$I_{(T|S,\,Z)}(\theta) = 0$ – 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 "free of apparent conflict of interest." The interpretation of this criterion in the context of membership for study investigators is an especially complex issue with proper resolution likely to depend on the setting. Professors Lagakos and Louis, who are leaders of statistical centers for the two government-sponsored AIDS clinical trials cooperative groups, along with Professor DeMets have indicated that study investigators should not be members of the DMC. Although he originally thought it would be "unnecessary and rather silly" to exclude the medical head of a clinical trial from interim results, Dr. Lagakos indicated his experience led to a dramatic change in his perspective. Two reasons he cites for excluding study investigators are the need to reduce ethical conflicts arising when one serves as a primary care physician for individual patients while also addressing more global patient care interests when monitoring evolving trial safety and efficacy results, and the need to avoid frequently arising professional conflicts that occur when the study investigator is also a consultant to the sponsoring pharmaceutical company. Other professional conflicts also exist for study investigators serving on their DMC. For example, their recommendation for premature trial termination could be influenced by the need to improve appearance of productivity before peer review or to bolster their curriculum vitae for more rapid promotion.

The atmosphere can differ considerably across different disease areas, as pointed out by Dr. Ellenberg when she contrasted the AIDS and Oncology settings. This could affect the balance of pros and cons for study investigator involvement on a DMC. In the setting of cancer cooperative groups, Professors Crowley and Green have given strong arguments that providing DMC membership for study investigators, who usually treat few if any of the patients in large multicenter clinical trials, has been a successful approach for more than a decade. Some cancer cooperative groups have also adopted conflict of interest guidelines that regulate the involvement study investigators and DMC members can have with companies sponsoring these trials.

Crowley and Green also note, in their experience where study investigators are DMC members, that "uninvolved" physicians who serve on the DMC too often are "uninterested." In contrast, in our experience in dozens of industry- and government-sponsored trials in which study investigators were not DMC members, the vast majority of "uninvolved" physicians have been consistently committed, knowledgeable, cognitive of the seriousness of their responsibilities and active valuable contributors to the review process. This favorable contrast in leadership provided by the uninvolved physician might indeed be caused by factors such as their inability to rely on study investigators to take the lead role during closed DMC sessions and the higher level of commitment required if they must be willing to travel specifically for the purpose of participation in a DMC meeting.

We look forward to further discussion about the relationship study investigators should have with the DMC. Other issues that should be discussed include guidelines about the compensation DMC members can be provided without inducing significant conflict of interest and procedures to expand the group of interested qualified statisticians to serve on DMCs. Members are entitled to receive compensation for their time and effort required to serve on a DMC, in addition to related expenses. For government-sponsored trials, honorariums provided are small. In the industry-sponsored setting, we agree with Professor Louis that members should receive no more than "customary compensation" for their participation, which he indicates "may be the standard per diem for consultation with industry." Contributing the honorarium to charity, suggested by Dr. Louis as one approach, certainly would reduce levels of conflict. Guidelines should also be discussed about the nature of the professional involvement a DMC member might have with the sponsoring company. The risks for significant conflict of interest would be increased if one served on a DMC for a trial sponsored by a company with which the member has had a lengthy ongoing consulting relationship.

Because DMC members assume considerable responsibility in helping to safeguard the rights of patients and to preserve the integrity and credibility of the trial, it is necessary that study sponsors and investigators be confident about the members' judgment, knowledge and experience. With increasing use of independent DMCs, an organized effort might be undertaken to expand the group of statisticians who are willing to be involved and who are experienced in the DMC process. One approach for rapid expansion would be for lists of interested statisticians to be compiled and made available on request to government or industry sponsors as they determine DMC membership. If these lists provide information about monitoring experience, efforts could be made to place two statisticians, one with and one without prior DMC experience, on as many committees as possible. These lists could be compiled by the PMA in the industry setting and by various institutes in NIH in the government-sponsored setting, or by statistical societies.

Other Monitoring Issues

To reduce the likelihood that prejudgment of early results would compromise the ability of the trial to obtain definitive conclusions, we have advised that members of the DMC should be the only individuals to whom the clinical trial's Data Analysis Center provides results on relative efficacy of treatments during the study. However, we concur with Professor Lagakos that the DMC should be responsive to the needs of sponsors. As recommended by Fleming and DeMets (1992), the DMC "should have procedures to evaluate and act on special requests from study investigators or sponsors to provide them limited access to some evolving study information. These procedures should not unblind non-DMC members to relative efficacy results."

We agree with Professors Farewell and Cook that a centralized DMC that monitors a series of studies in a disease area achieves a breadth of experience which provides obvious advantages to the Committee as it monitors additional studies. Committee members should be knowledgeable about available information external to the trial that is relevant to its decision-making responsibilities. Because important information might be from current blinded studies as well as from completed and published trials, we concur with Professor Louis that situations could arise in which the concurrent review of related clinical trials by two or more DMCs could be benefitted by their sharing some safety and efficacy data from these ongoing trials. Of course, DMCs would need to provide assurances that confidentiality would be preserved for any information received in this manner.

Professors Louis and DeMets provide strong arguments that careful attention should be given by the DMC to the accuracy and currentness of the database and to the clinical relevance of the questions being addressed by the trial. In a recent manuscript (Fleming, 1992), we explore these issues in greater depth and provide several illustrations from recently monitored trials. We indicate that "the role of the DMC can include some important activities which promote obtaining relevant information of high quality. In the final stage of study development, these include review of protocol specified study objectives and design, and review of procedures planned for capturing high-quality data and for reporting to the DMC. At analyses during the trial, these include careful review of data completeness and accuracy to assist in early detection and resolution of evolving problems."

Statistical monitoring guidelines were briefly considered by Crowley and Green, who discuss an informal approach proposed by Haybittle (1971). This approach has the desirable characteristics of simplicity and being conservative at early analyses. Even though we have investigated a more rigorous version of this boundary (Fleming, Harrington and O'Brien, 1984), we observe that the O'Brien-Fleming group sequential guideline has many additional important features (Emerson and Fleming, 1989; Simon, 1990; Fleming and DeMets, 1992). In addition to being conservative at early analyses when one should be very cautious, it is monotonically more liberal as we approach the final analysis, enabling one to obtain greater efficiency. Because final analysis levels are close to those of a fixed sample design, sample size is adequately approximated by that obtained from standard fixed sample design calculations. In addition, a Lan-DeMets (1983, 1989) use function implementation of the O'Brien-Fleming boundary does provide the flexibility needed by a DMC to alter the timing and number of interim analyses and properly adjust if accrual rates are unexpectedly fast or slow, while essentially preserving the type one error of the trial. Although it is true that one must specify the total information in the trial to employ a Lan-DeMets use function approach, this provides an important advantage relative to less rigorous guidelines since it removes ambiguity as to the timing of the "final analysis" and when the type one error can be used in a trial whose duration or accrual differs substantially from what was expected at trial design. Green and Fleming (1988) suggest guidelines for monitoring trials after completing the main phase of the trial during which the type one error of size α was spent. Finally, we note that the O'Brien-Fleming guideline is in a class of boundaries for which unbiased estimates have been developed for treatment parameters following a group sequential trial (Jennison and Turnbull, 1983; Tsiatis, Rosner and Mehta, 1984; Chang and O'Brien, 1986; Whitehead, 1986; Kim and DeMets, 1987; Emerson and Fleming, 1990).

ACTIVE CONTROL DESIGNS

In discussing group sequential designs and active control equivalence trials, Farewell and Cook have discussed the importance of giving careful consideration to both efficacy and toxicity, as well as concerns about resorting to global measures of these distinctly different response variables. One important setting, arising with increasing frequency over recent years, in which proper attention usually has not been given to both toxicity and efficacy is chemoprotection trials. In these trials, an agent is given with intent to reduce the side effects of toxic drugs that are known to provide important clinical benefits. Illustrations of such agents considered by FDA advisory committees over the past year include ICRF-187 to reduce the cardiotoxicity of adriamycin in advanced breast cancer, WR-2721 to reduce the neurotoxicity and myelosuppression of cisplatin and cyclophosphamide in advanced ovarian cancer, and macrophage colony stimulating factors to lessen the myelosuppressive effects of drugs in many oncology settings. The simplest and most cost effective approach to reducing toxic side effects, i.e., simply lowering the dose of the effective drug, usually is unacceptable due to the anticipated or documented reduced benefit obtained at lower doses. To establish a role for a chemoprotective agent, one must document that it selectively protects healthy cells and does not promote growth of tumor cells. This can be done by ruling out that it too will lead to reduced efficacy while providing reduced toxicity. Thus, one first must demonstrate that the agent significantly reduces the side effects of the drug. Second, one must show the efficacy of the combination of the drug and agent is equivalent to that of the drug alone. This second objective of establishing equivalent efficacy, which can be performed using the approach discussed in the main presentation, usually requires much larger sample sizes and longer follow-up than establishing reduced toxicity and very often is not addressed in the design of chemoprotection trials.

SURROGATE MARKERS

The discussants have expressed substantial reservations about therapeutic evaluations that rely on surrogate markers when attempting to evaluate clinical efficacy of treatment interventions. We concur with Professor DeMets, who refers to this reliance on surrogate markers as "the most disturbing, even threatening, issue today in clinical trials." After discussing several illustrations, he expressed frustration that "we do not seem to learn from our past" as we proceed to expand use of surrogate markers in AIDS trials. Crowley and Green, in turn, observe that an unfortunate result of the pressure to use surrogates in AIDS is pressure to approve cancer drugs using surrogate markers, such as tumor shrinkage, which have long been recognized to be unreliable markers for clinical endpoints such as patient survival.

Dr. Ellenberg observes that AIDS differs from some disease areas in that HIV-infected patients "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." There is a sense of urgency about the need to establish more effective treatments for patients infected with the HIV, but we again caution against less rigorous scientific approaches using surrogate markers that will provide rapid yet unreliable results. In cardiology, Hallstrom (1992) conservatively estimates that use of encainide and flecainide may have led to an additional 4,000 deaths per year in the U.S. alone. In AIDS, if surrogate markers are used to replace clinical endpoints, the public health consequences of false negative trials (as nearly occurred in CGD) could also be a staggering occurrence of potentially avoidable morbidity and mortality, whereas misleading results from false positive trials could slow our progress toward development of truly effective therapies and would encourage use of toxic and ineffective treatments that would contribute to spiralling health care costs without providing the desired therapeutic benefit.

STATISTICAL LEADERSHIP

Professors Lagakos, Lewis and Ellenberg have discussed the importance of statisticians increasingly providing strong and effective leadership in medical research. For illustration, one area of need is voting membership on FDA advisory committees. These disease-specific committees have substantial influence on establishing scientific standards for clinical research and on whether new drugs and biologics receive regulatory approval for marketing. Even though a substantial proportion of the most important issues confronted by these committees is statistical in nature, each committee has at most one statistician among their nine to twelve voting members. Opportunities for important and continuous statistical leadership would be substantially improved if membership on each advisory committee was increased to two voting statisticians serving staggered 4-year terms. In Europe, there is even greater need for obtaining statistical voting membership on regulatory advisory and decision-making bodies.

Several discussants have encouraged additional methodologic research to address important medical research problems. For example, one such problem identified by Dr. Ellenberg is that substantial rates of noncompliance to AIDS treatments requiring long-term administration might preclude evaluating long-term measures of clinical efficacy. Following the May 1991 FDA-sponsored "Workshop on Alternative Data Sources in AIDS," Paul Meier and colleagues have been exploring the concept of Large Simple Trials that might address several issues, including those caused by the noncompliance discussed by Ellenberg. Further statistical research in this and other areas of medical research could enable more rapid and efficient evaluation of treatments without compromising the reliability of conclusions.

ADDITIONAL REFERENCES

Cairns, J., Cohen, L., Colton, T., DeMets, D. L., Deykin, D., Friedman, L., Greenwald, P., Hutchison, G. B. and Rosner, B. (1991). Issues in the early termination of the aspirin component of the physician's health study. *Annals of Epidemiology* 1 395–405.

Carlin, B. P., Chaloner, K. C., Church, T., Louis, T. A. and Matts, J. P. (1993). Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *The Statistician.* To appear.

Chaloner, K., Church, T., Louis, T. A. and Matts, J. P. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician.* To appear.

Chang, M. N. and O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials* 7 18–26.

Colton, T., Freedman, L. S., Johnson, A. L. and Machin, D., eds. (1992). *Statistics in Medicine* 11.

Coronary Drug Project Research Group (1981). Practical aspects of decision making in clinical trials: The Coronary

Drug Project as a case study. *Controlled Clinical Trials* **1** 363–376.

DeMets, D. L. (1984). Stopping guidelines vs. stopping rules: A practitioner's point of view. *Commun. Statist. Theory Methods* **13** 2395–2417.

DeMets, D. L. and Gail, M. H. (1985). Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics* **41** 1039–1044.

DeMets, D. L., Hardy, R., Friedman, L. M. and Lan, K. K. G. (1984). Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Controlled Clinical Trials* **5** 362–372.

DeMets, D., Williams, G., Brown, B. W. and NOTT Research Group. (1982). A case report of data monitoring experience: The Nocturnal Oxygen Therapy Trial. *Controlled Clinical Trials* **3** 113–124.

Ellenberg, S. S., Myers, M. W. and Hoth, D. F. (1993). The use of external monitoring committees in clinical trials of the National Institute of Allergy and Infectious Diseases. *Statistics in Medicine*. To appear.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77** 875–892.

Fleming, T. R. (1992). Data Monitoring Committees and capturing relevant information of high quality. *Statistics in Medicine*. To appear.

Fleming, T. R. and DeMets, D. L. (1992). Monitoring clinical trials: Issues and recommendations. *Journal of Controlled Clinical Trials*. To appear.

Fleming, T. R., Harrington, D. P. and O'Brien, P. C. (1984). Designs for group sequential tests. *Journal of Controlled Clinical Trials* **5** 348–361.

Freedman, L. S., Graubard, B. I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11** 167–178.

Green, S. and Crowley, J. (1993). Data monitoring committees for Southwest Oncology Group clinical trials. *Statistics in Medicine*. To appear.

Green, S. J. and Fleming, T. R. (1988). Guidelines for the reporting of clinical trials. *Seminars in Oncology* **15** 455–461.

Hallstrom, A. (1992). Anti-arrhythmic drugs: Before and after the CAST. *Washington Public Health* **10** 48–50.

Harrington, D. P., Fleming, T. R. and Green, S. (1982). Procedures for serial testing in censored survival data. In *Survival Analysis* (J. Crowley and R. A. Johnson, eds.) 269–280. IMS, Hayward, Calif.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* **44** 793–797.

Hsieh, F. Y., Crowley, J. and Tormey, D. C. (1983). Some test statistics for use in multistate survival analysis. *Biometrika* **70** 111–119.

Jacobson, M. A., Besch, C. L., Child, C. C., Hafner, R., Matts, J. P., Muth, K., Wentworth, D. N., Neaton, J., Deyton, L. and CPCRA. (1992). Prophylaxis with pyrimethamine for toxoplasmic encephalitis in patients with advanced HIV disease: Results of a randomized trial. Unpublished manuscript.

Jennison, C. and Turnbull, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **25** 49–58.

Kim, K. and DeMets, D. L. (1987). Estimation following group sequential tests in clinical trials. *Biometrics* **43** 857–864.

Louis, T. A. (1982). Finding the observed information using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 226–233.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40** 1079–1087.

Packer, M., Carver, J. R., Rodeheffer, R. J., Ivanhoe, R. J., DiBianco, R., Zeldis, S. M., Hendrix, G. H., Bommer, W. J., Elkayam, U., Kukin, M. L., Mallis, G. I., Sollano, J. A., Shannon, J., Tandon, P. K. and DeMets, D. L., for the PROMISE Study Research Group. (1991). Effect of oral milrinone on mortality in severe chronic heart failure. *New England Journal of Medicine* **325** 1468–1475.

Peace, K. E. (1990). *Statistical Issues in Drug Research and Development.* Dekker, New York.

Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43** 487–498.

Second International Study of Infarct Survival (ISIS-2) Collaborative Group. (1988). Randomized trial of intravenous streptokinase, oral aspirin, both or neither among 17,187 cases of suspect acute myocardial infarction. *Lancet* **13** 349–360.

Simon, R. (1990). Commentary on "Interim Analyses in Clinical Trials" by S. S. Emerson and T. R. Fleming. *Journal of Oncology* **4** 134–136.

Tang, D. I., Gnecco, C. and Geller, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *J. Amer. Statist. Assoc.* **84** 776–779.

TIMI Study Group. (1985). The Thrombolysis in Myocardial Infarction (TIMI) Trial: Phase I findings. *New England Journal of Medicine* **312** 932–936.

Tsiatis, A. A., Rosner, G. L. and Mehta, G. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40** 797–803.

Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). *Statist. Sci.* **4** 298–340.

Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J. Amer. Statist. Assoc.* **79** 653–661.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73** 573–581.