Special Communication

Run-in Periods in Randomized Trials

Implications for the Application of Results in Clinical Practice

Ariel Pablos-Méndez, MD, MPH; R. Graham Barr, MD; Steven Shea, MD

Prerandomization run-in periods are being used to select or exclude patients in an increasing number of clinical trials, but the implications of run-in periods for interpreting the results of clinical trials and applying these results in clinical practice have not been systematically examined. We analyzed illustrative examples of reports of clinical trials in which run-in periods were used to exclude noncompliant subjects, placebo responders, or subjects who could not tolerate or did not respond to active drug. The Physicians' Health Study exemplifies the use of a prerandomization run-in period to exclude subjects who are nonadherent, while recent trials of tacrine for Alzheimer disease and carvedilol for congestive heart failure typify the use of run-in periods to exclude patients who do not tolerate or do not respond to the study drug. The reported results of these studies are valid. However, because the reported results apply to subgroups of patients who cannot be defined readily based on demographic or clinical characteristics, the applicability of the results in clinical practice is diluted. Compared with results that would have been observed without the run-in period, the reported results overestimate the benefits and underestimate the risks of treatment, underestimate the number needed to treat, and yield a smaller *P* value. The Cardiac Arrhythmia Suppression Trial exemplifies the use of an active-drug run-in period that enhances clinical applicability by selecting a group of study subjects who closely resembled patients undergoing active clinical management for this problem. Run-in periods can dilute or enhance the clinical applicability of the results of a clinical trial, depending on the patient group to whom the results will be applied. Reports of clinical trials using run-in periods should indicate how this aspect of their design affects the application of the results to clinical practice.

JAMA. 1998;279:222-225

RANDOMIZED CLINICAL trials are the soundest way to test the safety and effectiveness of a clinical intervention,[1,2] and their results are often pivotal in shaping clinical practice and drug policy.[3,4] While not all clinical trials seek to change clinical practice, either by providing clinically relevant evidence to physicians or by documenting effectiveness and safety in the context of the Food and Drug Administration (FDA) approval process,[4] many clinical trials do have this goal. Such trials are costly,[5,6] however, and clinical investigators naturally seek methods to optimize treatment effects. Nonadherence reduces the observed treatment effects of a drug and the statistical power of clinical trials with fixed sample size.[5-8] Simply increasing the sample size to overcome the impact of nonadherence raises the cost of a trial, while restricting analysis of the trial data to adherent subjects effectively "derandomizes" the study groups.[9,10] Investigators may instead exclude patients with traits or circumstances that hinder adherence[7,11] or use protocols that minimize nonadherence.[11] Alternatively, investigators may empirically probe patients' adherence during a run-in period and randomize only adherent subjects into the trial proper.[12,13]

More recently, run-in periods have also been used to screen potential subjects for placebo response[14] or for therapeutic response and adverse effects prior to randomization.[15-18] Despite the increasing use of these modifications to traditional approaches to subject selection, the implications of run-in periods for interpreting the results of clinical trials and applying these results in clinical practice have not been systematically examined. Using selected examples from the recent literature, this article examines 4 uses of run-in periods in clinical trials and analyzes the implications of each.

From the Division of General Medicine, College of Physicians and Surgeons, Columbia University (Drs Pablos-Méndez, Barr, and Shea), and the Division of Epidemiology, Columbia University School of Public Health (Drs Pablos-Méndez and Shea), New York, NY.

Reprints: Ariel Pablos-Méndez, MD, MPH, Division of General Medicine, PH 9 East Room 105, 622 W 168th St, New York, NY 10032-3702.

RUN-IN PERIODS THAT SCREEN POTENTIAL SUBJECTS FOR ADHERENCE

The Physicians' Health Study was the first large clinical trial to use a run-in period to screen for adherence.[5,12,19] This placebo-controlled study was designed to test the effects of aspirin and beta carotene in the primary prevention of ischemic heart disease and cancer, respectively, among male physicians in the United States. Pilot studies showed a nonadherence rate of 20% in the first 6 months, but of only 1% to 2% per year thereafter. After screening for nonadherence to beta carotene placebo and open-label aspirin during an 18-week run-in period, 33% of the subjects were precluded from randomization.[12] Reported adherence among the randomized subjects was 90% over the 5 years of the study,[19] which found that aspirin decreased the rate of myocardial infarction (relative risk [RR], 0.56; 95% confidence interval [CI], 0.45-0.70).[12] By contrast, the British Physicians' Study, which did not use a run-in period to exclude nonadherent subjects, achieved only 70% adherence and found no significant difference between groups receiving aspirin and placebo (RR, 0.97; 95% CI, 0.73-1.24).[20]

The use of a prerandomization run-in period to exclude nonadherent subjects does not compromise the validity of a properly conducted clinical trial because the randomized groups remain comparable, barring chance imbalances, with respect to both measured and unmeasured covariates. Validity, however, does not guarantee applicability. The efficacy observed in such trials is a "theoretical" ceiling for the potential effectiveness of the drug in clinical practice, where both adherent and nonadherent patients are encountered. Clinicians may argue that the results of such trials can be applied to adherent individuals, just as the results are applied to the demographic subset or clinical stage targeted by the study. However, studies show that physicians in clinical practice are quite inaccurate at predicting or detecting nonadherence.[7,21] Thus, the ability of physicians to identify the subset of patients to whom the results of such studies apply is qualitatively different from deciding if a patient is similar to

a demographically or clinically defined group of study subjects in a clinical trial.

This optimal treatment effect may be quite different from the effect observed in clinical practice. While more sophisticated models exist,[8,22-24] a simplified procedure can be used to illustrate the potential magnitude of this difference. In the Physicians' Health Study, which excluded 33% of otherwise eligible subjects because of nonadherence during the run-in period, the reported age-adjusted RR for aspirin prophylaxis was 0.56. Based on the reported absolute outcome rates, we calculated the number needed to treat (NTT) to prevent 1 myocardial infarction per year as 541.[12] If all subjects who entered the run-in period had been randomized, and if it is assumed that nonadherent subjects remained nonadherent and had no treatment effect, the recalculated RR and NNT are, respectively, 0.71 and 807, differences of 25% and 49% in the estimates of effect.

The basis for these calculations is as follows. The observed RR (RR_{obs}) and NNT to prevent an outcome (NNT_{obs}) in a clinical trial are derived from outcome rates observed in the placebo ($P\{O|Pl\}_{obs}$) and experimental ($P\{O|Exp\}_{obs}$) groups. The outcome rate in the placebo group does not change if nonadherers are allowed in the trial. In the experimental group, the revised outcome rate ($P\{O|Exp\}_{rev}$) depends on the randomized proportion, RP, and represents a weighted average of the rate among adherers and the nonadherers who would have the outcome rate of the placebo group if randomized:

$$P\{O|Exp\}_{rev} = P\{O|Exp\}_{obs}\,(RP) + P\{O|Pl\}_{obs}\,(1 - RP)$$

Once the revised outcome rate for the experimental group is obtained, revised estimates of RR and NNT can be easily calculated:

$$RR_{rev} = P\{O|Exp\}_{rev}/P\{O|Pl\}_{obs} = RR_{obs}\,(RP) + (1 - RP)$$

$$NNT_{rev} = 1/(P\{O|Pl\}_{obs} - P\{O|Exp\}_{rev}) = NNT_{obs}/RP$$

If nonadherers were included in the trial, the test statistics would be smaller, the P values larger, and the CIs wider than those derived from a study that only randomized adherent subjects. The SE of the revised estimate of RR (SE_{rev}) may be computed from the estimate of SE of the observed RR (SE_{obs}), the RP, and the number of subjects who entered the run-in period (N):[25]

$$SE_{rev} = [(RR_{obs}-1)^2 RP(1-RP)/N+(SE_{obs}^2 RP_{obs}^2)]^{1/2}$$

In practical situations, measured adherence ranges continuously between 0% and 100%, and the definition of nonadherence in clinical trials is based on adherence below some threshold. If the assumption is made that the relationship between adherence and the outcome rate is linear, then, based on the preceding equations, the revised estimates of the RR and NNT for the Physicians' Health Study for an adherence rate of 50% would be RR of 0.64 and NTT of 641, and for an adherence rate of 80%, RR of 0.61 and NTT of 579.

In summary, the use of a run-in period to select adherent patients for randomization does not compromise validity, but the applicability of positive results is diluted by the high prevalence and inaccurate prediction of nonadherence in clinical practice. Negative results from such studies, on the other hand, will be compelling.

RUN-IN PERIODS THAT SCREEN FOR PLACEBO RESPONSE

A recent randomized trial compared pindolol plus fluoxetine vs placebo plus fluoxetine in the treatment of major depression.[14] Because the depression rating scales used to measure the study end points were based entirely on self-reported data, there was potential in both groups for placebo responses to artifactually increase treatment response rates. A 7-day, single-blind (study patients) placebo run-in period, in which patients received placebo for both fluoxetine and pindolol, was used to exclude placebo responders, defined as those whose depression scores decreased by 25% or more or fell below 18 on the Hamilton Scale. Of the 132 eligible patients who entered the run-in, 111 were randomized and 21 were not, 19 because of placebo response and 2 because of patient withdrawal. The randomized study found a greater proportion of responders in the fluoxetine plus pindolol group (41/55 [75%] vs 33/56 [59%] [90% CI, 1.1-30.1; P=.04]). If the 19 placebo responders had entered the trial, 9.5 of them in each group, the results would still have trended in favor of the fluoxetine plus pindolol group but would not have reached statistical significance (50.5/64.5 [78%] vs 42.5/65.5 [65%] [90% CI, 0.5-26.1; P = .09]). In this case, the investigators were able to increase the statistical power and efficiency of their trial without compromising the applicability of the result, since the response rate in clinical practice, which will include some placebo responders, is unlikely to be lower than the reported rate.

RUN-IN PERIODS THAT USE THE TRIAL DRUG TO SCREEN FOR CLINICAL RESPONSE

Run-in periods have also been used to select subjects for randomization who have a positive clinical response to the trial drug or who do not deteriorate, die, or have other serious adverse effects. The Tacrine Collaborative Study used this type of run-in period in testing a cholinesterase inhibitor in patients with Alzheimer disease.[16] The initial study sample consisted of 632 patients who underwent a 6-week "enrichment" phase during which subjects were randomly allocated to 3 crossover sequences of placebo, low-dose tacrine, and high-dose tacrine. After a 2-week washout period on placebo, only the 215 patients (34% of those initially recruited) who had improved clinically (4-point decline in the Alzheimer Disease Assessment Scale [ADAS]) while receiving the active drug were randomized to either placebo or their best dose of tacrine for 6 weeks. During the randomized phase, patients receiving tacrine achieved a mean cognitive-subscale ADAS score of 30.3 vs 32.7 in the placebo group (P<.001). No significant differences were detected by the Mini-Mental State Examination or physicians' global assessment. The authors also reported the rates of adverse effects in each phase of the study: 42% of patients given tacrine at some point in the study had at least 1 abnormal alanine aminotransferase value, peaking after a mean of 10 weeks of first use; 113 participants (18%) were withdrawn from the study because of adverse effects. No deaths were reported. The study design features and the preliminary nature of the results were clearly described in the article's abstract and discussion, but the interpretation of the results did not specifically address their application in clinical practice.

Two large studies of carvedilol in the treatment of chronic heart failure further illustrate the use of this type of run-in period. The Australia–New Zealand Collaborative Group study excluded 6% of the 442 eligible patients who could not tolerate carvedilol during a 2- to 3-week run-in period.[17] The reported rate of adverse effects leading to withdrawal from the study among patients randomized to carvedilol was 14.5% (30/207).[17] Carvedilol, however, is known to worsen heart failure—in up to 44% of patients—during the initiation of therapy.[26] If the adverse events in patients receiving carvedilol during the run-in period are added to those occurring in the randomized phase of the trial, the overall incidence was 24.4% (57/234) vs 6.3% in the placebo group (RR, 3.9; 95% CI, 2.2-6.9). In addition, 2 sudden deaths occurred during the run-in period on carvedilol but none in the preceding 4 weeks of observation.[17]

The US Carvedilol Heart Failure Study Group study excluded 103 (8.6%) of 1197 otherwise eligible patients during the run-in period, 36 because of protocol violations and other administrative reasons and 67 because of adverse events, of whom 17 had worsening heart failure and 7 died.[18] Patients randomized to carvedilol after the run-in period had fewer

deaths than those in the placebo group (3.2% vs 7.8%; RR, 0.41; 95% CI, 0.24-0.69), and, as the authors pointed out, this mortality benefit held even if the 7 deaths occurring during the run-in period were ascribed to the group randomized to carvedilol. During the randomized phase of the study, the rates of adverse reactions leading to treatment discontinuation were 7.8% in the placebo arm and 5.7% among patients assigned to carvedilol. In the scenario least favorable to the drug, if the 60 nonfatal adverse events during the active-drug run-in period are attributed to the effect of the drug, the rate of adverse effects for carvedilol would be 5.1 per 1000 patient-weeks vs 2.8 in the placebo group (RR, 1.83; 95% CI, 1.22-2.73). However, because there was no control group during the run-in period, the number of adverse events attributable to the natural progression of disease, rather than to the drug, cannot be determined. While the article did not address this issue directly, the authors invited caution in interpreting a trial with "unusual characteristics" and emphasized that carvedilol must be initiated with extreme care in patients with heart failure.[18]

As illustrated by these studies, trials that use a run-in period to select patients on the basis of clinical response to the study drug may preclude a full assessment of both the therapeutic effectiveness and the safety of the drug when used in clinical practice. As with trials that use run-in periods to exclude nonadherent patients, comparisons between the 2 randomized groups of patients remain valid, but estimates of treatment effects and tests of statistical significance will differ from what would have been observed without the run-in period. No methods for adjusting the results of such studies have, to our knowledge, been published.

There are additional complexities in interpreting results from clinical trials that use run-in periods to exclude subjects because of nonresponse or adverse effects that arise when the probabilities of therapeutic response and adverse effects are not independent, that is, when therapeutic responsiveness and sensitivity to the adverse effects of a drug are correlated. This issue is of concern because the assumption of independence is difficult to test and unlikely to hold true in many specific drug trials. It is apparent that exclusion of subjects who fail to respond positively during a run-in, as in the tacrine study, will lead to overestimation of the treatment benefit compared with what would have been observed in a trial without a run-in period. However, this exclusion may also influence the estimate of treatment risk. If, instead of selectively excluding nonresponders, a trial randomized this group, the adverse event rate may be higher, if this group were more susceptible to adverse effects, or lower, if this group were refractory to both adverse and beneficial effects of the drug. The magnitude and even the direction of this influence may be difficult to predict.

Similar reasoning applies to exclusion of subjects who have adverse effects during an active-drug run-in period prior to randomization. It is apparent how this exclusion will lead to underestimation of the rate of adverse effects if this rate is based solely on the randomized phase, but this exclusion may also influence the estimate of treatment benefit when the probabilities of adverse event and treatment response are correlated. For example, in the tacrine study, it is hard to know whether those among the 66 subjects excluded during the run-in because of alanine aminotransferase elevations who would have been randomized to active drug and would have failed to complete treatment because of the adverse effect would have been more or less likely to have a cognitive response than others in the treatment group.

The use of an active-drug run-in period to select subjects based on clinical response also poses other problems in interpreting study results. First, the effects of the active drug during the run-in period may carry over after patients are switched to placebo.[27] While a washout interval at the end of the run-in period might allow patients to return to their baseline clinical state, there is no guarantee that this will occur or that their response to retreatment would be the same as when new to the drug. For example, the tacrine study documented a rebound withdrawal effect among patients randomized to placebo after the active-drug run-in, whereby their cognitive function declined to below baseline.[16] Second, the hypothesis actually tested by clinical trials using active-drug run-ins is that, among patients tolerating the drug during a run-in period, withdrawal (ie, switching to placebo) is not associated with clinical deterioration. This hypothesis differs subtly from a direct test of the initiation of drug treatment vs placebo. The difference is illustrated by the contrasting conclusions of randomized, placebo-controlled studies in patients with heart failure, where digoxin withdrawal induced striking clinical deterioration (RR of worsening heart failure, 5.9; 95% CI, 2.1-17.2),[28] while initiating digoxin treatment has shown less benefit (RR of hospitalization due to heart failure, 0.72; 95% CI, 0.66-0.79; and no survival improvement).[29]

RUN-IN PERIODS THAT ENHANCE CLINICAL APPLICABILITY

The complexities posed by run-in periods are further illustrated by the Cardiac Arrhythmia Suppression Trial (CAST),[15] in which a run-in period was purposively used to enhance the clinical applicability of the results. During the 1980s many cardiologists used anti-arrhythmic drugs to suppress ventricular ectopy in patients after myocardial infarction, based on observational data showing a strong association between ventricular ectopy in this setting and sudden death. A common management strategy was to obtain a Holter monitor test several weeks after myocardial infarction, to treat if high-density or complex ectopy was present, and to judge the effect of treatment by whether repeat Holter monitoring showed a reasonable degree of suppression of the ectopy. If the ectopy suppressed, the patient would be committed to long-term treatment. CAST was designed to test this clinical practice in a randomized controlled trial.[15] Eligible subjects underwent an open-label run-in phase on one of the study drugs to determine a drug and dosage that resulted in reasonable suppression of ectopy as judged by repeat Holter monitoring. Only those patients who suppressed were randomized, either to their best drug and dose or to placebo. The results of CAST were therefore directly applicable to a widespread clinical management strategy. CAST was terminated early because of higher mortality in the treatment group.[15] Had CAST not incorporated this run-in period to document suppressability of the arrhythmia, the results might not have been so immediately accepted and translated into clinical practice, particularly if the study had only shown no difference between the groups, since practicing clinicians might have wondered if those randomized to treatment merely failed to suppress and thereby failed to benefit. This issue could not readily have been addressed by subgroup analysis, since suppressability may have a favorable prognosis, even in the absence of a treatment benefit.

COMMENT

The validity of a randomized trial is necessary but not sufficient for the application of the results to clinical practice. The methodologic standards of modern trials, including randomized treatment assignment, blinding, and analysis based on intention to treat, help protect their validity.[30,31] However, because these methodologic standards do not address the issue of applicability, fully valid study results may nonetheless be compromised with respect to their application in clinical practice.[32] A clinician faced with the results of an internally valid clinical trial can consider the characteristics of the patient population sampled, such as exclusion of patients with certain comorbidities, and the conditions under which the trial was conducted,

such as provision of free care and strict adherence to protocols, and compare these with what is realistic for patients in his or her own clinical setting.[33] This is an imperfect process, and it probably leads to an overestimation of the benefits of many treatments by both clinicians and patients, but all these circumstances are described by information that is available at the time the clinical decision to use a drug is made. The use of a run-in period to select study subjects removes the study from clinical practice by an additional, qualitatively different step that requires prospective information about a patient's response over time. In most circumstances, this information is not available at the time of the clinical decision to initiate treatment. Thus, practicing clinicians or patients may wish to have evidence from studies with run-in periods recalculated as if the run-in had not been used, in order to have a more directly applicable estimate of the risks and benefits of the drug. As the examples illustrate, however, such recalculations are usually not possible and, when possible, require highly artificial assumptions even when the numbers excluded during the run-in period are known.

As shown by the CAST study, the use of a prerandomization run-in period to focus a study on a highly selected group of patients can also enhance clinical applicability. This example illustrates the principle that it is the relationship between the characteristics of the randomized study sample and those of patients encountered in clinical practice, not the presence of the run-in period per se, that determines whether the run-in period enhances applicability or dilutes it.

It has been suggested that design issues in clinical trials may be conceptualized along 3 dimensions, validity, generalizability, and efficiency.[31] Validity is inviolable, but in this schema there may be trade-offs between efficiency and generalizability, and earlier analyses of run-in periods have used statistical models to focus on the gains in efficiency that might be achieved by using run-in periods to exclude nonadherent patients.[13,19,22-24] Similar models could be constructed for run-in periods to exclude placebo responders. However, in situations where the run-in period dilutes applicability, use of a run-in period does not increase efficiency, if efficiency is understood as estimating the right quantity with the smallest sample size. Estimates of treatment effects that cannot readily be applied in clinical practice do not represent increased efficiency, even if sample size and cost are reduced.

Editors of medical journals should consider requiring abstracts as well as methods to state the type of run-in period used, the proportions excluded during the run-in, and the causes for exclusion. The discussion sections of papers reporting such trials should explicitly examine the implications of the run-in strategy for clinicians seeking to apply the trial results to everyday patients, in order to avoid an overly optimistic view of the effectiveness and safety of a therapeutic intervention. Meta-analyses[34] and cost-effectiveness evaluations[35] using results from such trials should make explicit the methods used to take into account the effects of the run-in period on parameter estimates reported from the randomized phase of those trials. Finally, *run-in period* should be made an indexable term and included in the checklist for reporting clinical trials in the biomedical literature.[36,37]

The authors wish to thank Ana Diez-Roux, MD, PhD, Daniel Rabinowitz, PhD, Richard Defendini, MD, and Katherine Nickerson, MD, for helpful comments.

References

1. Pocock SJ. Statistical aspects of clinical trial design. *Statistician.* 1982;31:1-18.
2. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, part I. *Br J Cancer.* 1976;34:585-612.
3. From trial outcomes to clinical practice. *Drug Ther Bull.* 1996;34:38-40.
4. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation (CBER). *Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products (Draft).* Available at: http://www.fda.gov/CBER/cberftp.html. Accessed March 13, 1997.
5. Hennekens CH. Issues in the design and conduct of clinical trials. *J Natl Cancer Inst.* 1984;73:1473-1476.
6. Meinert CL, Tonascia S. *Clinical Trials: Design, Conduct, and Analysis.* New York, NY: Oxford University Press; 1986.
7. Haynes RB, Dantes R. Patient compliance and the conduct and interpretation of therapeutic trials. *Control Clin Trials.* 1987;8:12-19.
8. Schork MA, Remington RD. The determination of sample size in treatment-control comparisons for chronic disease studies in which drop-out or non-adherence is a problem. *J Chronic Dis.* 1967;20:233-239.
9. Peduzzi P, Detre K, Wittes J, Holford T. Intent-to-treat analysis and the problem of crossovers. *J Thorac Cardiovasc Surg.* 1991;101:481-487.
10. Newell DJ. Intention-to-treat analysis. *Int J Epidemiol.* 1992;21:837-841.
11. Lipid Research Clinics Program. The Lipid Research Clinics Coronary Primary Prevention Trial results, I: reduction in incidence of coronary heart disease. *JAMA.* 1984;251:351-364.
12. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med.* 1989;321:129-135.
13. Lang JM. The use of a run-in to enhance compliance. *Stat Med.* 1990;9:87-95.
14. Perez V, Gilabert I, Faries D, Alvarez E, Artigas F. Randomised, double-blind, placebo-controlled trial of pindolol in combination with fluoxetine antidepressant treatment. *Lancet.* 1997;349:1594-1597.
15. The CAST (Cardiac Arrhythmia Suppression Trial) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med.* 1989;321:406-412.
16. Davis KL, Thal LJ, Gamzu ER, et al, and the Tacrine Collaborative Study Group. A double-blind, placebo-controlled multicenter study of tacrine for Alzheimer's disease. *N Engl J Med.* 1992;327:1253-1259.
17. Australia-New Zealand Heart Failure Research Collaborative Group. Effects of carvedilol, a vasodilator-β-blocker, in patients with congestive heart failure due to ischemic heart disease. *Circulation.* 1995;92:212-218.
18. Packer M, Bristow MR, Cohn JN, et al, for the US Carvedilol Heart Failure Study Group. The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. *N Engl J Med.* 1996;334:1349-1355.
19. Lang JM, Buting JE, Rosner B, Cook N, Hennekens CH. Estimating the effect of the run-in on the power of the Physicians' Health Study. *Stat Med.* 1991;10:1585-1593.
20. Peto R, Gray R, Collins R, et al. Randomised trial of prophylactic daily aspirin in British male doctors. *BMJ.* 1988;296:313-316.
21. Stephenson BJ, Rowe BH, Haynes RB, Macharia WM, Leon G. Is this patient taking the treatment as prescribed? *JAMA.* 1993;269:2779-2781.
22. Schechtman KB, Gordon ME. A comprehensive algorithm for determining whether a run-in strategy will be a cost-effective design modification in a randomized clinical trial. *Stat Med.* 1993;12:111-128.
23. Brittain E, Wittes J. The run-in period in clinical trials: the effect of misclassification on efficiency. *Control Clin Trials.* 1990;11:327-338.
24. Blackwelder WC, Hastings BK, Lee ML, Deloria MA. Value of a run-in period in a drug trial during pregnancy. *Control Clin Trials.* 1990;11:187-198.
25. Casella G, Berger RL. *Statistical Inference.* Pacific Grove, Calif: Wadsworth & Brooks/Cole Advanced Books & Software; 1990:section 7.4.2.
26. Sackner-Bernstein JD, Krum H, Goldsmith RL, et al. Should worsening heart failure early after initiation of beta-blocker therapy for chronic heart failure preclude long-term treatment? *Circulation.* 1995;92(suppl I):395. Abstract 1881.
27. Louis TA, Lavoru PW, Bailar JC, Polansky M. Crossover and self-controlled designs in clinical research. *N Engl J Med.* 1984;310:24-31.
28. Packer M, Gheorghiade M, Young JB, et al, for the RADIANCE Study. Withdrawal of digoxin from patients with chronic heart failure treated with angiotensin-converting enzyme inhibitors. *N Engl J Med.* 1993;329:1-7.
29. The Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med.* 1997;336:525-533.
30. Senn SJ. Clinical trials and epidemiology. *J Clin Epidemiol.* 1990;43:628-632.
31. Sackett DL. The competing objectives of randomized trials. *N Engl J Med.* 1980;303:1059-1060.
32. Liberati A. The relationship between clinical trials and clinical practice: the risks of underestimating its complexity. *Stat Med.* 1994;13:1485-1491.
33. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about a therapy or prevention, B: what were the results and will they help me in caring for my patients? *JAMA.* 1994;271:59-63.
34. Stienen U. The once-daily dose regimen of carvedilol: a meta-analysis approach. *J Cardiovasc Pharmacol.* 1992;19(suppl 1):S128-S133.
35. Rutten-van Molken MP, van Doorslaer EK, van Vliet RC. Statistical analysis of cost outcomes in a randomized controlled clinical trial. *Health Econ.* 1994;3:333-345.
36. The Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Checklist of information for inclusion in reports of clinical trials. *Ann Intern Med.* 1996;124:741-743.
37. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA.* 1996;276:637-639.