21. Alliance for Cervical Cancer Prevention. Effectiveness, safety, and acceptability of cryotherapy: a systematic literature review. Available at: http://www.path.org/files/RH_cryo_white_paper.pdf. Accessibility verified October 5, 2005.
22. Cox JT. Management of cervical intraepithelial neoplasia. *Lancet*. 1999;353:857-859.
23. Samson SL, Bentley JR, Fahey TJ, McKay DJ, Gill GH. The effect of loop electrosurgical excision procedure on future pregnancy outcome. *Obstet Gynecol*. 2005;105:325-332.
24. Goldie SJ, Kuhn L, Denny L, Pollack A, Wright TC. Policy analysis of cervical-cancer screening strategies in low-resource settings: clinical benefits and cost effectiveness. *JAMA*. 2001;285:3107-3115.
25. Crum CP. The beginning of the end for cervical cancer? *N Engl J Med*. 2002;347:1703-1705.
26. Skjeldstad FE, Koustsky L; for the Merck Phase III HPV Vaccine Steering Committee (FUTURE II). Phase III trial of prophylactic quadrivalent HPV 6, 11, 16, 18 L1 virus-like particle (VLP) vaccine: prevention of cervical intraepithelial neoplasia (CIN) 2/3 including adeno- and squamous-cell carcinoma in situ (CIS). Presented at: Infectious Diseases Society of America Late Breaker Session 66, LB-8a; October 7, 2005; San Francisco, Calif.
27. Blumenthal PD. Immunization against cervical cancer: Who? When? Where? Available at: http://www.medscape.com/viewarticle/444979. Accessibility verified October 5, 2005.

When (Not) to Stop a Clinical Trial for Benefit

Stuart J. Pocock, PhD

In this issue of *JAMA*, Montori and colleagues[1] provide a valuable extensive and critical systemic review of clinical trials that were stopped early for benefit. Readers of the reports of such trials often feel a sense of excitement, especially when phrases such as "a major treatment advance," "ethical need to stop the inferior treatment," and "vital to tell the world immediately" are used. However, experience suggests that early results and enthusiasm, especially for modestly sized trials terminated early for apparent major benefit, are often moderated as subsequent reports arise.[2]

The skeptic should ask first whether correct and appropriate structures were in place for analyzing and reviewing, and making decisions based on, the trial's accumulating interim data. Having the members of an effective independent data monitoring committee (DMC) or data and safety monitoring board as the only individuals accessing and interpreting interim data split by treatment group is now considered an essential part of good practice for major randomized trials.[3-5] Still, a substantial minority of reported major trials appear not to have a DMC in place.[6]

Second, with or without a formal DMC recommendation, another question is whether the decision to stop a trial early and report the results was an appropriate judgment. This decision should be aided by a predefined statistical stopping boundary for a primary outcome,[7-9] but some trials have no such guideline. It is important that such a boundary is sufficiently stringent (eg, very strong evidence of a treatment difference with a very small *P* value) to match the ethical and public health implications of a decision to stop the trial. In a spirit of requiring proof beyond reasonable doubt that a treatment difference is sufficient to affect future clinical practice, some lenient statistical boundaries are not a sensible choice in the direction of benefit. For instance, the so-called Pocock boundary[9] and the O'Brien-Fleming boundary's last interim look[9] both typically require values around $P = .02$ for stopping, which is usually insufficient strength of evidence to stop a trial for benefit. Both boundaries can be made more appropriate if the overall type I error is set at 1% rather than the conventional 5%.

Many complex methods exist for statistical stopping boundaries, whereas in practice there is considerable merit in the simple Haybittle-Peto boundary,[9] which requires $P < .001$ as evidence required to consider stopping a trial early for benefit. Even so, such a boundary should not be applied too soon, when few outcome events have been observed.

Decisions on early stopping (or not) need to be based on wise judgments interpreting the totality of available evidence, both in the current trial (considering primary and other efficacy outcomes and safety issues) and in other external evidence (especially from related trials).[10] Accordingly, a statistical stopping boundary is only one useful objective component in an inevitably more challenging decision-making process. The ethical dilemma is to safeguard the interests of patients randomized in the current trial while also protecting society from overzealous premature claims of treatment benefit.[11] For instance, if a trial is evaluating a treatment meant to be given long-term for conditions such as hypertension or chronic arthritis, short-term benefits, no matter how statistically significant, may not merit early stopping. If a trial is for regulatory approval, the sponsor and trialists should be encouraged not to stop early unless there is overwhelming evidence of treatment superiority, since the regulators require substantial evidence of both efficacy and

See also p 2203.

Author Affiliation: Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, England.
Corresponding Author: Stuart J. Pocock, PhD, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, England (stuart.pocock@lshtm.ac.uk).

safety, often in at least 2 trials reaching their intended full size and patient follow-up.

Montori et al[1] rightly draw attention to some reports of trials that were stopped early but that did not document the planned size and circumstances of the relevant interim analysis and stopping boundary. Such deficiencies need correcting by authors, peer reviewers, and editors in line with CONSORT recommendations.[12] Indeed, journals should consider rejecting the report of any trial potentially stopped prematurely and lacking adequate documentation, and access to trial protocols by journals would help in making this decision. There is probably less need to present adjusted analyses that attempt to correct for the interim monitoring and early stopping, since stopping depends on more than a statistical boundary, and complexities of adjustment can clutter the presentation of results and make interpretation of the findings more difficult. Real insight rests more on a full understanding of the circumstances at the time of stopping. Also, between the moment of making the decision to stop and locking the final database used for analysis and publication, substantial additional and corrected data may become available for analysis. Indeed, such data cleaning may justify a pause before any definite decision to stop the trial.

From a reader's perspective, the key problem is whether to believe the treatment benefit is truly as great as the data imply. Montori et al[1] appropriately emphasize that trials stopping early will tend to be on a "random high" of observed benefit, and if further data had been collected in either this or another trial, some "regression to the truth" to a more modest effect estimate would occur.[2,13] These issues are more pronounced in smaller trials.

Montori et al reported a median of 66 events observed at the time trials were stopped. To achieve a difference between treatment that is significant at $P<.001$ requires a split by treatment group of at least 46 vs 20 events, which means that risk happens to be reduced by 57% or more. In most therapeutic areas, this is highly implausible and is often associated with relatively short patient follow-up time. Thus in many settings, trials should not stop so soon, because it is highly likely that the therapeutic claim is exaggerated.

The data monitoring experience in the CHARM program in 7599 patients with heart failure provides a thought-provoking example.[14] At the fourth interim analysis with a median 1-year follow-up, there were 260 vs 339 deaths in the candesartan and placebo groups, respectively, a 24% risk reduction that crossed the $P<.001$ stopping boundary. For several documented reasons,[14] the DMC voted to continue until the next interim analysis. The treatment mortality difference was then attenuated in subsequent interim analyses so that at the trial's intended completion with a median of 3.1 years of follow-up, there were 886 deaths in the candesartan group vs 945 deaths in the placebo group, a 9% risk reduction ($P=.055$). Early stopping was resisted, and hence an exaggerated claim of survival benefit was avoided and important long-term benefits in other outcomes, such as cardiovascular death and heart failure hospitalization, were realized in each of the 3 component trials of the CHARM program.

So when is it appropriate to stop a trial early? The ASCOT factorial trial's data monitoring experience provides useful insights.[15,16] First, in 10305 patients with hypertension, the comparison of atorvastatin with placebo was halted when the difference in the primary end point, major coronary events, at interim analysis reached $P<.001$, the stopping boundary. With 100 vs 154 primary events in the atorvastatin and placebo groups, respectively, and a risk ratio of 0.64 ($P=.0005$), the published result was clear-cut.[15] The appropriateness of stopping early was supported by other trials of statins in other populations and by important benefits in other outcomes, such as stroke.

A more difficult stopping decision arose in the ASCOT trial for the 19342 patients randomized to receive amlodipine-based and atenolol-based regimens. The predefined primary end point was major coronary events, whereas it is well known that the key effect of antihypertensive treatment is in reducing risk of stroke. Thus, when there emerged a highly significant reduction in stroke for amlodipine-based compared with atenolol-based treatment ($P<.001$), much debate ensued on whether to stop the trial, resulting in a decision to continue to the next interim analysis. Some months later, the trial was stopped early when there was also a significantly higher rate of mortality in the atenolol-based group, although still no significant difference existed for the primary end point. This example illustrates the complexities and tough decisions that can arise in data monitoring.[17]

Can a trial be stopped on the basis of secondary end points? Perhaps not, but on occasion, such as with the ASCOT-BPLA study, results of secondary end points (327 strokes with amlodipine vs 422 with atenolol, a 23% risk reduction [$P=.0003$]) provide convincing evidence of great public health importance.[16] In lay terms, "when early results proved so promising it was no longer fair to keep patients on the older drugs for comparison, without giving them the opportunity to change."[18] However, the data in these 2 examples are more substantial compared with those in the majority of trials reviewed by Montori et al. The message is clear: most trials stopped early for benefit should not have been stopped at that point. Stopping for harm or futility is another matter[19] that equally importantly requires future systematic review and comment. Inappropriate stopping of trials for commercial reasons raises additional serious concerns.[20]

In summary, all major randomized trials should have an independent DMC that functions effectively and makes wise judgments aided by stringent statistical stopping boundaries for benefit. It is critical that the DMC, principal investigators, executive committees, and sponsors all recognize

the full public health implications of their recommendations and decisions.

Financial Disclosures: None reported.

REFERENCES

1. Montori VM, Devereaux PJ, Adhikari NKJ, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294:2203-2209.
2. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218-228.
3. Ellenberg S, Fleming T, DeMets D. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. Chichester, England: John Wiley & Sons; 2002.
4. Draft guidance for clinical trial sponsors on the establishment and operation of clinical trial data monitoring committees, 66 *Federal Register* 58151-58153 (2001).
5. DAMOCLES Study Group. A proposed charter for clinical trial data monitoring committees: helping them to do their job well. *Lancet*. 2005;365:711-722.
6. Sydes M, Altman DG, Babiker AB, Parmar M, Spiegelhalter DJ; DAMOCLES Study Group. Reported use of data monitoring committees in the main published reports of randomised controlled trials: a cross-sectional study. *Clin Trials J*. 2004; 1:48-59.
7. O' Brien P. Data and safety monitoring. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Chichester, England: John Wiley & Sons; 1998:1058-1066.
8. Fleming TR, Harrington DP, O'Brien PC. Designs for group sequential tests. *Control Clin Trials*. 1984;5:348-361.
9. Schulz KF, Grimes DA. Multiplicity in randomised trials, II: subgroup and interim analyses. *Lancet*. 2005;365:1657-1661.
10. Brocklehurst P, Elbourne D, Alfirevic A. The role of external evidence in monitoring clinical trials: reflections from a perinatal trial. *BMJ*. 2000;320:995-998.
11. Pocock SJ. When to stop a clinical trial. *BMJ*. 1992;305:235-240.
12. Moher D, Schulz KF, Altman DG; CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285:1987-1991.
13. Pocock S, White I. Trials stopped early: too good to be true? *Lancet*. 1999;353: 943-944.
14. Pocock S, Wang D, Wilhelmsen L, Hennekens CH. The data monitoring experience in the Candarsartan in Heart failure Assessment of Reduction in Mortality and morbidity (CHARM) program. *Am Heart J*. 2005;149:939-943.
15. Sever P, Dahlof B, Poulter NR, et al; ASCOT Investigators. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial—Lipid Lowering Arm (ASCOT-LLA): a multicentre randomised controlled trial. *Lancet*. 2003;361:1149-1158.
16. Dahlöf B, Sever PS, Poulter NR, et al; ASCOT Investigators. Prevention of cardiovascular events with an antihypertensive regimen of amlodipine adding perindopril as required versus atenolol adding bendroflumethiazide as required, in the Anglo-Scandanavian Cardiac Outcomes Trial—Blood Pressure Lowering Arm (ASCOT-BPLA): a multicentre randomised controlled trial. *Lancet*. 2005;366: 895-906.
17. DeMets DL, Furberg CD, Friedman L. *Data Monitoring in Clinical Trials: A Case Studies Approach*. Heidelberg, Germany: Springer; 2005.
18. Hall C. Heart attacks may be cut by half. *Daily Telegraph*. September 5, 2005:1.
19. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. *Lancet*. 1999;354:1983-1988.
20. Psaty BM, Rennie D. Stopping medical research to save money: a broken pact with researchers and patients. *JAMA*. 2003;289:2128-2130.

