GROUP SEQUENTIAL ANALYSES

In addition to the need for methods of monitoring multiple measures of therapeutic effect, more research is needed on methods of analyzing trials with three or more treatment arms. In a traditional two-arm trial, the determination that one arm is inferior to the other is tantamount to terminating the trial. Yet in a trial involving three or more treatment arms, the possible decisions to be made at each interim analysis are greater in number and in complexity, and will depend on the overall goals of the trial (e.g., to exclude clearly inferior treatments or to identify the single best treatment). Professor Michael Hughes, a colleague in my department, has made some important initial steps in this area, but additional approaches are needed.

SURROGATE MARKERS

In the field of HIV/AIDS, there is great interest in identification of "surrogate markers." Usually, this refers to a laboratory marker that can be used as a surrogate for a clinical outcome such as the development of an AIDS-defining opportunistic infection. The potential value of a valid marker is obvious, yet, as Professor Fleming notes, use of an invalid marker could lead to the widespread use of ineffective drugs and/or the non-use of effective drugs. Thus, the validation of markers becomes very critical. Some of the early investigations of surrogate markers in HIV/AIDS have attempted to determine whether a treatment's effect on a particular marker can fully explain its effect on clinical endpoints. And, in all cases, only part of the clinical effect could be explained by the effect on markers. In retrospect, this is not surprising because it is unrealistic to expect that any single laboratory marker could fully explain all of an AIDS drug's beneficial effect because of the complex nature of this disease and how it can be affected by intervention. Thus, it may be that a battery of several markers needs to be determined that collectively can explain most of a drug's beneficial clinical effect. Given the high cost of the assays that are needed to evaluate some of the virological and immunological markers in HIV/AIDS, the design of studies to assess the "surrogate marker" question becomes critical. More statistical research on this topic is urgently needed.

In closing, I would like to thank Professor Fleming for his excellent article.

Comment

Thomas A. Louis

INTRODUCTION

Professor Fleming has considerable experience in conducting clinical trials and serving on Data Monitoring Committees (DMCs). We are fortunate that he has prepared a debriefing. It reinforces the impact of statistical science potentiated by subject area expertise and of both technical and broad viewpoints. In complex applications, relevant disciplines must be represented, and statistics is central to the enterprise. As Fleming notes, we must provide strong and effective leadership. To do so, we must educate collaborators on the role of statistics and be educated on a study's scientific and clinical basis. Statistical philosophies, principles and methods (frequentist/Bayes, multiple comparisons, choice of tests and estimators) need to guide deliberations, but in the complex world of clinical trials absolute dictums are seldom appropriate.

Thomas A. Louis is Professor and Head, Division of Biostatistics, University of Minnesota School of Public Health, Box 197 Mayo, Minneapolis, Minnesota 55455.

Statisticians and other DMC members are truly on the line. Stopped trials are very difficult to restart, and the decision to terminate can essentially freeze out other, similar trials. Continuing a trial beyond what many think is a reasonable stopping point puts study participants at unnecessary risk and delays dissemination of important information. Ware (1989) and related discussion show the heat generated by these issues. Contrast this situation to analysis of a stable data base: investigators can analyze, reanalyze, critique other analyses and sustain the give and take for years or decades. A DMC must make important decisions in an acute time frame.

GENERAL DISCUSSION

Data Quality

Building trust with patients and clinicians being recruited for a trial often is, and should be, a sensitive negotiation. Of course, all stake-holders need to be convinced that the question is clinically relevant. Of equal importance is assurance that everyone's interests

will be protected. Timely, accurate and multidisciplinary data monitoring is the keystone of this protection. Data must be accurate and current. Nothing in my statistical experience is quite so unsettling as wondering whether the data supporting monitoring analyses are accurate and up to date. To take an extreme case, consider the impact of a potentially miscoded treatment indicator variable when the DMC has agreed to unblind the study, keep it going if "A" is the placebo, but stop it if "A" is the active intervention. Jacobson et al. (1992) add to the list of trials where timely and accurate data monitoring served the best interests of patients and physicians.

Meta-analysis

A DMC on which Fleming serves reviews both ACTG and CPCRA HIV-related protocols. The committee has been very effective, in part because it has direct access to data from related trials in the two investigative groups and summary information from other, ongoing trials. This information supports both formal and informal meta-analyses that significantly improve the monitoring process. Though Fleming cites examples of problems generated by early release of relative efficacy results, broad communication among DMCs will be beneficial, so long as confidentiality can be maintained. Unfortunately, broad communication and confidentiality are antithetic, and I await Fleming's views on this issue.

The DMCs role as a nexus for meta-analysis should be formalized. Monitoring will benefit a great deal from consideration of a wide body of relevant information. Important sources of information include completed studies, ongoing studies and the prior opinions of a community of stake-holders (designers of the study, eventual consumers of the information). Flexible and robust Bayesian approaches have great potential to structure and document information. For example, investigators treating patients in a trial should not see monitoring information, but each can be "represented" during monitoring by the posterior distribution computed from their prior. The collection of posterior distributions will complement standard analyses. Carlin et al. (1992), Chaloner et al. (1992) and a recent issue of *Statistics in Medicine* (Colton et al., 1992) document recent developments and provide entry points to the literature.

DMCs in Industry

Pharmaceutical companies are beginning to follow the NIH lead and constitute "arms-length" DMCs for their trials. It stands to reason that private industry and the government want to populate these committees with experts – specifically, to seek participation from statisticians with both technical and substantive expertise. Frequently, those asked to serve are also involved in designing and conducting related studies, many times testing modalities developed by the company requesting participation on a DMC.

Increased use of DMCs is a healthy trend, but it does raise a potential conflict of interest. Biomedical research will not be served by restricting statisticians and others to choose between conducting and monitoring trials in a disease area. Full disclosure of activities is a necessary prerequisite to reducing conflict of interest, but disclosure does not eliminate the problem. In addition, statisticians and others should not have a financial interest in the outcome of the study and should receive no more than "customary compensation" for our participataion. "Customary" may be the standard per diem for consultation with industry or the usual federal rate. Some may wish to receive only expenses, and others may wish to have their honorarium given directly to charity. Guidelines and precedents are necessary, with the primary goal preservation of our honest-broker status.

Surrogate Endpoints

Any device to reduce delay in making valid conclusions from a clinical trial is most welcome. Therefore, surrogate endpoints that require shorter follow-up than waiting for the ultimate outcome are very attractive. However, Fleming's and others' examples of the potential for false positives and false negatives give pause. Potential surrogates must be carefully validated (for an example, see Freedman, Graubard and Schatzkin, 1992), but even a valid (though imperfect) surrogate may not produce an overall advantage. For example, the surrogate may require shorter follow-up, but a larger sample size.

To structure an evaluation, consider a parametric model where the scalar θ describes a relation between treatment (Z) and outcome (T). We have the following representation of the joint distribution of (Z, S, T), where S is the surrogate:

$$P_\theta(Z, S, T) = P(Z)P_\theta(S|Z)P_\theta(T|Z, S).$$

We assume that the marginal distribution of the treatment indicator does not depend on θ and that $\theta = 0$ is equivalent to no parametric relation between Z and T: $P_0(T|Z) = P(T)$. Perfect parametric surrogacy is equivalent to $P_\theta(T|Z, S)$ not depending on θ.

Conditioning in two different orders and using the Fisher information decomposition in Louis (1982) gives

$$\begin{aligned} I_{(T,S|Z)}(\theta) &= I_{(T|Z)}(\theta) + I_{(S|T,Z)}(\theta) \\ &= I_{(S|Z)}(\theta) + I_{(T|S,Z)}(\theta). \end{aligned}$$

Equating the right-hand sides and solving for $I_{(T|Z)}(\theta)$ produces:

$$(1) \qquad I_{(T|Z)}(\theta) = I_{(S|Z)}(\theta) + [I_{(T|S,Z)}(\theta) - I_{(S|T,Z)}(\theta)].$$

Notice that when S is a perfect (parametric) surrogate,

$I_{(T|S,\,Z)}(\theta) = 0$ – a highly unrealistic situation. Therefore, perfect surrogacy serves only as an interesting theoretical construct. Equation (1) structures an assessment of the cost/benefit of using a surrogate. If the term in square brackets is negative, then use of the surrogate will be more efficient for inferences on θ than using the ultimate endpoint. Even when this term is positive, the surrogate may still be attractive. Use of it will require additional patients (or events), but total trial duration and person-months on study may be shorter than having to wait for the ultimate endpoint.

Candidates for surrogates abound, but validation is usually elusive. In AIDS research, lab values such as CD4, neopterin and β_2 microglobulin, and disease status indicators such as Karnofsky score, weight and intermediate clinical events are contenders for surrogate status. CD4 currently has top billing, but several challenges remain. The measurement process produces considerable intra- and interlab variability. The true value reacts to short-term infections, can be influenced by smoking and has a pronounced circadian rhythm. In addition, we don't know the best method of using a CD4 *trajectory* to define a surrogate endpoint, and different classes of treatments can have differential effects on CD4, but equivalent therapeutic value. Accruing information from treatment studies linking potential surrogates to long-term follow-up will pin down their status.

Methods Development

A DMC must be ready for anything, and the challenges of monitoring have spawned a variety of methods. Most notable has been introduction of the alpha-spending function, which eliminates the need for specifying in advance the number of monitoring looks. Even with the spending function we can get into difficult and exceedingly unproductive deliberations about how many looks *have been* performed. Fortunately, monitoring boundaries based on a large number of looks (even after each observation) are only slightly broader than the usual, and the broader boundaries should be used.

Fleming presents other exciting recent developments resulting from the wide variety of analyses required for proper monitoring (e.g., multiple measures of treatment effect), the need to react to interesting leads and the need to increase precision (e.g., use of auxiliary variables). Their use in monitoring puts special importance on robustness of validity and efficiency.

CONCLUSION

I enthusiastically thank Professor Fleming for preparing his article. I have learned a great deal and have been energized to give careful thought to technical and broad issues related to clinical trials.

The exigencies of clinical trial design and conduct, especially those associated with monitoring, will continue to seed conceptual and methodologic research that crosses disciplinary and philosophical boundaries. Monitoring and other components of clinical trial design and analysis must balance robustness and efficiency; each trial gets stopped only once. Striking this balance will continue to challenge clinical trialists from all disciplines.

ACKNOWLEDGMENT

Partial support was provided by contract NO1-AI-05073 from the National Institute of Allergy and Infectious Diseases.

Rejoinder

Thomas R. Fleming

MONITORING CLINICAL TRIALS

I appreciate the comments, clarifications and extensions of my distinguished colleagues who have long provided extensive statistical scientific leadership to this area of evaluating therapeutic interventions. I thank the editors for this opportunity for further discussion of some issues related to their comments.

Data Monitoring Committees

The discussants uniformly endorse the concept of Data Monitoring Committees (DMCs), with Professor DeMets specifically advocating their use "for any comparative (Phase III) trial that is pivotal and has either mortality or irreversible morbidity as a primary outcome." With the increasing implementation of such committees pointed out by Professors Ellenberg and Louis, certain issues will need further attention. These include guidelines for membership in various settings and for financial compensation and procedures for expansion of the group of interested, qualified statisticians.

We have stated that DMCs should be "independent," specifically indicating that DMC members should be