LAN, K. K. G. and DEMETS, D. L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics* **45** 1017–1020.

LAURIE, J. A., MOERTEL, C. G., FLEMING, T. R., WIEAND, H. S., LEIGH, J. E., RUBIN, J., MCCORMACK, G. W., GERSTNER, J. B., KROOK, J. E., MALLIARD, J., TWITO, D. I., MORTON, R. F., TSCHETTER, L. K. and BARLOW, J. F., FOR THE NORTH CENTRAL CANCER TREATMENT GROUP AND THE MAYO CLINIC (1989). Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil: The North Central Cancer Treatment Group and the Mayo Clinic. *Journal of Clinical Oncology* **7** 1447–1456.

LIN, D. Y. (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations. *Biometrika* **78** 123–131.

LIN, D. Y., FISCHL, M. A. and SCHOENFELD, D. A. (1992). Evaluating the role of CD4-lymphocyte change as a surrogate endpoint in HIV clinical trials. *Statistics in Medicine*. To appear.

MACHADO, S. G., GAIL, M. H. and ELLENBERG, S. S. (1990). On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection. *Journal of AIDS* **3** 1065–1073.

MARX, J. L. (1989). Drug availability is an issue for cancer patients, too. *Science* **245** 346–347.

MEDICAL RESEARCH COUNCIL WORKING PARTY (1984). The evaluation of low-dose preoperative x-ray therapy in the management of operable rectal cancer: Results of a randomly controlled trial. *British Journal of Surgery* **71** 21–25.

MOERTEL, C. G., FLEMING, T. R., MACDONALD, J. S., HALLER, D. G., LAURIE, J. A., GOODMAN, P. J., UNGERLEIDER, J. S., EMERSON, W. A., TORMEY, D. C., GLICK, J. H., VEEDER, M. H. and MAILLIARD, J. A. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine* **322** 352–358.

O'BRIEN, P. C. and FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

PEPE, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79** 355–365.

PEPE, M. S. and FLEMING, T. R. (1991). A non-parametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* **86** 108–113.

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* **8** 431–440.

RIDER, W. D., PALMER, J. A., MAHONEY, L. J. and ROBERTSON C. T. (1977). Preoperative irradiation in operable cancer of the rectum: Report of the Toronto Trial. *Canadian Journal of Surgery* **20** 335–338.

VOLBERDING, P. A., LAGAKOS, S. W., KOCH, M. A., PETTINELLI, C., MYERS, M. W., BOOTH, D. K., BALFOUR, H. H., REICHMAN, R. C., BARTLETT, J. A., HIRSCH, M. S., MURPHY, R. L., HARDY, D., SOEIRO, R., FISCHL, M. A., BARTLETT, J. G., MERIGAN, T. C., HYSLOP, N. E., RICHMAN, D. D., VALENTINE, F. T., COREY, L. and THE AIDS CLINICAL TRIALS GROUP OF THE NATIONAL INSTITUTE OF ALLERGY AND INFECTIOUS DISEASES (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine* **322** 941–949.

WHITEHEAD, J. (1986). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics* **42** 461–471.

WITTES, J., LAKATOS, E. and PROBSTFIELD, J. (1989). Surrogate endpoints in clinical trials: Cardiovascular diseases. *Statistics in Medicine* **8** 415–425.

Comment

John Crowley and Stephanie Green

Dr. Fleming has been instrumental in implementing monitoring committees and stopping guidelines for randomized clinical trials in both cancer and AIDS. Through his research, educational activities and service on Government committees, he serves as a model of statistician involvement in important clinical research. We whole-heartedly agree with the general principles Tom has discussed in this article. We welcome the opportunity to expand on some of the specific issues he raises.

John Crowley and Stephanie Green are Members, Program in Biostatistics, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, MP-557, Seattle, Washington 98104-2092.

DATA MONITORING COMMITTEES

Structure

The model of committees composed of independent investigators meeting every 6 months with open hearings beforehand is not practical in every setting, nor is it necessarily desirable. Funds are not available for committees of this sort for the 150 or so randomized trials being conducted in the cancer cooperative groups. Further, we believe that those who know the most about the trial are among those in the best position to judge it. In particular, it seems important to include some members who treat patients with the regimens being studied (and who thus face the ethical issues directly), as well as those who are most familiar with any problems with the data. Tom and we were involved in the development of the Southwest Oncology Group monitoring committee policy in 1985. Since then, the group has had good results using monitoring commit-

tees consisting of the clinician(s) in charge of the trial, the responsible statistician, the clinician in charge of the disease area (e.g., lung cancer) in the group, the group chair, the group statistician, a physician in the group who is not involved with the trial and a representative of the National Cancer Institute (the funding agency) (Green and Crowley, 1993). We try to ensure that committee members are free of any financial conflict of interest, but we recognize that other kinds of conflicts are inevitable if the committee is to be knowledgeable. In fact, our experience is that the least useful member of the committee is the uninvolved physician, who is chosen to be disinterested but too often is simply uninterested.

Conduct

Although we agree that leaks from data monitoring committees are dangerous, we would like to note that we think it is reasonable to use interim results for planning new studies. Using current information to arrive at the best guess as to the best treatment in order to include this in a new design will often be desirable. We would also like to note that ethical people may disagree. A monitoring committee member who strongly disagreed with the rest of the committee would be in a difficult position and might conclude his/her ethical obligation was to disclose results.

EARLY STOPPING/REPORTING GUIDELINES

Use of appropriate early stopping rules has been an important development in cancer clinical trials over the past 10 years. The O'Brien-Fleming boundaries are a common choice. We find that a conservative, flat boundary early on, with final analysis done at the 0.04–0.045 level, is also a useful option, particularly when several analyses are planned (Haybittle, 1971). By comparison, the O'Brien-Fleming boundary is more conservative at early analyses, but more liberal at later ones. In light of all the approximations that are used to arrive at boundaries (including the assumption that the total information at the end of the trial is known in advance) and of the observation that test characteristics are not sensitive to exact timing of interim analyses (DeMets and Gail, 1985), we generally specify interim tests rounded to sensible levels (e.g., 0.005, 0.01) and do not adjust if the tests are not done at exactly the right number of deaths. We also note that it is possible with the spending function approach to run out of error before the trial is complete if accrual is much slower than anticipated and that the approach is subject to bias. For instance, additional tests should not be done on the basis of results, but this can easily happen inadvertently (e.g., if an additional test is done due to slow accrual, but accrual is slow because clinicians have decided based on their clinical observations which treatment is best). We agree that follow-up after results are first published is important and would suggest that in most cases following indefinitely is reasonable since treatment may change the natural history of the disease.

EQUIVALENCE

We also have had difficulty in explaining that "nonsignificant" is not the same as equivalent. Our approach to designing equivalence trials when the new agent being tested is less toxic (or otherwise less objectionable) has been to set up a null hypothesis of a moderate improvement due to standard treatment, with the alternative being less than a moderate improvement (Harrington, Fleming and Green, 1982). Sample size is based on having adequate power to reject the null when the survival distributions are the same. "Moderate" varies from trial to trial. We use standard interim testing guidelines for early stopping, which is equivalent to the confidence interval approach Tom used in the PCP trial, but differs from the Jennison and Turnbull approach.

MULTIPLE ENDPOINTS

We appreciate the discussion of surrogacy and agree the concept generally is not useful in practice. In the cancer field, tumor shrinkage has been used for over 30 years as a short-term endpoint to screen active agents from inactive ones, but few would argue that tumor shrinkage is in any real sense a surrogate for patient survival. We are also not optimistic that auxiliary variables will prove any more useful. For instance, in Hsieh, Crowley and Tormey (1983), it is shown that very little is gained by using information on tumor shrinkage in analyzing survival unless this auxiliary variable is very highly correlated with survival and the relationship between the two can be modeled correctly. We are also not enthusiastic about the prospects for smoothing in this context. We share Tom's interest in the development of methods to incorporate multiple endpoints of clinical importance, but add a caution that any system of weights suffers from questions of interpretability. From our perspective, an unfortunate result of the pressure to use surrogates and auxiliary variables in AIDS is pressure to approve cancer drugs based on uncontrolled trials using tumor shrinkage as an endpoint, despite the well-known lack of association between tumor shrinkage and either survival or quality of life.

IMPORTANCE OF THE ROLE OF STATISTICIANS

We agree that we have an important role to play and that statistical scientists are often underrepresented on key advisory committees. But it is possible that too

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.