Clinical Trials in the Wake of Vioxx
Requiring Statistically Extreme Evidence of Benefit to Ensure the Safety of New Drugs

Michelle D. Roth-Cline, BS

Clinical trials have become the preeminent tool for investigating new diagnostic, therapeutic, and preventive treatment strategies. During these trials, periodic interim data analyses are conducted by data monitoring committees (DMCs) to ensure participant safety. If accumulated data indicate harm to participants, clear evidence of superior efficacy, or futility of collecting additional data, studies may be terminated before their planned conclusion.

These termination decisions are invariably complex and involve important subjective and ethical elements. Of greatest concern when studies are terminated because of clear benefit is the possibility that the superior intervention has toxicities that remain undiscovered. Early termination entails the abbreviation of clinical experience that would permit assessment of the longer-term safety profile. However, study continuation when treatments are not equally effective also means that some participants will continue to receive a less effective therapy.

The tension between protecting both trial participants and future patients has been increasingly acute since rofecoxib (Vioxx [Merck]) and other cyclooxygenase-2 inhibitors were implicated in a greater risk of myocardial infarction. Thrombotic side effects of cyclooxygenase-2 inhibition were postulated as a result of the propensity for selective antagonists to decrease vascular prostacyclin production and possibly affect the balance between prothrombotic and antithrombotic eicosanoids. Little evidence of toxicity was evident in preapproval trials, however, and rofecoxib was approved in 1999 for relief of the signs and symptoms of osteoarthritis, the management of acute pain, and the treatment of primary dysmenorrhea.

The possibility of cardiac toxicity in human trials was recognized in the Vioxx Gastrointestinal Outcomes Research (VIGOR) trial. Meta-analyses, database studies, and prospective trials further corroborated cardiac toxicity, and Merck voluntarily withdrew rofecoxib from the US market on September 30, 2004, amid widespread criticism of both Merck and the Food and Drug Administration.

In the wake of the withdrawal, drug safety has received increasing public attention. The need to ensure the reliability of safety information has prompted proposals to increase the length of preapproval trials to obtain longer-term safety information. One such strategy recently used by the planning committees for some definitive cardiovascular trials, in consultation with the Food and Drug Administration, is to require statistically extreme evidence of benefit on several end points before considering early termination (personal communication, David DeMets, PhD, 2005). Typical stopping guidelines suggest termination when statistically conservative but more moderate evidence of benefit is attained on key end points, provided that other outcomes are favorable and consistent. Whereas as others have noted the possibility exists that early termination may lead to premature adoption of study findings, at issue here is the likelihood that toxicities would be observed with trial continuation. Although the utility of extending trials may seem self-evident, this analysis will show that obtaining toxicity information in this manner is less certain than it initially appears.

The argument for requiring statistically extreme evidence appears to be based on 4 premises: (1) If statistically extreme evidence of benefit were required for early termination, clinical trials would be longer; (2) if clinical trials were longer, additional toxicity data might be obtained; (3) if additional toxicity data were obtained as a result of requiring statistically extreme evidence of benefit, toxicities that had a bearing on future approvability might be observed that would otherwise remain hidden; and (4) the policy of requiring this additional toxicity data for the purpose of preventing the approval of therapies with unobserved toxicities is coherent.

In what follows, these premises are examined by considering the likely effects of imposing the requirement on the likelihood of approving drugs with hidden toxicities.

Two True Premises
(1) If statistically extreme evidence of benefit were required for early termination, clinical trials would be longer. This first premise is prima facie true. Suppose a trial were conducted that compared 2 therapies intended to decrease...
mortality among patients with heart failure. Sample size is unaffected by the extreme monitoring boundary, because at the end of the study, the critical value for statistical significance remains at conventional levels. Suppose that at the time of interim analysis, patients on treatment A had a lower death rate than those on treatment B and that the probability value for benefit was 0.01. If more evidence of benefit were required ($P=0.0001$), additional evidence of benefit on treatment A would have to accumulate. Because of the time required to observe these additional outcomes, the length of the trial must increase, and termination would be delayed. Therefore, requiring statistically extreme evidence generally entails a longer trial, and the first premise is true.

(2) If clinical trials were longer, additional toxicity data might be obtained.

This second premise is also prima facie true. One of the disadvantages of early termination is that less long-term clinical experience with the new therapy is obtained. The probability of observing more adverse events (AEs) generally increases with patient exposure, whether the toxicity is based on cumulative dose or occurs at a steady state. So, combining the first 2 premises, if statistically extreme evidence of benefit were required for early termination, additional toxicity data might be obtained.

However, this apparently straightforward conclusion masks one obvious caveat. If preapproval trials are shorter than the necessary exposure period or do not administer a sufficient dose, toxicities would not be observed regardless of stopping criteria. The utility of toxicity data also may be limited by a lack of external validity or by inadequacies in data monitoring and analysis, but these limitations exist regardless of stopping criteria. Even if we assume that preapproval trials would capture all of these toxicities, the question remains of how much additional information would be gained by forgoing early termination.

The amount of toxicity information that could be obtained by delaying termination is unknown and unknowable a priori. However, phase III trials for life-threatening conditions often include mortality as an end point and are expected to follow up on a large number of participants for several months or even years. Early termination of these trials for benefit may reasonably be expected to shorten trial length by a year or more$^{14–16}$ and thereby decrease the toxicity data obtained. So, using statistically extreme stopping criteria in these cases might increase toxicity information, subject to 2 additional caveats.

First, requiring highly significant evidence in some phase III trials would not have yielded additional toxicity information. Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure (MERIT-HF),$^{15}$ Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS),$^{17}$ and Cardiac Insufficiency Bisoprolol Study II (CIBIS-II)$^{16}$ are examples of trials in cardiology in which statistically extreme evidence of benefit was attained. In these trials, the evidence developed so rapidly between scheduled DMC meetings that termination before accumulation of highly significant evidence was unfeasible. For similarly efficacious therapies, statistically extreme evidence may be expected to develop regardless of initial stopping guidelines.

Second, one of the important considerations involved in the early termination of trials is previous clinical experience with the intervention. The β-Blocker Heart Attack Trial (BHAT) is an example of a phase III study that was terminated for benefit with a nominal probability value on the primary outcome of $0.005,^{14}$ which is already substantially less than a nominal probability value of 0.05. If the observed mortality trend had been allowed to continue until the probability value reached 0.0001, the trial would have been an estimated 92% longer. Trial continuation in this case might have been expected to yield a corresponding increase in the possibility of discovering unknown toxicities.

However, propranolol was approved as an antihypertensive before BHAT, so the safety profile was already established in a population reasonably comparable to the targeted study population. The probability of detecting an unknown toxicity with trial continuation was probably small. Smaller phase III trials are sometimes terminated with moderate evidence of benefit when the safety profile is unknown, but in these cases, relatively few patient-years of follow-up would have been gained with trial continuation.$^{18}$

As these examples illustrate, using statistically extreme stopping criteria may lead to additional toxicity information on longer trials of therapies for which the toxicity profile is not well established. For the requirement to be applicable, the evidence of benefit must develop gradually so that the DMC has the opportunity to examine the data before extreme evidence is attained. The next question is whether toxicities with a bearing on future approvability would be observed under these conditions.

The Third Premise

(3) If additional toxicity data were obtained as a result of requiring statistically extreme evidence of benefit, toxicities that had a bearing on future approvability might be observed that would otherwise remain hidden.

For the toxicity of an agent to have a bearing on future approvability, the toxicity must be serious enough to alter the risk-benefit profile substantially and must be detectable during clinical trials. For a serious toxicity to be detected, 1 of 2 things must occur: Either the agent must be associated with a toxicity that is rare in the population under study, or it must substantially increase the frequency of a common toxicity relative to its comparator. If the study population were adults >60 years of age, a few cases of a rare AE such as rhabdomyolysis might be enough to raise safety concerns. If the AE were myocardial infarctions, only a substantial imbalance in the number of events between treatment groups might be considered indicative of toxicity. Examples of both scenarios are considered below.

Cerivastatin was an HMG Co-A reductase inhibitor (statin) that was withdrawn from the market in August 2001 because of high reporting of rhabdomyolysis associated with its use. A database study of claims from large health plans indicated that to observe 1 case of hospitalized rhabdomyolysis, 1873 patients needed to be treated with cerivastatin for 1 year.$^{19}$ If hundreds of additional person-years of exposure could be gained by forgoing early termination as suggested previously, we might believe that the probability of observing 1 or more
cases of rhabdomyolysis is reasonably high and might there-
participant entry is staggered, so there would be no uniformity to the individual follow-up periods that were lost. Although requiring statistically extreme evidence may entail additional person-years of exposure, it is less clear that it would entail more information about a particular toxicity.

Even if additional person-years of exposure would invari-
ably result in relevant toxicity information, the exposure necessary to influence drug approval is substantial, as illustrated in the Figure. This figure was generated by assuming a dichotomous response variable and computing a fixed number of AEs using a standard binomial probability distribution. Specifically, the calculations assume a 90% probability of detecting at least 1 to 5 or more AEs occurring at rates ranging from 1:10 to 1:10,000. When each estimate is multiplied by 1 person-year and the simple assumption noted earlier is made, the estimates are also equal to the number of person-years of additional follow-up necessary to detect AEs 90% of the time at each rate. A more complicated model could be generated using an exponential function for time to events, but the conclusion would be the same: An untenably large number of participants are required to detect very rare AEs.

As noted earlier, a database study suggests that 1873 patients need to be treated with cerivastatin for 1 year to observe 1 case of rhabdomyolysis. From the above calculations, \( \approx 4300 \) participants would have to be treated for 1 year to detect at least 1 case, 10,000 to detect at least 3 cases, and 15,000 to observe at least 5 cases. Only 1 statin trial was large enough to detect these outcomes. Although the incidence of rhabdomyolysis is greater with combined statin-fibrate therapy than with monotherapy, participants taking the combination in preapproval trials were at most a subgroup of those in the treated arm. In either case, statin-induced rhabdomyolysis would probably not have been observed in preapproval trials frequently enough to affect approvability.

Other drugs withdrawn since 1990 also had AEs that were too rare or incidental to be detectable even with longer-term clinical trials and thus would not have been detected regardless of stopping criteria. Bromfenac-associated hepatotoxicity occurred in only 1 in 20,000 patients taking the drug for \( > 10 \) days. About 50 reports of any serious AE relating to temafloxacin occurred in the 6 months after the drug was marketed, although perhaps 500,000 patients received the drug. Even if only 1 in 5 of these very serious adverse effects were reported, the incidence of AEs would still only have been 1 in 2000. Terfenadine, mibefradil, astemizole, and cisapride were all recalled primarily because of drug interaction–induced toxicity rather than because monotherapy with the drug was unsafe.

We now turn to the scenario concerning the detection of an increased relative frequency of a common toxicity. Rofecoxib-associated myocardial infarctions are consistent with this profile. As with the statin example, trial length and design considerations would obviate application of the requirement for shorter trials. In larger trials, the difference in the rate of AEs between arms of the trial would have to be unexpected, reasonably large, and practically detectable. As another example illustrates, this is not always the case.

The antibiotic grepafloxacin was withdrawn in 1999 because of ventricular arrhythmias stemming from a prolonged QT interval, even though clinical trials of grepafloxacin did not exclude patients at risk for QT interval prolongation. QT prolongation is a common side effect of approved drugs in many classes, including other fluoroquinolones. Although effects vary between drugs, a summary of animal QT interval data related to fluoroquinolone use suggests that grepafloxacin was comparable to other marketed fluoroquinolones. Some QT prolongation in grepafloxacin trials might have been expected and probably allowable. Finally, assessments of the mean change in QTc intervals in clinical studies suffer from a number of technical limitations. Even if preapproval trials had been long enough to use statistically extreme stopping criteria, it is still not clear that grepafloxacin toxicity sufficient to bar approval would have been detected.

In longer-term phase III studies that include major morbidity or mortality as outcomes, it is more probable that early termination for benefit could reduce the probability of detecting toxicities with a bearing on approvability. However, DMCs look carefully at accrued data before trials are terminated and, regardless of early termination guidelines, would probably not recommend discontinuation for benefit if the efficacious study intervention were also associated with an increased risk of AEs. Therefore, the requirement for extreme evidence would be important only if treatment-associated toxicity were absent or minor at the time statistically moderate evidence of benefit was realized.

Slightly different questions arise when the reason for extending the trial is to obtain a more precise statistical estimate of an anticipated risk. In a trial of a new anticoagulant, for example, it might be important to characterize the risk of uncontrolled bleeding. The underlying assumption relating to approvability decisions is that a precise estimate of the risk would afford an accurate risk-benefit assessment and that an accurate assessment would prevent approval of a drug with side effects sufficient to warrant later withdrawal.

The veracity of the assumption is more uncertain than is initially evident. For the requirement to yield valuable information, the trial must be large and the toxicity relatively

Approximate number of participants necessary to observe AEs 90% of the time that are expected to occur with frequencies of 1 in 10 to 1 in 10,000.
frequent but not frequent enough to warrant termination for harm. Of the drugs that have been withdrawn since 1990, only 2 had primary indications that would clearly support long-term phase III studies and had a toxicity that was relatively frequent. The requirement per se would not have prevented approval in either case. Encainide proved so toxic in the Cardiac Arrhythmia Suppression Trial that the trial was terminated for harm.22 Similarly, flosequinan was withdrawn in 1993 after the interim results of the PROFILE trial indicated that patients taking the drug had an increased risk of hospitalization or death.30 Termination criteria for benefit in these cases would have been immaterial. Therefore, by definition, the relatively imprecise estimates obtained in trials to date have been adequate to ensure that drugs such as the anticoagulant in the above example were rightly approved. It is unclear why future trials merit a different standard.

Even if precise estimates of risk were necessary, a well-known limitation of clinical trials is the generalizability of results. Issues such as compliance, comorbidities, genetics, concomitant medications, age, and socioeconomic status may mean that trial volunteers are different from nonvolunteers.31–33 Some cardiovascular studies have suggested comparability between trial results and outcomes in clinical practice,34,35 whereas others found substantial variation.31,36–38 Systematic differences may exist in the quality of care received by participants in clinical trials compared with patients in general practice,38,39 and physicians’ adherence to evidence-based practice guidelines also may be uneven.40,41 For all these reasons, risk estimates generated in a clinical trial may be expected to vary in a different population. Although trial results may indicate the general acceptability of the risk-benefit profile, a mere increase in the precision of the estimate would not necessarily make it robust.

Finally, the uncorroborated results of a single long-term trial that suggests the toxicity of a new drug, although worrisome, might not impede approval unless the evidence of harm was unequivocal or the risk-benefit profile was clearly unacceptable relative to other treatment options. The interim results of the Adenomatous Polyp Prevention on VIOXX (APPROVe) trial that prompted the rofecoxib withdrawal46 appeared momentous in part because prospective trials, database studies, and meta-analyses had previously suggested an increased risk of cardiac events in a drug marketed for pain relief.

In summary, requiring statistically extreme evidence would increase the likelihood of observing toxicity data with a bearing on approvability only in select circumstances. The trial must be long enough to permit interim analysis. Early termination would have to substantially reduce the number of toxicities observed. For the requirement to yield useful information, a rare toxicity must be sufficiently common with drug treatment to be observed in the period after the DMC would have otherwise terminated the trial. Alternatively, a greater number of events in the treated arm would have to occur during the same period. Any imbalance in the number of AEs also would have to be small enough when moderate evidence of benefit was achieved that the DMC would recommend termination in the absence of the extreme evidence requirement. Finally, to fulfill the third premise, the toxicity information observed as a result of the requirement must be sufficiently compelling to impede approval.

The Fourth Premise

(4) The policy of requiring statistically extreme evidence of benefit for the purpose of preventing the approval of therapies with unobserved toxicities is coherent.

The availability of alternatives and risk-benefit determinations lie at the heart of the decision to market drugs. Despite previous knowledge of drug interactions leading to cardiac arrhythmias, for example, terfenadine was withdrawn only after fexofenadine, the active metabolite, became available. Astemizole was explicitly withdrawn because of the availability of other prescription antihistamines.42 Although ≈1 in 1000 people taking alosetron is expected to have serious constipation and 1 in 350 to have ischemic colitis, it was reintroduced to the market in 2002 because of the paucity of other effective therapies for irritable bowel syndrome.43 Therefore, even if serious AEs were observed because of the extreme evidence requirement, those toxicities would affect approvability only pending an assessment of the overall risk-benefit profile.

Additionally, lengthening all trials to impose this requirement would entail a delay in dissemination of the vast majority of therapies that are helpful, because only ≈2.5% of approved drugs are later withdrawn.44 From a purely utilitarian standpoint, more people would be harmed by delaying dissemination of all drugs for a year than by releasing a small number of drugs with undesirable risk-benefit profiles. Previously recalled drugs have been from varying classes, for varying indications, with toxicities that may be undetectable regardless of any realistic increase in the length of preapproval trials.

Finally, there are issues to consider even if the imposition of the requirement were limited to studies that are currently long enough to sustain multiple interim analyses. Trial continuation in the face of a substantial difference in apparent efficacy entails that some participants would continue to receive a therapy that is likely to be inferior. Many trials in which questions of termination arise also have serious or irreversible outcomes as end points. For additional evidence to accumulate, participants on the less effective regimen would necessarily continue to suffer those outcomes at a higher rate.

Toward a Coherent Policy

For reasons of generalizability noted earlier, the range of AEs observed in a clinical trial is likely to be artificially small. Policy discussions must therefore extend to the optimization of active postmarketing surveillance efforts, which are clearly necessary to capture AEs that are not likely to be observed in clinical trials. In the wake of the rofecoxib withdrawal, these complex issues have been discussed substantively.45,46 At issue here is what stopping criteria provide the optimal marriage of safety information and proof of efficacy given the inherent limitations of preapproval trials.

The concern about unduly foreshortening clinical trials is not new. Peto stopping boundaries proposed in 1976 require probability values for early termination in the range of 0.001.
Most commonly used boundaries require highly significant evidence for termination early in the trial,\textsuperscript{12} both to guard against termination in the face of fluctuating early trends and to ensure some minimum of follow-up data. Adequate follow-up was a particular concern in MERIT because of substantial questions about the safety of β-blockers in patients with heart failure. The trial was designed with 2 separate criteria that were met before termination: crossing of a conservative upper boundary for benefit and a prespecified requirement that at least 50% of the total number of expected deaths were observed.\textsuperscript{48}

Because the stated purpose of the requirement for extreme evidence is to ensure adequate follow-up, the most direct approach is to add a second criterion to early termination guidelines in lieu of requiring extreme evidence, as was done in MERIT. This additional criterion could be customized to suit the needs of an individual trial and might be based on follow-up time, number of events, or even some limit on the toxicities observed. A second termination criterion for the hypothetical anticoagulant trial noted earlier might stipulate that the rate of major bleeds was within a clinically acceptable range, analogous to a noninferiority design. Other benefits of the second criterion approach include increased transparency of the study aims and enhanced flexibility to terminate the trial on the basis of a range of safety and efficacy considerations.

Limited circumstances might exist in which allowing trials to progress beyond usual significance levels is reasonable, as when outcome measures trend in opposing directions. In the Prostate Cancer Prevention Trial, for example, finasteride significantly reduced the incidence of prostate cancer, but a greater proportion of men who developed cancer had high-grade disease.\textsuperscript{49} Notably, even in this complex scenario, the trial was terminated with probability values of 0.001. Additional evidence probably would not have changed treatment recommendations. Although there may be trials in which requiring statistically extreme evidence could provide toxicity information with a bearing on approvability, these situations are more limited than is initially evident, and widespread application of the requirement may have undesirable consequences.\textsuperscript{50}

Acknowledgments

The author gratefully acknowledges thesis advisors David L. DeMets, PhD, and Norman Fost, MD, MPH, for their helpful comments and suggestions on the manuscript, as well as Thomas Cook, PhD, for his guidance on statistical questions.

Disclosures

None.

References

30.  
46. Roth-Cline Clinical Trials in the Wake of Vioxx 2259.

**KEY WORDS:** clinical trials    ethics    statistics
Clinical Trials in the Wake of Vioxx: Requiring Statistically Extreme Evidence of Benefit to Ensure the Safety of New Drugs
Michelle D. Roth-Cline

Circulation. 2006;113:2253-2259
doi: 10.1161/CIRCULATIONAHA.105.604512
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2006 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/113/18/2253

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/