Multi-Arm Clinical Trials of New Agents: Some Design Considerations

Boris Freidlin,[1] Edward L. Korn,[1] Robert Gray,[2] and Alison Martin[1]

Abstract A major challenge in the development of anticancer therapies is the considerable time and resources needed for conducting randomized clinical trials (RCT). There is a need for more efficient RCT designs that accelerate development, minimize costs, and make trials more appealing to patients. We review the statistical and logistical characteristics of multi-arm designs that compare several experimental treatments to a common control arm. In particular, we present a rationale for not requiring multiplicity adjustment in multi-arm trials that are designed for logistical efficiency. Relative to conducting separate RCTs for each experimental agent, this multi-arm design is shown to require a lower total sample size than multiple two-arm trials.

The definitive stage of drug development requires convincing demonstration of clinical benefit, the gold standard being the phase III randomized clinical trial (RCT) in which patients are randomly assigned to the experimental or control treatment. As RCTs require considerable patient, infrastructure, and financial resources (1), the number of experimental agents that can be tested in a given disease setting at a given time may be extremely limited. It is therefore imperative to optimize the phase III evaluation process. One attractive possibility for improving efficiency is to perform multi-arm trials with multiple experimental treatment arms and a single control arm. By sharing a control arm, the total required sample size can be dramatically reduced, allowing for agents to be tested in a more rapid fashion. There are, however, a number of statistical and logistical issues involved in the design of multi-arm trials that can lessen their efficiency compared with performing separate two-armed trials of each agent. We discuss these issues using, as an example, the design of the intergroup trial E2805 led by the Eastern Cooperative Oncology Group.

Sharing a Control Arm Reduces the Total Sample Size

If experimental agents X and Y are tested separately in their own RCTs, then there will be four treatment arms altogether (X, Y, and two control arms), whereas a single multi-arm trial will have three treatment arms (X, Y, and a single control arm), resulting in a reduction in total sample size of 25% (assuming no multiplicity adjustment; see below). With three experimental agents, the corresponding numbers of arms are six (for separate trials) and four (for a multi-arm trial), yielding a 33% reduction in total sample size. With four experimental agents, there is 37% reduction in total sample size; in general, with K experimental agents, there will be a $(K - 1) / (2K)$ reduction in total sample size.

An additional benefit to using a multi-arm design is that because there is a higher probability that a patient will be randomized to one of the experimental treatments than if separate two-armed trials are done (for which the probability would be 50%), the multi-arm RCT may be more attractive for patient participation (2–4). To further increase the probability of receiving an experimental treatment, some multi-arm trials are designed to randomize fewer patients to the control arm. For example, the SORCE trial (5) randomized fewer patients to the placebo arm. However, the marginal increase in the probability of receiving an experimental agent is offset by a larger overall sample size required in this approach. In a two-armed trial, the most efficient randomization is 1:1. This is no longer true for a multi-arm trial. In fact, the sample size reduction is maximized if one allows for the possibility of randomizing a larger proportion of patients to the (single) control arm than to each of the experimental arms (6). However, the incremental gain from increasing the proportion of control arm patients is small (usually <5%) and thus does not justify the negative aspect of reducing the probability of experimental arm assignment and the added complexity of unequal randomization.

Multiplicity Adjustment

It is well recognized by the medical research community that when multiple statistical tests are carried out in a single experiment, the probability of a false-positive result is inflated (7, 8). With a multi-arm trial with K (>1) experimental treatment arms, this multiplicity question arises because each experimental arm will be compared with the control treatment arm, requiring K statistical tests. However, whether a multiplicity adjustment is required for this application is not straightforward. First, we note that if one wanted to make a multiplicity adjustment, one possibility would be to use a Bonferroni adjustment and require a nominal significance level

Authors' Affiliations: [1]Biometric Research Branch and Clinical Investigations Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland and [2]Eastern Cooperative Oncology Group, Harvard School of Public Health, Boston, Massachusetts

Received 2/6/08; revised 3/4/08; accepted 3/26/08.

Requests for reprints: Boris Freidlin, Biometric Research Branch, EPN-8122, National Cancer Institute, Bethesda, MD 20892. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov.

doi:10.1158/1078-0432.CCR-08-0325

of α/K for each of the K treatment-control comparisons to limit the overall probability of any false-positive result to be $\leq\alpha$. The effect of this Bonferroni adjustment is to dramatically reduce the efficiency advantage of using the multi-arm trial. For example, for a trial designed with 90% power with one-sided α = 0.025, the sample size reductions for using a multi-arm trial instead of K two-armed trials are 11%, 14%, and 15% for K = 2, 3, and 4, respectively, when a Bonferroni adjustment is used. The reductions (Table 1) are slightly smaller for a trial designed with 80% power, and can disappear for less power (9). These sample size reductions can be compared with 25%, 33%, and 37% if no adjustment were made. Several less conservative multiplicity adjustment methods have been proposed (10–12); however, these methods provide only a marginal improvement over the Bonferroni adjustment.

But is it appropriate to make a multiplicity adjustment for this multi-arm trial application? A popular argument against such adjustments is that if the experimental arms were compared with controls in separate two-armed trials, then no adjustment would be made (13). Although this is an appealing argument, it requires refinement and qualification. In particular, the relationship between the individual treatment questions needs to be examined. Was the study designed to support treatment recommendation(s) by a combined evaluation of individual comparisons or was each comparison intended to result in a separate decision? To put it another way: were the experimental arms put together in a single trial because the corresponding questions are related, or primarily, because of efficiency and logistical reasons? (14) An example of when the questions are related would be a trial that evaluates the addition of an experimental agent to several backbone regimens versus the control arm: an RCT (15) evaluated whether the addition of docetaxel to two platinum regimens is beneficial in advanced non–small cell lung cancer, docetaxel + cisplatin or docetaxel + carboplatin versus the standard regimen (vinorelbine + cisplatin). One should make a multiplicity adjustment here because the two comparisons are related in that they are both part of the primary study question "addition of docetaxel to a platinum agent." Therefore, the treatment recommendation will be based on a joint interpretation of treatment comparisons. For example, if one of the two comparisons showed a marginally significant treatment effect but the other comparison was flat negative, we would need a smaller P value from the positive comparison to be convincing.

Another setting with related questions is when a trial evaluates several different schedules (16) or doses (17) of an agent versus the control. Here, again, each experimental arm is a component of a primary overall question, and the results of the treatment-control comparisons reflect on each other, suggesting that a multiplicity adjustment would be appropriate. In fact, some regulatory agencies explicitly require the adjustment for studies testing multiple doses (18).

For the multi-arm trial application being considered here, several experimental agents share a control arm for the purpose of improving efficiency and the trial is focused on answering the efficacy question for each drug separately; the interpretation of the results of one comparison have no direct bearing on the interpretation of the others. In this situation, we believe no multiplicity adjustment is required (19). An example is E2805 which randomly assigned locally advanced renal cancer patients to nine cycles of sunitinib (arm A), sorafenib (arm B), or placebo (arm C) after radical or partial nephrectomy. This study was designed to address two separate questions in resected renal cancer at high risk of recurrence: (a) is adjuvant sunitinib better than placebo and (b) is adjuvant sorafenib better than placebo? These questions could easily have been tested in different two-armed trials and the answer to each question is unrelated to the other. As we argue above, multiplicity adjustment is not required for this design. However, E2805 was designed using multiplicity adjustment to control the overall probability of a false-positive result to be 0.025 (one-sided type I error). The design calls for enrolling a total of 1,332 patients to provide ~80% power to detect a 25% reduction in the hazard for disease-free survival. Without the adjustment, the same power can be achieved with only 1,110 patients. For comparison, conducting separate trials for each drug would require 1,480 patients.

The abovementioned issue of relatedness between the individual questions should not be confused with statistical correlation between the corresponding statistical tests associated with each question (20–22). In a multi-arm study, individual comparisons are positively correlated due to the use of the same control arm (this is not the case for separate trials). Because of this, a multi-arm trial has a lower overall probability of any false-positive result but a higher probability of making more than one false-positive conclusion (relative to separate trials; ref. 9). However, these probability differences are small (9), especially when the number of experimental arms is in a practical range (two to four arms). Therefore, the fundamental issue for the purpose of multiplicity adjustment is the relatedness of clinical questions with the statistical correlation having minimal relevance.

Table 1. Reduction in sample size in a multi-arm trial relative to conducting K independent two-armed trials assuming trial designed with one-sided significance level (type I error) of 0.025, power as stated (for each experimental agent), and the accrual and follow-up periods being the same for all trials

No. of experimental arms (K)	No multiplicity adjustment to significance level	With Bonferroni multiplicity adjustment to significance level		
		50% power	80% power	90% power
2	25%	2%	9%	11%
3	33%	0%	11%	14%
4	37%	(1%)*	12%	15%

*In this situation, the required sample size is 1% larger for a multi-arm trial relative to four independent two-armed trials.

Interim Monitoring

RCTs incorporate formal interim-monitoring guidelines to allow stopping early for strong evidence of benefit of the experimental agent (efficacy) or when it seems that the experimental agent will not be better than the control treatment (futility). The presence of K experimental arms introduces additional complexity. We restrict the discussion here to the situation in which, as described above, the experimental arms are being incorporated into a multi-arm trial for efficiency reasons. As we noted above, in this situation, a multiplicity adjustment is not needed for the nominal significance levels associated with the individual treatment-control comparisons. However, even though this is true, a multiplicity adjustment is required for efficacy interim monitoring. This is because if sufficiently strong evidence of experimental treatment versus control efficacy is observed for at least one of the experimental treatments, this treatment will be deemed an improvement over the control treatment and the control arm will have to be stopped. This, in turn, dictates stopping accrual to other experimental arms because without concurrent controls, the subsequent patient results cannot be interpreted. Because the trial will be stopped if any of the treatment-control comparisons cross an efficacy interim-monitoring boundary, a multiplicity adjustment is required for the efficacy boundaries. A simple Bonferroni approach can be used, by dividing a P value stopping boundary for each treatment-control comparison by K. This leads to a more conservative efficacy boundary (one requiring stronger evidence for stopping) than if no multiplicity adjustment was made. After accrual is completed and all patients are off the study treatment, the effect of early reporting of one treatment-control comparison may only have a negligible effect on the other comparisons, especially if the possibility of crossover is low (e.g., in an adjuvant setting). Therefore, a less conservative efficacy monitoring approach may be appropriate after all study treatments have been completed.

In contrast to efficacy monitoring, no multiplicity adjustment is required for futility monitoring. This is because if one of the experimental arms crosses a futility interim-monitoring boundary, this treatment arm can be discontinued and the results of that treatment-control comparison can be made public without interfering with the continuation of the trial for the other treatment arms.

The efficacy monitoring for E2805 is based on the O'Brien-Fleming boundary (23), adjusted to control the overall false-positive error of concluding the efficacy of either experimental arm. The first interim analysis is planned at the time when 34% of the total number of events are observed. Accrual is expected to be completed at that time (although with some patients still on study treatment), minimizing the potential effect of efficacy stopping on the other treatment-control comparison. In theory, at the later interim analyses, after all patients are off the study treatment, one could have considered not making multiplicity adjustment in efficacy boundary. The protocol also specifies a futility rule without multiplicity adjustment that allows the discontinuation of an experimental arm.

Use of Placebos for Blinding

Many RCTs are blinded, especially when the primary outcome is not overall survival. In a two-armed trial, this is achieved by randomizing patients to receive either the experimental agent or a placebo for the agent. With a multi-arm trial, blinding must ensure that none of the arms can be distinguished. When the multiple experimental arms employ different doses, administration schedules, or administration modalities (e.g., p.o. versus i.v.), blinding can become cumbersome. For example, in E2805, sorafenib and sunitinib are administered orally at different doses and schedules for nine cycles: sunitinib 50 mg (4 × 12.5 mg capsules) p.o. daily ×4 weeks followed by rest ×2 weeks; sorafenib 400 mg (2 × 200 mg tablets) p.o. twice daily ×6 weeks. Therefore, a distinct placebo was required for each drug: patients were randomized to receive (a) sunitinib and placebo for sorafenib (at sorafenib schedule), (b) placebo for sunitinib (at sunitinib schedule) and sorafenib, and (c) placebo for sunitinib (at sunitinib schedule) and placebo for sorafenib (at sorafenib schedule). The situation would be even more complicated if one of the drugs required i.v. administration. Generally, as the number of experimental arms increases, blinding can become less feasible.

One possibility to avoid the use of multiple placebos is to use partial blinding in which the patients randomized to the experimental treatments receive no placebos, and the control patients randomly receive one of the placebos for the experimental treatments. For example, if this approach had been used in E2805, one-third of the patients would have been randomized to sunitinib, one-third to sorafenib, one-sixth to a placebo for sunitinib, and one-sixth to a placebo for sorafenib. Each experimental arm would be compared against pooled placebo arms. This type of partial blinding is better than no blinding, but is not perfect; patients will not know whether they are on the control arm or an experimental arm, but they will know that they are not on some specific experimental arms. In addition, if one or more experimental arms are stopped for futility, then some patients will know that they are on either the control treatment or a treatment that has been deemed no better.

Definitive evaluation of an experimental agent in a RCT often requires data collection and/or logistical considerations tailored to the particular agent. For example, if certain toxicity was observed in early development, a more intensive monitoring focused on that toxicity may be required in the phase III study (especially early in the trial; ref. 24). In a multi-arm blinded trial, this will require identical procedures to be used on all arms, potentially implying additional intensive and costly monitoring that may include invasive tests.

Other Logistical Issues with Multi-Arm Studies

Often, experimental arms have eligibility restrictions due to the specific safety/toxicity profile concerns. Therefore, combining the arms in a multi-arm study has the potential to narrow the pool of eligible patients (24). Restrictive eligibility criteria may adversely affect accrual and generalizability (25).

Pharmaceutical Company Issues

Pharmaceutical companies devote considerable resources to the development of new agents, so any properties of a multi-arm trial design that lessens the chance of a positive phase III

result will diminish companies' willingness to participate. One issue of major concern is that the company's agent will be compared with the other experimental agents in the trial, and its status will be diminished if it has less efficacy than some of the other agents. As the trial is not designed to have sufficient precision to distinguish activity between the experimental treatments, this is an unfair comparison as one agent may falsely seem better than another. Although there is no way to keep readers from interpreting published trial results in various ways, the trial protocol can specify that the analyses for this trial are restricted to comparisons of each agent to the control.

Another possible company concern may be that one of the other agents being tested will cross an interim efficacy boundary and force the closure of the trial before the company's agent has been fully evaluated. However, as noted above, the interim-monitoring boundary for efficacy should be conservative in the multi-arm trial design. Therefore, it is less likely that a multi-arm trial would stop than a separate two-armed trial of another agent would stop early for efficacy—an event that could also put a two-armed trial of the company's agent at risk of early closure.

A challenge to the industry is the potential that their investment in a RCT may be jeopardized by results from a competitor's trial which change the standard of care. A multi-arm trial serves, in effect, to level the playing field by standardizing the assessment of the agents.

It is essential that the sponsor of the RCT exercises the same degree of attention to confidentiality of results and intellectual property rights that accompany separate two-armed trials. In E2805, both companies will receive at the end of the trial the raw data for the control arm and their agent.

Discussion

Phase III RCTs are an expensive and time-consuming component of drug development (1). Improving the efficiency of phase III designs is an important public health issue. In this article, we provide a rationale for not requiring multiplicity adjustment in trials that were designed to achieve logistical efficiency by comparing several experimental treatments to a common control arm. This multi-arm approach is shown to require a lower total sample size compared with conducting separate RCTs for each agent. This design is also more appealing to patients and physicians because it provides an increased probability of receiving an experimental agent rather than the control treatment.

The multi-arm design efficiency gains offer the possibility of being able to test more experimental agents in a quicker fashion. However, rapid accrual of a large number of patients will still be required. If the multi-arm trial captures the most interesting agents available in the field at a time into a single trial, the trial can be seen as compelling and have enhanced accrual (E2805 has been able to accrue significantly ahead of schedule). Individual pharmaceutical companies usually do not have multiple drugs ready for phase III testing in the same disease setting. Therefore, several pharmaceutical companies with competitive products would likely need to be involved in a multi-arm trial, but may lack incentives to take advantage of this approach. These and other logistical complexities limit the pool of organizations that can successfully facilitate and carry out such a design. Large multi-institutional consortia (like the National Cancer Institute Cooperative Groups) are especially positioned to use this multi-arm approach. The groups have (*a*) wide access to patients and a proven track record for rapid accrual, (*b*) infrastructure for the design and conduct of large complex clinical trials, and (*c*) a unique academic/government status needed to address the logistics to accomplish complex blinding, drug distribution from a central source, data management, confidentiality, and preservation of intellectual property across several companies.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Roberts TG, Jr., Lynch TJ, Jr., Chabner BA. The phase III trial in the era of targeted therapy: unraveling the "go or no go" decision. J Clin Oncol 2003;21:3683–95.
2. Benson AB III, Pregler JP, Bean JA, et al. Oncologists' reluctance to accrue patients onto clinical trials: an Illinois Cancer Center study. J Clin Oncol 1991;9:2067–75.
3. Avins AL. Can unequal be more fair? Ethics, subject allocation, and randomised clinical trials. J Med Ethics 1998;24:401–8.
4. Jenkins V, Fallowfield L. Reasons for accepting or declining to participate in randomized clinical trials for cancer therapy. Br J Cancer 2000;82:1783–8.
5. Eisen T. Adjuvant therapy in renal cell carcinoma: where are we? Eur Urol Suppl 2007;6:492–8.
6. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986. p. 96.
7. Godfrey K. Statistics in practice. Comparing the means of several groups. N Engl J Med 1985;313:1450–6.
8. Bauer P. Multiple testing in clinical trials. Stat Med 1991;10:871–89.
9. Proschan MA, Folmann DA. Multiple comparisons in a single experiment versus separate experiments: why do we feel differently? Am Stat 1995;49:144–9.
10. Dunnett CW. A multiple comparisons procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50:1096–1121.
11. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
12. Bristol DR. Designing clinical trials for two-sided multiple comparisons with a control. Control Clin Trials 1989;10:142–52.
13. Hughes MD. Multiplicity in clinical trials. In: Kotz S, Johnson NL, Read CB, editors. Encyclopedia of biostatistics. New York: Wiley; 1998.
14. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials 2000;21:527–39.
15. Fossella F, Pereira JR, von Pawel J, et al. Randomized, multinational, phase III study of docetaxel plus platinum combinations versus vinorelbine plus cisplatin for advanced non-small-cell lung cancer: the TAX 326 study group. J Clin Oncol 2003;21:3016–24.
16. Perez EA, Suman VJ, Davidson NE, et al. Effect of doxorubicin plus cyclophosphamide on left ventricular ejection fraction in patients with breast cancer in the North Central Cancer Treatment Group N9831 Intergroup Adjuvant Trial. J Clin Oncol 2004;22:3700–4.
17. Rosen LS, Gordon D, Tchekmedyian S, et al. Zoledronic acid versus placebo in the treatment of skeletal metastases in patients with lung cancer and other solid tumors: a phase III, double-blind, randomized trial—the Zoledronic Acid Lung Cancer and Other Solid Tumors Study Group. J Clin Oncol 2003;21:3150–7.
18. EMEA CPMP. Points to consider on multiplicity issues in clinical trials. http://www.emea.europa.eu/pdfs/human/ewp/090899en.pdf.
19. Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. J R Stat Soc [Ser A] 1996;159:93–110.
20. Miller R. Simultaneous statistical inference. New York: Springer-Verlag; 1981. p. 26.
21. Dixon DO, Pennello G. Comment on: Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials 2001;22:548–52.
22. Proschan MA, Waclawiw MA. Authors reply: Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials 2001;22:548–52.
23. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.
24. Vermorken JB, Parmar MK, Brady MF, et al. Clinical trials in ovarian carcinoma: study methodology. Ann Oncol 2005;16 Suppl 8:20–9.
25. George SL. Reducing patient eligibility criteria in cancer clinical trials. J Clin Oncol 1996;14:1364–70.