much more difficult to demonstrate the superiority of an additional intervention, because the patients may be closer to any physiological limit in terms of how much they can improve (or slow in their decline). In order to demonstrate statistical significance with smaller differences, larger sample sizes are required. For example, to show a statistically significant difference with an effect size (ES) of 0.5, 63 patients per group are required. If the ES is only 0.25, though, the required sample size is nearly four times as large – 252 in each group. Moreover, because the potential for drug interactions and an increased incidence of adverse events may make the combination worse than TAU, two-tailed tests should be mandatory, obviating the advantage of superiority trials.

Equivalence and non-inferiority trials similarly become meaningless with add-on trials. There is no sense in adding another drug, with all of the resultant costs and opportunities for adverse events and interactions, if the only consequence were to show that the combination is the same as or not worse than taking a single medication.

Thus, with add-on trials, the only meaningful approach is a traditional two-group parallel study.

Ethics and Placebos

I will not get into an extended discussion of the ethical issues of placebos when existing treatments exist; that is covered elsewhere. However, in addition to the possible ambiguity of results discussed above when there is no placebo comparison, there is another consideration arguing in favour of a placebo control. Paradoxical as it may seem, placebo controlled trials may actually expose fewer people to adverse reactions than active control designs; and may result in a similar proportion of people who are untreated.[14] The paradox arises because of sample size. As mentioned above, if the expected ES between a new treatment and a placebo is 0.50, then 63 people are required per group. If the rate of adverse drug reactions (ADRs) is 20%, that means that there will be none in the placebo group and 13 in the treated group. When the comparison condition is another effective drug, then two things happen. First, the expected ES is smaller, resulting in a larger sample size; and second, both groups are exposed to ADRs. If we use the example from above, where the ES is 0.25 and 252 patients are required in each group, then 91 people will experience ADRs; that is, a rate 7 times as high as in a placebo-controlled trial.

https://doi.org/10.1017/S0317167100005540 Published online by Cambridge University Press

Conclusions

There is little question that, when effective treatments exist, there are major ethical issues when new drugs are tested against placebos. However, there are many methodological problems when the comparison group consists of an active treatment: (a) poorly executed trials with low power can be mistaken for "proving" equivalence; (b) when the two arms yield comparable results, there is no guarantee that either one was effective in that particular trial; (c) there may be a tendency (conscious or unconscious) to use wide equivalence intervals to decrease sample size; (d) successive non-inferiority trials may lead to a gradual reduction of effectiveness; and (e) the requirement for large sample sizes with superiority trials and equivalence trials with narrow intervals. These difficulties raise ethical concerns themselves; not only from possibly erroneous findings, but also, if larger sample sizes are needed, from having more patients on a possibly less effective treatment and a delay in getting the drug to market.

It would be fatuous to say that all of these problems would be eliminated if only excellent trials were conducted. Such pleas have been made ever since RCTs were first run, to little avail; poorly designed, underpowered studies appear to be as prevalent now as when Cohen[15] first said that most studies did not have sufficient power to test their main hypotheses.[16] My recommendation would be that, when an existing therapy exists, and if certain conditions apply:

_ Studies should consist of three arms: the new drug, the existing drug, and a placebo group.

_ The study should be adequately powered to detect a clinically important difference in superiority trials, or to rule out a Type II error in equivalence and non-inferiority trials between the two drug arms.

_ The placebo arm need only be large enough to determine that the study as a whole was successful (i.e., to detect a difference between it and the pooled effect of the two treatments).

The conditions that should apply would include:

_ The placebo arm should be as brief as possible.

_ The patient's condition is not expected to deteriorate rapidly.

_ Patients are withdrawn if their deterioration "is greater than that expected for normal clinical fluctuation in a patient with that diagnosis who is on standard therapy".[17]

_ Patients are automatically withdrawn if they begin to exhibit behaviours that may be dangerous to themselves or others, "even if there is not sufficient deterioration in the overall monitoring to trigger disenrollment".[17]

_ There is full and informed consent from the patient and/or the substitute decision maker.

A third arm would increase sample size somewhat, and does not address the issues of the size of the equivalence interval, or the potential gradual reduction in effectiveness, but would solve the major problem of the ambiguity of no difference. While placebo-controlled trials may be unethical, it is even more unethical to do active control studies when they can be scientifically meaningless or misleading.

References

1. Feinstein AR. Clinical biostatistics: XLVIII. Efficacy of different research structures in preventing bias in the analysis of causation. Clin Pharmacol Ther. 1979; 26: 129-41.
2. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. New Engl J med.. 1994; 331: 394-8.
3. Streiner DL, Norman GR. PDQ Epidemiology (2nd ed.). Toronto: B. C. Decker; 1996.
4. Bok S. The ethics of giving placebos. Sci Am. 1974; 231(5): 17-23.
5. Fleischhacker WW, Czobor P, Hummer M, Kemmler G, Kohnen R, Volavka J, et al. Placebo or active control trials of antipsychotic drugs? Arch Gen Psychiatry. 2003; 60: 458-64.
6. Norman GR, Streiner DL. Biostatistics: The bare essentials. 2nd ed. Toronto: B. C. Decker, 2000.
7. Akiyama T, Pawitan Y, Greenberg H, et al. Increased risk of death and cardiac arrest from encainide and flecainide in patients after non-Q-wave acute myocardial infarction in the Cardiac Arrhythmia Suppression Trial. CAST Investigators. Am J Cardiol. 1991; 68: 1551-5.
8. Strandberg TE, Salomaa VV, Naukkarinen VA, Vanhanen HT, Sarna SJ, Miettinen TA, et al. Long-term mortality after 5-year multifactorial primary prevention of cardiovascular diseases in middle-aged men. JAMA. 1991; 266: 1225-9.
9. Streiner, DL. Unicorns do exist: a tutorial on "proving" the null hypothesis. Can J Psychiatry. 2003; 48: 756-61.
10. Steiger JH. Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychol Methods. 2004; 9: 164-82.
11. Clinical Psychiatry Committee. Clinical trial of the treatment of depressive illness: report to the Medical Research Council. BMJ. 1965; 1: 881-6.
12. Peet M, Bethell MS, Coates A, Khamnee AK, Hall P, Cooper SJ, et al. Propranolol in schizophrenia: I. Comparisons of propranolol, chlorpromazine and placebo. Brit J Psychiatry. 1981; 139: 105-11.
13. Streiner DL. Propranolol in schizophrenia [letter to the editor]. Brit J Psychiatry. 1982; 141: 212-13.
14. Streiner DL. The lesser of two evils: the ethics of placebo-controlled trials. Can J Psychiatry (in press).
15. Cohen J. The statistical power of abnormal-social psychological research: a review. J Abnorm Soc Psychol. 1962; 65: 145-53.
16. Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychol Methods. 2004; 9: 147-63.
17. Orr JD. Guidelines for the use of placebo controls in clinical trials of psychopharmacologic agents. Psychiatr Serv. 1996; 47: 1262-4.

https://doi.org/10.1017/S0317167100005540 Published online by Cambridge University Press