much is being asked of statisticians. New scientific approaches that accomplish rapid new drug development seem unlikely to exist. Although statisticians will make improvements, limited data sets only yield limited amounts of information. We cannot change that without making extensive unverifiable assumptions. Further, we often find ourselves in the position of making decisions concerning study conduct that should involve more extensive input from clinicians and others. We must make sure that expectations of statisticians remain reasonable and balanced.

We welcome the opportunity to make these additions to an excellent discussion of the issues facing us in evaluating therapeutic interventions.

Comment

David L. DeMets

I appreciate the opportunity to comment on this paper by Professor Fleming and want to compliment him on a timely and very relevant discussion of current issues in clinical trials.

In general, I agree with Professor Fleming's key points, so my remarks are similar in spirit, based on my experience with cardiovascular clinical trials and, more recently, with cancer and AIDS trials. In particular, I will comment on two points: the data monitoring committee and surrogate outcomes.

Clinical trials play an important role in the long and complex process to develop and evaluate new drugs, devices or procedures. Because patients are involved, ethical issues as well as scientific and economic factors must be considered in the design, conduct and analyses. In order to establish a model for conducting such trials, the National Heart Institute in the 1960's formed a committee chaired by the late Professor Bernard Greenberg. This committee's report, typically referred to as the Greenberg Report (Heart Special Project Committee, 1988), became the framework for NIH-sponsored cardiovascular trials as well as many other disease areas. One of the first trials to implement this model was the Coronary Drug Project (Coronary Drug Project Research Group, 1981). A key component to this clinical trial model was the data monitoring committee (DMC), an independent body not directly participating in the conduct of the trial at the clinic level and charged with the responsibility of patient safety as well as monitoring accumulating data for early evidence of benefit. If either treatment safety or benefit becomes convincing, consideration should be given for early termination. The Coronary Drug Project foresaw that this decision process would be very complex and formed a committee with a diversity of expertise. The complexity of this monitoring process and the need for this expertise is best illustrated by reading accounts of several examples of the data monitoring experience (Coronary Drug Project Research Group, 1981; DeMets et al., 1982, 1984; Cairns et al., 1991). This model has now been used in dozens of trials, especially in heart, lung, blood, eye and cancer. Recently, the NIH AIDS clinical trials groups also adopted a variation of this model.

Looking back on over 25 years of experience with this data monitoring committee, I would argue strongly that it has been very successful. Where it has not been used, problems have often occurred, as Professor Fleming points out. I would also argue that this clinical trial model should be used for any comparative (Phase III) trial that is pivotal and has either mortality or irreversible morbidity as a primary outcome.

One demand of this monitoring process not always appreciated is the need for a timely and reasonably clean data base, at least for the critical endpoint and safety variables. Not having current data could lead to incorrect or inappropriate decisions and inferences, a process almost experienced by the Nocturnal Oxygen Therapy Trial (DeMets et al., 1982). In addition, we cannot always anticipate the direction or rapidity in which convincing trends emerge. Such an example is provided by the Cardiac Arrhythmia Suppression Trial (Cardiac Arrhythmia Suppression Trial Research Group, 1989), a trial briefly discussed by Professor Fleming for which I served on the data monitoring committee. With less than 10% of the expected number of deaths, the results were already trending strongly in a negative direction. The DMC requested the statistical center to contact all clinical sites and obtain up-to-date mortality data before the critical meeting of the DMC. Fortunately, the statistical center was able to provide such analyses, even at this early stage. Results were even more convincing with the up-to-date data, and the trial was stopped, declaring the treatment to be harmful. It would have been much more difficult, perhaps impossi-

David L. DeMets is Chair, Department of Biostatistics, and Associate Director, Comprehensive Cancer Center, 6775 Medical Sciences Center, 1300 University Avenue, University of Wisconsin, Madison, Wisconsin 53706.

ble, to make such a decision with data several months old. Although reasonably up-to-date data are necessary, it does not come without planning. Data acquisition, processing and analysis systems must be in place at the start of the trial. The DMC must also be in place from the start and be familiar with the protocol and analysis plans. Although such activity costs money, not having such a structure could be even more costly.

Over the past 15 years, dozens of papers have been published on statistical procedures for monitoring interim data. Although useful and essential, none sufficiently reflect the complexity of the decision-making process and thus must be viewed as helpful guidelines but not strict rules (Coronary Drug Project Research Group, 1981; DeMets, 1984). As described, several issues must be considered in a careful review of interim data. For example, the DMC must rule out that other factors are not responsible for the emerging trends or, in some cases, lack of any trend. Randomization should produce comparable treatment groups, but this must be evaluated using baseline covariates. Compliance to therapy and the protocol, completeness of data, unbiasedness of patient evaluation, internal and external consistency and the risk-to-benefit ratio are but a few of the issues to be considered. Repeated significance testing or data evaluation, and multiple outcomes are important factors but certainly not the only ones. Many randomized clinical trials now utilize the data monitoring statistical methodology but in the context of a DMC.

A second major topic addressed by Professor Fleming is that of surrogate markers or outcome variables. The basic idea is to substitute a surrogate outcome (e.g., exercise tolerance or arrhythmia suppression) for the relevant clinical outcomes (e.g., nonfatal heart attack, morbidity, AIDS). From my experience, this is the most disturbing, even threatening, issue today in clinical trials. As stated in the paper, "one can rarely establish that surrogate endpoints are valid." Statistical regression models often used to identify markers are only models, establishing association and not causality, and even then surrogates usually explain only a small proportion of the variability of the clinical outcome. Three recent examples in cardiology are sobering. The first, and certainly the most dramatic, is provided by the CAST study already mentioned (Cardiac Arrhythmia Suppression Trial Research Group, 1989). Arrhythmias are associated with sudden death, and it was widely believed that suppression of these arrhythmias would prevent or reduce the incidence of sudden deaths. Drugs were approved and in widespread use based on their ability to modify a surrogate outcome, arrhythmias. CAST confirmed that these drugs do suppress arrhythmias. Unfortunately, these drugs suppressed some of them permanently, resulting in a higher rate of sudden death and total mortality. This is a classic example of a drug successfully modifying the surrogate but having a negative effect on the clinical event of most interest.

Another example is provided in congestive heart failure. A promising new drug, milrinone, had been shown to have beneficial effects on exercise testing – a possible surrogate outcome in this class of patients. Yet a clinical trial called PROMISE (Packer et al., 1991) clearly showed this drug to be harmful for general use in congestive heart failure patients, based on mortality as the primary outcome. In this case, however, milrinone was not yet widely used.

A third example is in the use of clot busters or biologics with thrombolytic activity in evolving or imminent heart attacks. The basic idea is for these biologics to break up the blood clots that can cause heart attacks, reperfuse the heart and prevent any damage to the myocardium or heart muscle. Reperfusion rates or time to reperfusion might be thought of as a surrogate outcome. A clinical trial (TIMI Study Group, 1985) established that one agent, TPA, had a higher reperfusion rate than another, streptokinase. Yet a very large multicenter trial (Second International Study of Infarct Survival Collaborative Group, 1988) could find no mortality difference between these two agents. Although this issue is still the subject of ongoing debate and research, it seems obvious that cardiology cannot depend on reperfusion rate alone to evaluate thrombolytic therapy.

Recently, in the pressure to find new effective therapies in the treatment of AIDS, surrogate outcomes such as CD4 counts are being considered. In fact, approval of therapies have already been based on this surrogate. Yet our knowledge of CD4 as a valid surrogate is limited. We do not fully understand if the therapies affect the patient in other ways, probably not well studied. In the desire to seek rapid solutions to treatment of the AIDS epidemic, shortcut methods may not prove to be as desirable as we hope. There also seems to be discussion to even expand the use of surrogates, despite examples such as those from recent cardiovascular clinical trials. This trend is quite worrisome. Where would future patients with arrhythmias or congestive heart failure be without a trial such as CAST with the more clinically relevant outcome? Must we hastefully establish as a standard of care a therapy on an invalid or marginal surrogate and then years later discover it had little or no real benefit or, worse yet, produce harm?

What is frustrating is not that we seek better clinical trial methodology but that we do not seem to learn from our past. Furthermore, we as statisticians must clearly point out to our clinical colleagues the limitations of our mathematical models with which they, like us, can become fascinated and forget how little we can depend on them for establishing true valid surrogates.