as primary clinical trial endpoints. Although I fundamentally share his concerns, I would cite one important aspect of the AIDS context that is relevant to the use of such endpoints in trials designed to obtain marketing approval for new drugs. HIV-infected patients, unlike patients recovering from myocardial infarction or suffering from chronic granulomatous disease, will inevitably die of their disease within a short time relative to their otherwise expected remaining lifetime. The best we can hope for from current therapies is a modest to moderate prolongation of survival. In this circumstance, it does not seem inappropriate to accept a higher level of risk in deciding what therapies might be made available. Whether therapies that have only shown positive effects on early markers should be distributed in "expanded access" or "parallel track" programs, or whether the FDA should permit their manufacturers to market them, may be more of an economic than a scientific issue. Whatever mechanism is used, it will ultimately fall to federally funded research programs of the Public Health Service to mount trials that compare available regimens and move toward defining optimal treatment strategies for patients at various stages of disease. In these trials, it will be essential to study clinical efficacy—that is, physical rather than laboratory manifestations of disease—until and unless we discover markers that come much closer to meeting the Prentice criteria.

It is encouraging to learn of the innovative investigations by Fleming and colleagues of the potential use of the auxiliary information present in early markers of disease to strengthen evaluation of therapies when only limited long-term clinical data are available. As Fleming notes, the circumstances under which this type of approach will significantly add to our ability to assess treatments reliably are somewhat limited. Nevertheless, it would be of interest to test out such approaches in data sets in which the relationship between the surrogate and the "true" endpoint is fairly well characterized—for example, if S were blood pressure and T were heart attack or stroke. The problem is complicated in AIDS because there has been experience with relatively few treatments and therefore little data regarding the correlation between S and T in the presence of different therapies. If this correlation varies greatly according to the particular regimen being administered, it would be difficult to use this approach in any routine way.

In conclusion, I would like to congratulate Dr. Fleming for highlighting some of the issues biomedical statisticians are struggling with, and hope that his paper will inspire more statisticians to become actively involved in, and even leaders of, the process of planning and carrying out medical research programs.

Comment

Vern T. Farewell and Richard J. Cook

INTRODUCTION

In this paper, Dr. Fleming provides an excellent review of some current methodological problems facing health scientists involved in clinical trials. Some issues considered in detail are monitoring clinical trials, the analysis of equivalence trials, multiple endpoints and surrogate markers. We will remark on each of these in turn.

MONITORING

The examples cited clearly demonstrate the importance of a monitoring committee for moderate to large-scale sequential clinical trials. In particular, a specialized and centralized Data Monitoring Committee (DMC) for the AIDS Clinical Trial Group (ACTG) is discussed. Such a specialized monitoring committee has immediately obvious advantages. As more trials are passed through the DMC, the disease-specific knowledge gained from early trials can be applied to later studies.

In principle, there are a variety of other diseases that require DMCs. For fields with less trial activity and experience, it may be advantageous to provide access to less specialized DMCs. Although it may be necessary to supplement the available expertise for individual trials, this more general DMC could provide statistical expertise on monitoring and advice on termination to a wide range of clinical investigators. Such a committee, perhaps under the sponsorship of a funding agency, would help to make the most efficient use of available research funds.

Vern T. Farewell is Professor and Richard J. Cook is Ph.D. student, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

In monitoring a single study, a DMC must often deal with the influence of ongoing studies of related therapies or similar studies in slightly different patient populations. This situation might be alleviated if larger stratified trials focusing on different subgroups of the population could replace several smaller studies. For example, recent HIV-positive therapy trials evaluating continued use of AZT versus a switch to DDI were undertaken in separate patient populations with different levels of CD4 counts. A single coordinated study may have been helpful. However, monitoring problems may still be difficult in stratified trials with separate stratum analyses, since blinding of the results may be difficult to maintain once one stratum reaches statistical significance and is terminated. Investigators may have to be more proactive in ensuring sufficient accrual rates and minimize protocol violations to maintain validity and quality in the subsequently collected data.

The urgency associated with HIV research has resulted in some Phase III trials which involve therapies that have not passed through the more traditional stages of drug development and testing (i.e., Phase I and Phase II trials). This creates a risk of potentially toxic treatments being applied to large groups of patients. Although there are problems with subjectivity and "soft" outcomes in toxicity analyses, this is now receiving more attention (Peace, 1990), and explicit consideration of toxicity at the design stage may encourage further improvement. Assuming a reliable and reproducible measure of drug toxicity, Cook, in submitted work, has proposed a modified sequential design allowing formal monitoring of efficacy and toxicity outcomes. Unlike established sequential methods for multiple response data involving global measures of treatment effects (Tang, Gnecco and Geller, 1989; Lin, 1991), the proposed design maintains the individuality of the component test statistics. When two responses are as distinct as efficacy and toxicity, it is undesirable to resort to global measures. The procedure operates as follows.

Let θ represent a measure of the relative efficacy of the new treatment to the standard therapy and μ represent a measure of the relative toxicity. Suppose one is interested in testing H_{01}: $\theta = 0$ versus H_{a1}: $\theta \neq 0$ and H_{02}: $\mu \leq 0$ versus H_{a2}: $\mu > 0$ and one is willing to stop early due to at least one of the following reasons: (i) greater efficacy, (ii) lower efficacy or (iii) greater toxicity.

As in the univariate Fleming, Harrington and O'Brien (1984) sequential design, one must specify the overall size a, the maximum number of analyses N and a vector of stopping probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$, where π_i represents the probability of stopping the trial at stage i when $(\theta, \mu) = (0, 0)$.

In the bivariate sequential design, it is further necessary to specify the probability of stopping due to the different outcomes. This can be achieved by specifying a vector of conditional probabilities, f, where

$$f_n = pr(\text{reject } H_{01} \text{ at stage } n \mid \text{stopped at stage } n)$$

Note that one can select π and f such that $f^t\pi = a_\theta$ to allow the marginal efficacy analysis to operate with size a_θ at $(\theta, \mu) = (0, 0)$. As in the univariate case, the power of the analyses depends on the group size. Selecting the group size that satisfies the most stringent marginal power specifications will satisfy both conditions.

The primary advantages of this procedure are that it allows power and sample size calculations for both responses and provides a formal justification for early termination due to treatment effects on efficacy or toxicity responses, or both. In fact, it is a general procedure for sequential analysis of bivariate response data when early termination is desired based on either or both outcomes. As in the univariate case, it should be viewed as providing a guideline rather than a strict rule for termination.

ACTIVE CONTROL EQUIVALENCE TRIALS

When an effective treatment is available, it is sometimes desirable to find new treatments that are equally effective but less toxic. Thus, both standard and experimental therapies are active. Fleming (1990) outlines an approach for the analysis of such a trial that involves specification of a point θ on the relative efficacy axis that is termed the point of "equivalent therapeutic index" and a distance δ that determines the range of therapeutic equivalence. These are determined by prior knowledge of the relative toxicities and costs of the two treatments. However, there may be difficulty in determining θ and δ when little is known about the experimental therapy. Again, an analysis based on both efficacy and toxicity responses could be useful in this setting. One could consider an equivalence-based analysis for the efficacy response. Depending on the frequency and relative severity of the toxicity response, one may or may not want an equivalence analysis based on this outcome. A conservative Bonferroni-type adjustment or correlation adjustment can be made to ensure that the analyses do not result in an overall type I error rate greater than a. In ongoing work, we are exploring extensions of this approach to a sequential design with repeated confidence intervals and confidence regions.

MULTIPLE MEASURES

Recently, there has been an increased interest in the design and analysis of clinical trials with more than one response. Multiple measures of treatment effect can provide more detailed and descriptive information about the relative performance of the experimental

therapy. However, in a trial based on multiple endpoints, one must again address the problem of a potentially inflated type I error. To address this, one can construct global test statistics. This is an appealing approach when there are several endpoints that are related in some manner since it often provides more power for detecting small but consistent treatment effects across several outcomes. O'Brien (1984) introduces a method of analysis for such multiple response data that is based on a statistic formed by taking a linear combination of scores from a GLS analysis of treatment effects on each outcome. Wei and Lachin (1984) propose a global test statistic for multivariate failure time data. These approaches were subsequently extended to allow a sequential design by Pocock, Geller and Tsiatis (1987) and Lin (1991), respectively. Relevant issues in the sequential analysis of multivariate failure time data are the choice of the (possibly data-dependent) weights used in constructing the global test statistic and the choice in the timing of the analyses. Analyses could be timed based solely on events from one response or based on requirements for the frequency of events from both responses.

Although global measures are often more sensitive to smaller effects, there is some difficulty in the interpretation of the final test statistics, particularly with data-dependent weights. An alternative is to perform univariate analyses adjusted for the multiple responses. This is perhaps the most descriptive approach although it is only really feasible when the number of endpoints is small, say two or three. Methods for normal data are well established (Pocock, Geller and Tsiatis, 1987), and the distribution theory outlined by Lin (1991) allows for correlation-adjusted marginal significance testing in a failure time setting.

SURROGATE MARKERS

Fleming's discussion on the use of surrogate endpoints in clinical trials clearly indicates the dangers of extrapolating treatment effects from analyses based on invalid surrogates to the primary endpoint. Machado, Gail and Ellenberg (1990) demonstrate this problem by a simulation study in the setting of HIV-positive therapy trials. Nevertheless, there is still considerable pressure to adopt designs based on surrogate endpoints.

A valid surrogate for one treatment may not be for another. Also, a surrogate may be valid for one patient group but not for another. For example, a temporary rise in CD4 counts may have clinical significance for patients with low counts but may be of limited value for patients with initially higher counts. Since such situations cannot always be anticipated, methodology is required that allows evaluation of the utility of a potential surrogate and incorporation of the estimated surrogate treatment effect into some final analysis as appropriate. We are currently investigating such an approach that appeals to the asymptotic multivariate normal distribution theory outlined by Lin (1991) and involves correlation-based weights for a potential surrogate and a primary endpoint analysis. The approach does allow investigators to prespecify bounds on the weighting scheme to some extent. The idea is that highly correlated test statistics indicate a useful surrogate and result in an increased weighting of this endpoint. Conversely, uncorrelated or negatively correlated endpoints indicate a surrogate marker with a treatment effect that is inconsistent with the primary endpoint. This results in a decreased weighting of the surrogate responses.

Let $Z_1(t)$ and $Z_2(t)$ represent standard log-rank statistics at time t for the treatment effect on the surrogate marker and the primary endpoint, respectively. A global test statistic can be defined as

$$R(t) = p_1(t)Z_1(t) + p_2(t)Z_2(t),$$

where $p_1(t)$ and $p_2(t)$ are possible data-dependent weights. If $\rho(t)$ represents the correlation of the test statistics at time t, possible weights are of the form

$$p_1(t) = j_1 + \rho(t)$$

$$p_2(t) = j_2 - \rho(t)$$

where j_1 and j_2 are chosen to reflect the relative weighting of the marginal test statistics assuming zero correlation. Simulations have been performed to evaluate the results of such a weighting scheme with the timing of the analyses dictated by the primary endpoint events. The timing of the analyses is particularly relevant here, as a sufficient number of events must occur for both endpoints to properly assess the correlation. Results indicate that even with such correlation-based weights, there is still a danger in the use of marker responses if there is the possibility of discordant treatment effects. It appears that the choice of j_1 and j_2 is the driving factor in the analysis, since $\rho(t)$, derived from Lin's (1991) asymptotic covariance matrix, is not sufficiently sensitive to properly assess the utility of a potential surrogate.

ACKNOWLEDGMENTS

The authors would like to acknowledge support from SIMS through Grant No. DA04722 (U.S. National Institution on Drug Abuse) and from the National Science and Engineering Research Council (NSERC) of Canada.