Monitoring the ACTIVE-W trial: Some issues in monitoring a noninferiority trial
John A. Cairns, MD,a Janet Wittes, PhD,b D. George Wyse, MD, PhD,c Janet Pogue, MSc,d Michael Gent, DSc,d Jack Hirsh, MD,e JOHN MARLER, MD,e and Edward L.C. Pritchett, MDf

Vancouver, British Columbia; Calgary, Alberta, and Hamilton, Ontario, Canada; Washington, DC; Bethesda, MD; and Durham, NC

Noninferiority comparisons of new to current therapies and the use of composite outcomes represent significant advances in the design of clinical trials. They increasingly characterize trials of new cardiovascular agents, posing new challenges to Data Safety Monitoring Boards (DSMB) and principal investigators. The ACTIVE-W study was a noninferiority comparison of the combination of clopidogrel and acetylsalicylic acid versus oral anticoagulant among patients with atrial fibrillation. When unexpectedly high rates of stroke and then of the composite outcome of stroke, non–central nervous system systemic embolism, myocardial infarction, and vascular death emerged, the DSMB modified its monitoring plan and conducted its first formal interim analysis much earlier than had been planned in the DSMB charter. The study was terminated when only 27% of the anticipated outcomes had occurred. This paper discusses issues of appropriate stopping guidelines for noninferiority trials and the early emergence of significant harm in relation to one component (stroke) of a composite outcome. Conditional power was not determined concurrently with the HRs during the monitoring of ACTIVE-W; however, the members of the DSMB now believe that such calculations should be considered as useful adjuncts to the calculation of HRs and could lead to earlier termination of noninferiority trials whose interim results suggest futility, without the need for convincing proof of harm. [Am Heart J 2008;155:33-41.]

The ACTIVE study comprises 3 separate interrelated trials to evaluate the combination of clopidogrel plus aspirin (C/ASA), oral anticoagulation (OAC), and irbesartan in patients with atrial fibrillation (AF).1 ACTIVE-W used a noninferiority design to compare C/ASA to OAC. The principal investigators (PI) terminated ACTIVE-W on the recommendation of the Data and Safety Monitoring Board (DSMB) when only 27% of the anticipated events had occurred, because of unacceptably high rates in the C/ASA group of the primary outcome, a composite of stroke, non–central nervous system systemic embolism, myocardial infarction, and vascular death. In this paper, the members of the DSMB review some challenges inherent in monitoring for safety and futility in noninferiority trials that we faced in monitoring ACTIVE-W. We also discuss dilemmas that arose from the emergence of a large difference between the treatment arms for stroke, one clinically important component of the primary composite outcome, at an information fraction when the data were not showing convincing evidence of a difference for the composite outcome itself.

The ACTIVE-W trial
ACTIVE-W compared C/ASA to OAC among OAC-eligible patients with AF, in whom OAC is the standard of care. The PIs anticipated that C/ASA would be noninferior to, that is, “not unacceptably worse than,”2 OAC with respect to the composite outcome, and that, compared to OAC, the C/ASA would be easier to use clinically and would lead to a lower rate of major bleeding.1 The trial was designed to randomize 6500 patients in a 1:1 ratio to the 2 treatment arms and to follow them for an average of approximately 3 years, with a minimum of 1450 adjudicated primary outcomes and approximately 90% power to declare noninferiority if C/ASA maintained 50% of warfarin's efficacy. Selection of the noninferiority margin in trials of this type poses challenges. Too narrow a margin leads to an impractically large sample size; too wide a margin may lead to declaring a new treatment “noninferior” when it is actually considerably less effective than the standard therapy. Following a frequently used standard
methodology, the PIs chose the noninferiority margin for ACTIVE-W in 2 steps. First, from a meta-analysis of all 6 available randomized controlled trials of adjusted-dose OAC in AF (Dr Carl Van Walraven, University of Ottawa, unpublished), they calculated an estimated hazard ratio (HR) of 1.73 comparing control (placebo or usual care) versus OAC for the composite outcome (95% CI 1.37-2.18). Selecting the lower limit of the 95% CI (HR 1.37) gave a conservative estimate of the benefit of OAC over placebo. An HR of 1.186 (ie, 1 + 0.372/2), equivalent to preservation of 50% of the benefit of OAC over placebo, was accordingly set as the noninferiority margin (Figure 1). Thus, if at the end of the trial the estimated HR of C/ASA to OAC were statistically significantly <1.186, then C/ASA was to be declared noninferior to OAC. When an agent is found to be noninferior to the comparator, the point estimate of the HR is usually <1.0 (Figure 1, examples 2a and 2b); however, the point estimate of the HR may be above 1.0 if the CIs are sufficiently narrow and the noninferiority margin is sufficiently >1.0 (Figure 1, example 2c). The problems in selecting noninferiority margins, conducting noninferiority trials, and interpreting their results have been thoughtfully reviewed.

The DSMB and PIs agreed upon a detailed charter, which for ACTIVE-W included extremely stringent stopping guidelines for noninferiority of C/ASA to OAC as follows (Figure 2, panel A):

Greater than expected (and early) efficacy: an estimated treatment benefit that lies more than a specified number of standard deviations from the
noninferiority margin (4.0 SD at ½ of the anticipated events and 3.0 SD at ¾ of the anticipated events, on at least 2 consecutive time points, 3 months apart).

Given the undesirability of allowing the trial to continue until it unequivocally demonstrated substantial inferiority of C/ASA, the DSMB charter expressed the stopping guidelines for harm in terms of deviation of the point estimate of the HR from 1.0 rather than from the noninferiority margin (Figure 2, panel B) as follows:

Futility/safety: there is clear evidence for harm, for example, a deviation of the point estimate of the hazard ratio for clopidogrel/ASA versus OAC exceeds 1.0 by more than 3 SD at either of the formal analyses done when ½ and ¾ of the expected number of total events has accumulated.

**General considerations in monitoring superiority trials**

The DSMB is responsible for protecting the interests of the participants in a trial, typically monitoring for serious adverse events and for an unexpected strongly negative trend for the major study outcome. If at an interim analysis with a sizeable number of events, the data show a very small effect size or point in the direction of harm, the likelihood is low that the trial will demonstrate statistically significant benefit and the experimental agent may, in fact, be harmful. In such a case the DSMB may consider recommending early termination.
The DSMBs of most trials are also concerned with whether a new agent is more efficacious than (ie, superior to) a placebo or the current standard of care. Typical formal statistical stopping guidelines allow early termination with protection of the prespecified type I error rate if interim data from the trial show much greater than expected efficacy. The rationale for stopping a trial early for unequivocal efficacy includes the opportunity to provide effective new therapies to patients as soon as possible and the potential to reallocate human and financial resources. One important consequence of stopping a trial early is that the resulting estimates of effect size and competing risks may be imprecise.

Superiority trials may undertake different types of comparisons, each of which will affect the DSMB’s decision making during monitoring. In the most common instance the trial will be comparing a therapy which is new or not in widespread use to placebo. Here the interest is in proving efficacy; there is generally neither interest in nor justification for proving harm. Therefore, DSMBs typically use asymmetric stopping guidelines that require strong evidence to stop early for benefit but much weaker evidence to stop for harm. When interim data from a trial of a new agent or one not in widespread use show that the new therapy is likely to be significantly inferior to placebo, early termination is desirable.

A second type of trial compares a commonly used but unproven agent to placebo. If the interim data suggest the agent is causing harm, the DSMB is in a particularly difficult situation and must think particularly carefully about continuation, regardless of any prespecified stopping guidelines. The DSMB’s primary responsibility is the safety of patients in the study; however, if the evidence of harm is not convincing, practice may not change. The CAST and the estrogen/progesterone component of the WHI provide 2 such examples. In both cases, the primary outcomes were irreversible serious clinical events: death in the CAST and myocardial infarction in the WHI. Without convincing evidence of harm, cardiologists would have been likely to continue using antiarrhythmic drugs for patients with asymptomatic arrhythmias after myocardial infarction and the widespread use of hormone replacement therapy might not have changed.

Figure 3

The outcome of stroke in ACTIVE-W. The HR of 1.0 and the noninferiority margin are indicated as in Figure 1. The HRs and their 99.7% CIs (3 SD) calculated from the data available to the DSMB at each of its meetings including at the time of the first formal interim analysis (August 2005) and at the conclusion of the study (November 2005).
A third type of trial compares 2 therapies in general use in the absence of rigorous comparative data. A trial that demonstrates the superiority of one agent logically implies inferiority of the other (although it may still be superior to placebo). Such demonstration may be readily justified when the outcomes are relatively benign. Even for irreversible clinical outcomes like stroke or myocardial infarction, such an approach may be justified because only unequivocal evidence of superiority of one of the agents is likely to change medical practice.

Stopping guidelines for safety generally take into consideration the potential for adverse events (eg, major bleeding), requiring the DSMB to assess competing risks and benefits of the new therapy. In addition, formal stopping guidelines are commonly in place that address the possibility of inferiority of the new therapy. If the new therapy is being compared to a placebo, a finding of inferiority means that the new therapy is actually increasing the rate of the outcome and causing harm. There is no need to prove that a new agent is inferior to placebo. If one could predict at some point before completion of the trial that, if completed, the trial is unlikely to conclude the new treatment is more effective than the placebo, then the trial could be stopped for “futility”. Stopping early would avoid ongoing inconvenience to patients, waste of resources, and, in extreme cases, actually proving harm. The challenge is to avoid premature termination, which risks failure to detect a true benefit. Stochastic curtailment is one approach that may aid in this decision. It involves calculating conditional power, which is the probability of rejecting the null hypothesis given the results observed at a given point in the trial under specified assumptions about subsequent results were the trial to continue. Calculation of conditional power allows quantification of the chance of detecting a true benefit of the new therapy.
therapy if the trial stops early and may provide guidance for decisions about stopping for futility or potential harm before harm is actually proven.

**General considerations in monitoring noninferiority trials**

After regulatory approval of a therapy, a related new therapy may be developed that offers potential advantages (eg, greater convenience, lower cost, or an improved side-effects profile). Regulatory approval of the related new therapy requires evidence that its effect on an important clinical outcome is not importantly inferior to that of the proven standard agent. The noninferiority trial paradigm is increasingly used to provide such evidence. Whereas proof of efficacy in a superiority trial is gained from evidence that the HR of the new therapy versus placebo is significantly <1.0 by a clinically important amount (Figure 1, example 2a), proof of efficacy in a noninferiority trial comes from evidence that the HR of the new therapy versus proven standard therapy is significantly less than a prespecified value (the noninferiority margin) which is >1.0 (Figure 1, examples 2b and 2c). The question asked is not whether the new therapy is superior to the proven standard therapy, but whether the new therapy is “noninferior to” (ie, not unacceptably worse than) the proven standard therapy. To calculate the noninferiority margin, the investigators must first estimate the benefit of the proven standard therapy over placebo, preferably from an overview of the results of prior randomized placebo-controlled trials (Figure 1, example 1). The investigators assume “constancy,” that is, they assume the previously observed benefit of a now-approved therapy compared to placebo is still relevant to current circumstances. They next judge how much of this benefit the new agent must preserve to conclude that the new therapy is acceptable, that is, “noninferior.” On study completion, efficacy is judged by comparing the calculated HR and its CI to the noninferiority margin (Figure 1).

Our experience in ACTIVE-W suggests that monitoring for harm and for futility in noninferiority trials is more difficult than in superiority trials. By analogy with superiority trials, to consider recommending early termination for efficacy, the DSMB would require unequivocal evidence in early data that the new treatment is noninferior to the proven standard treatment. The noninferiority margin is a reasonable choice for the guideline for early efficacy (Figure 2, panel A). However, the noninferiority margin may be inappropriate as a reference for early futility or harm, given that the comparison is with proven standard therapy. A point estimate of the relative risk lying far to the right of (worse than) the noninferiority margin means that the data have shown statistically significant evidence of substantial harm with the new treatment. We believe a more appropriate strategy relates the observed HR and its confidence interval to an HR of 1.0, rather than to the noninferiority margin (Figure 2, panel B). Even so, such a requirement would still mean that the data would have to show convincing evidence of harm before the trial would be stopped for futility. Therefore, the case for considering formal calculations of futility is even stronger than it is for superiority trials.

**The ACTIVE-W monitoring experience: early harm, early futility, and the composite outcome**

The first patient in ACTIVE-W was randomized in June 2003. At the DSMB’s second meeting in October 2004, the available data suggested a higher event rate on C/ASA than on OAC, with an observed HR for stroke of 2.34 (Z = 2.38) (Figure 3), although the HR for the composite outcome was only 1.34 (Z = 1.5) (Figure 4). Somewhat surprisingly, for major bleeding, the HR for C/ASA versus OAC was 0.79 (Z = −0.39).

When the board reviewed the data again on January 17, 2005, the HR for stroke was 1.88 (Z = 2.50) and the HR for the composite outcome was 1.42 (Z = 2.36) (Figures 3 and 4). C/ASA continued to show numerically less major bleeding (HR 0.86, Z = −0.77). Concerned for the safety of the participants, the DSMB engaged the PIs, who were not provided with any treatment-specific outcome data, in a general discussion about monitoring noninferiority trials and possible courses of action. The possibilities included conducting the first formal interim analysis earlier than specified in the DSMB charter and using various adjusted z values, and modifying the composite outcome to include major bleeds. The board decided to conduct its first formal interim analysis using the total of adjudicated and unadjudicated composite outcomes when approximately 25% of the anticipated events had occurred.

When the DSMB met again in August 2005, the HRs for stroke, for stroke or nonsystemic embolus, and for the composite outcome were 1.97 (Z = 3.73), 2.18 (Z = 4.49), and 1.49 (Z = 3.66), respectively (Figures 3 and 4). Now for the first time the data showed numerically more major bleeding with C/ASA (HR 1.09, Z = 0.54). Although in accepting the formal stopping guidelines the DSMB had not anticipated the need for an analysis when only 25% of the anticipated events had occurred, the DSMB concluded that there was an extremely low likelihood that C/ASA would be found to be noninferior to OAC if the trial continued. Moreover, the DSMB was concerned that the data were showing increasingly strong evidence of harm. Detailed review of the HRs for all components of the composite outcome and of many additional outcomes failed to provide a rationale to continue the study. The DSMB therefore recommended termination of ACTIVE-W.
and, after discussions, the Steering Committee promptly implemented this recommendation. ACTIVE-W had another feature that made monitoring challenging—its composite outcome. Clinical trials use composite outcomes to increase power, to decrease the required sample size, or to decrease the duration of follow-up. In addition, a composite outcome may provide a more complete assessment of the efficacy of the active therapy. Interpreting the results of a trial and making sensible clinical decisions based upon them can be difficult if the clinical importance of various components of the composite differs substantially. Weighting the various outcomes may help to guide clinical decision making. Clearly, death will be more important than nonfatal myocardial infarction, and both will be more important than a transient ischemic attack. However, it is not clear how to weigh the relative importance of various components of the composite differently. Weighting of the various outcomes may help to guide clinical decision making. Clearly, death will be more important than nonfatal myocardial infarction, and both will be more important than a transient ischemic attack. However, it is not clear how to weigh the relative importance of various components of the composite differently. Weighting of the various outcomes may help to guide clinical decision making.

The complexities in interpreting composite outcomes and their components came into sharp focus during the monitoring of ACTIVE-W. The DSMB of ACTIVE-W had to evaluate early strong evidence for increased stroke with C/ASA when the results for the composite outcome were much less certain. As indicated earlier, the primary efficacy outcome of ACTIVE-W was a composite of stroke, non-CNS systemic embolism, myocardial infarction, and vascular death. In designing ACTIVE-W, the investigators anticipated that the relative risk of C/ASA to OAC might vary among the individual outcomes, but that C/ASA would be noninferior to OAC for the composite outcome. They further expected that C/ASA would lead to lower rates of major bleeding than would OAC, although the composite did not include this outcome.

The DSMB's charter did not address the early emergence of harmful trends in individual components of the composite outcome, nor did it establish a formal hierarchy among them. Insufficient data were available on disability and quality of life to influence the DSMB decisions. Although major bleeding had been expected to be less frequent with C/ASA than with OAC, by the time of the August interim analysis and at the conclusion of the study, the data showed numerically more major bleeding with C/ASA (HR 1.09, Z = 0.54 and HR 1.10, Z = 0.63, respectively).

Reflections on the DSMB's monitoring and decisions

The DSMB's stopping guideline for harm for C/ASA specified that the point estimate of the HR lies 3.0 SDs to the right of 1.0. The first formal analysis of efficacy was conducted when only 24% of the expected events (adjudicated and unadjudicated) had occurred. In retrospect, however, these guidelines were problematic, for they required strong evidence that C/ASA was inferior to OAC. Given that C/ASA was not widely used in patients such as those in the trial, and that there would have been no case for using the new therapy unless it were noninferior to OAC, we members of the DSMB have discussed how guidelines that include calculations of conditional power might better protect patients in future noninferiority trials.

At a given information fraction, conditional power in this case would be the probability of declaring noninferiority of C/ASA to OAC at the end of the trial. Calculating this probability requires assumptions about future data. Two reasonable approaches present themselves. One method would have been to have assumed that the true HR was indeed 1 and to have calculated, under this assumption, the probability of declaring noninferiority at the end of the trial. As Table I shows, by November 2005 the conditional power under this assumption was 38%, whereas before November it was 50% or higher. Thus, had we calculated conditional power under this set of assumptions and used the calculation of conditional power to complement our calculations of the z value associated with the observed HRs, we would have had stronger evidence for our recommendation of early termination after the August formal interim analysis, although such evidence would not have been likely to prompt us to make such a recommendation earlier than we did.

Another approach, and one we prefer, would have been to project the future trend on the basis of an HR calculated from the data at the time of each meeting. Table I shows 4 columns depicting projections based on the observations. The column that projects into the future assuming that the observed HR is the true parameter would lead to very low conditional power. We would have not used this projection because it fails to account for variability. At the end of the trial, we would be looking at the upper end of the 95% CI to establish whether the data were consistent with noninferiority. During the trial, we probably would have elected to project the as yet unobserved data on the basis of the lower end of a CI.
Table I. Conditional power as a function of HR

<table>
<thead>
<tr>
<th>Date of DSMB meeting</th>
<th>Information fraction (proportion of expected events)</th>
<th>Assumed HR = 1</th>
<th>HR, z value, and CP</th>
<th>Lower bound of 80% CI</th>
<th>Lower bound of 90% CI</th>
<th>Lower bound of 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr '04</td>
<td>0.02</td>
<td>1.00</td>
<td>0.95 1.45 0.95 0.00</td>
<td>0.88 1.00 0.76 1.00</td>
<td>0.67 1.00</td>
<td></td>
</tr>
<tr>
<td>Oct '04</td>
<td>0.07</td>
<td>1.00</td>
<td>0.93 1.34 1.49 0.00</td>
<td>1.04 0.88 0.97 0.99</td>
<td>0.91 1.00</td>
<td></td>
</tr>
<tr>
<td>Jan '05</td>
<td>0.13</td>
<td>1.00</td>
<td>0.84 1.42 2.36 0.00</td>
<td>1.17 0.52 1.11 0.69</td>
<td>1.06 0.81</td>
<td></td>
</tr>
<tr>
<td>May '05</td>
<td>0.21</td>
<td>1.00</td>
<td>0.51 1.56 3.85 0.00</td>
<td>1.35 0.12 1.29 0.17</td>
<td>1.25 0.22</td>
<td></td>
</tr>
<tr>
<td>Aug '05</td>
<td>0.24</td>
<td>1.00</td>
<td>0.50 1.49 3.66 0.00</td>
<td>1.30 0.20 1.25 0.25</td>
<td>1.20 0.31</td>
<td></td>
</tr>
<tr>
<td>Nov '05</td>
<td>0.27</td>
<td>1.00</td>
<td>0.38 1.51 3.95 0.00</td>
<td>1.32 0.13 1.27 0.17</td>
<td>1.23 0.21</td>
<td></td>
</tr>
</tbody>
</table>

The table shows the date of each meeting of the DSMB. The DSMB did not meet in May 2005, but the italicized line provides retrospectively calculated data for this point in time. Follow-up concluded in November 2005. Beside each date is the information fraction (the ratio of actual to anticipated outcomes). The next pair of columns show an assumed HR = 1.0 for each point in time, with the conditional power calculated on that assumption. For each date, the next 3 columns show the observed point estimate of the HR, its z value, and CP. The lower bound of the 80%, 90%, and 95% CIs of the observed HRs and the conditional power calculated for each. CP, Conditional power.

Had we chosen the lower bound of the 80%, 90%, or 95% CI as a reasonably optimistic estimate of the effect of C/ASA relative to OAC, by January 2005 we would have seen conditional power between 52% and 81%, and might well have recommended a formal interim analysis earlier than at the information fraction of 24%, the point we actually chose. Although we recognize that we cannot know what we would have done, we have considered the action we might have taken had we chosen to perform the interim analysis at an information fraction of about 21% (May 2005). At that time, we would have found a z value of 3.85 and conditional power ranging from 12% to 22%. With this combination of calculations, we might well have recommended early termination at that time, about 3 months earlier than the actual date of the recommendation.

We now believe that complementing the calculation of z values with an analysis of conditional power would have increased the statistical formality of our deliberations about the possibility of recommending early termination and have strengthened our conviction in August to recommend early termination. The observation of low power for detecting noninferiority in January might even have precipitated an earlier formal interim analysis which might have led to an earlier recommendation to terminate the study.

In monitoring a trial, the DSMB relies heavily on the timely acquisition of outcome data. In trials that, like ACTIVE-W, have many clinical centers, reports of outcomes are often delayed by many months. Therefore, in considering how to improve the monitoring of such trials, especially if decisions about early termination might be made early in the trial, the investigators must devise methods for rapid reporting of outcomes.

Our combination of biostatistical and clinical judgment led us to believe that protection of the subjects enrolled in the trial required us to recommend terminating ACTIVE-W, even though our charter had not anticipated the circumstances observed. This experience has led us to conclude that the discipline of clinical trials needs a more appropriate method for monitoring noninferiority trials. We suggest considering stochastic curtailment and the calculation of conditional power or some other measure of futility which would allow the quantitative prediction of the likelihood of a positive outcome (in this case, a finding of noninferiority) when a given HR is observed at a given information fraction. If the likelihood of observing a positive outcome is extremely low, then a DSMB could recommend stopping the trial at less cumulative harm to those patients in the trial than was the case with the approach that we used, which was based on finding an HR a given number of SDs from 1.0. Potential downsides to be considered in calculating conditional power and acting on the results include the potential to increase the probability of a type II error and the increased likelihood of early termination of studies that are unlikely to achieve conventional statistical significance, but could contribute useful data to future meta-analyses.

The PIs chose the composite outcome for this trial to capture the aggregate effects of both C/ASA and OAC on individual serious outcomes that were expected to respond differentially to the 2 regimens. They anticipated that stroke and non-CNS systemic embolus would be more frequent with C/ASA, but the observed frequency required the DSMB to develop additional monitoring and stopping guidelines. By the time of the first formal analysis of efficacy, the z values for stroke and for the
composite outcome both exceeded the guidelines set before the review of unblinded data and the appropriate course of action was clear. The appropriate recommendation would have been unclear, however, had the HR for the composite outcome continued to show borderline clinical significance in the face of highly significant HRs for the components of stroke and non-CNS systemic embolus.

Statistical theory alone is an insufficient guide to decision making around these latter issues and must be complemented by the clinical judgment and experience of the DSMB members. Therefore, DSMB charters must allow the DSMB to consider complex clinical factors, leading to the possibility that the DSMB and the investigators may not agree. The DSMB charter for the ACTIVE trials specified procedures for managing such disagreement, although it was not necessary to implement them.

The advantages of noninferiority comparisons of new therapies to current therapies, and the rationale for the use of composite outcomes, ensure that these design features are increasingly likely to characterize trials of new cardiovascular therapeutic agents. Appropriate monitoring of such trials will present DSMBs with new challenges to ensuring a reasonable balance of the risks to patients in the trial and the interests of those in the wider community.

References