Statistical Science
1992, Vol. 7, No. 4, 428–456

Evaluating Therapeutic Interventions: Some Issues and Experiences

Thomas R. Fleming

Abstract. In frequently occurring life-threatening diseases such as cancer, AIDS and cardiovascular disease, there is a need of significant public health importance for rapid yet reliable evaluation of promising new therapeutic interventions that might provide greater efficacy and reduced toxicity. Leadership from statistical scientists is essential to effectively address many of the challenges resulting from this need. By discussing recent experiences, primarily in the area of oncology and AIDS clinical trials, we will illustrate several of these challenges. We also will review some designs and methods that have been implemented in these settings. Particular attention will be given to experiences from involvement with FDA Advisory Committees and with Data Monitoring Committees for clinical trials sponsored by industry or by the National Institutes of Health. Among issues to be discussed will be the role of independent monitoring committees and group sequential guidelines in randomized clinical trials, the evaluation of equivalence trials and the use of surrogate and auxiliary endpoints.

Key words and phrases: Group sequential designs, interim analyses, data monitoring committees, administrative analyses, repeated confidence intervals, equivalence trials, active control designs, auxiliary endpoints, surrogate markers, estimated likelihood, augmented score, estimating equations.

INTRODUCTION

Worldwide, the demand for health care resources continues to grow at an alarming rate. In the United States alone, according to fiscal year 1992 U.S. budget projections, an estimated $830 billion will be spent on health care. In frequently occurring life-threatening diseases such as cancer, AIDS and cardiovascular disease, there is a need of significant public health importance for more effective therapeutic interventions.

Strong leadership from statistical scientists is essential to develop and to guide the application of methods that allow rapid, yet reliable, evaluation of promising new treatments. These methods should minimize the use of limited patient and health care resources as well as the length of time required to definitively establish the efficacy of a new intervention. However, in our haste for answers, we should keep in mind that there are substantial negative consequences if we compromise the reliability of conclusions through a less rigorous scientific approach. False positive evaluations of new therapies resulting from an inappropriately streamlined scientific process could provide toxic and ineffective treatments to large numbers of patients, and could lead to unnecessary and significant additional demands on a health care budget that already is on the verge of exceeding available resources.

We will consider some of the important issues in evaluating therapeutic interventions, and discuss recent experiences to illustrate approaches that have been taken and some areas of future research. Our primary focus will be on some controversial issues arising in the use of surrogate marker data, in particular in AIDS clinical trials, and on issues arising in monitoring ongoing clinical trials, such as the need for independent Data Monitoring Committees and for protocol-specified group sequential guidelines. We will also discuss active control designs and the analysis of multiple measures of treatment effect.

Thomas R. Fleming is Professor of Biostatistics and Statistics, University of Washington, and Member, Fred Hutchinson Cancer Research Center, Seattle, Washington 98195.

MONITORING CLINICAL TRIALS

Motivation for Proper Monitoring Procedures

In randomized trials designed to provide definitive assessments of the effects of therapeutic interventions,

periodic interim analyses of accumulating data enable investigators to discontinue a treatment as soon as its efficacy or toxicity has been established to be unacceptable. This not only satisfies ethical requirements, but also allows more efficient use of limited research resources. However, if inappropriate procedures are used in this process of monitoring interim data, the integrity and credibility of the trial can be compromised. To illustrate some of the problems that can arise from an unstructured approach to performing interim monitoring of clinical trials, we begin with an example from the setting of treatment for rectal cancer. In a collection of patients about to receive surgical treatment for their disease, clinicians from Toronto's Princess Margaret Hospital had offered randomization to preoperative radiation treatment versus a control regimen involving surgery alone. After prematurely stopping the randomized trial, these investigators reported that early results from the study had shown "no difference between the two groups" in patient survival (Rider et al., 1977). In fact, the authors stated the trial of 125 patients was smaller than intended because interim results had been regularly available to all participating clinicians and because "the absence of any trend in survival during the early years caused the study to die a natural death." However, when exploratory analyses were conducted after trial termination and a favorable $p = 0.01$ treatment difference in survival was identified in the subgroup of 38 patients having Duke's Stage C cancer, they revised earlier conclusions to state the study "demonstrates that preoperative radiation treatment materially benefits the patient suffering from Duke's Stage C rectal cancer . . . thus there can be few arguments against its universal use." Fortunately, the Medical Research Council in the United Kingdom conducted a confirmatory trial having 552 patients randomized to either the Princess Margaret regimen for preoperative radiation therapy or to a "surgery only" control (Medical Research Council Working Party, 1984). The study revealed no evidence of any survival effect overall or in its Duke's C subgroup of 221 patients.

This example illustrates the value of confirmatory trials and the undesirable consequences of an unstructured approach to interim monitoring. Wide dissemination of results on relative efficacy of treatment regimens led to an inappropriate early loss of interest, and unstructured data exploration provided misleading conclusions. Several authors, including Armitage, McPherson and Lowe (1969), Fleming, Green and Harrington (1984) and Fleming and Watelet (1989), have demonstrated the substantial increase in the likelihood of obtaining false positive or false negative conclusions if one does not properly adjust for multiple testing, which occurs when interim monitoring is conducted in clinical trials. Pocock (1977) and O'Brien and Fleming (1979) proposed group sequential designs that allow early termination of a trial if initial results are extreme, while preserving the type I and type II error rates. Due to the complexity of randomized clinical trials, these procedures are intended to provide helpful guidelines, rather than rigid rules, about when early trial termination should occur.

To provide additional structure to preserve study integrity and credibility, randomized trials designed to definitively establish treatment efficacy and safety (i.e., Phase III trials) should have independent Data Monitoring Committees responsible for making recommendations about early termination. This is particularly important in the setting of diseases that are life threatening or produce irreversible morbidity. The Data Monitoring Committee (DMC) should have multidisciplinary representation, including statistical scientists, ethicists and physicians from relevant medical disciplines, to ensure that all relevant efficacy and safety issues are adequately addressed in any decisions concerning conduct of the trial, including early stopping. Members of the DMC ideally should also be the only individuals to whom the clinical trial's Data Analysis Center provides results on relative efficacy of treatments during the study, in order to minimize the likelihood that prejudgment of early results would compromise the ability of the trial to obtain definitive conclusions.

In the setting of oncology clinical trials, Green, Fleming and O'Fallon (1987) performed a matched analysis of larger randomized clinical trials from two major cancer cooperative groups, one that did and one that did not reveal interim results on efficacy only to members of a DMC. This matched analysis provided substantial evidence that DMCs do contribute positively to preserving the integrity of prospective clinical trials. The group without DMCs had 50% of studies showing declining patient accrual rates over time, had inappropriate early termination of studies yielding equivocal results and had completed studies with final results that were inconsistent with prematurely published early positive results. The studies in the group with DMCs were nearly or completely free of these problems.

Data Monitoring Committees

History in the United States. The Greenberg Report (1988), written in 1969 and published later in the *Journal of Controlled Clinical Trials*, established many of the guiding principles for the proper function of DMCs. These principles have been followed since the early 1970s by the major cardiovascular disease studies sponsored by the National Institutes of Health (NIH). In cancer cooperative groups sponsored by the National Cancer Institute (NCI), DMCs have been in place in all randomized comparative trials of the North Central Cancer Treatment Group since 1979, of the South-

west Oncology Group since 1984 and of most other major groups currently conducting studies.

When the AIDS Clinical Trial Group (ACTG) sponsored by the National Institute of Allergy and Infectious Diseases (NIAID) was established in 1987, a single DMC was formed to monitor all ACTG randomized trials. The DMC meets quarterly and has nine members, including three statisticians, two ethicists and four clinicians with infectious disease specialization. Membership formally excludes anyone from industry, the Food and Drug Administration (FDA) or the NIAID AIDS Program, as well as those directly involved in treating ACTG patients. However, to insure that important interactions can occur, Open Sessions are held during which industry/government sponsors, the FDA and study investigators can provide information to the DMC. Closed Sessions then occur at which data on the relative efficacy of treatments are openly discussed by committee members. This single committee now also monitors all randomized trials conducted by NIAID's second AIDS cooperative group, the recently formed Community Program for Clinical Research in AIDS (CPCRA).

Experiences. Experiences in monitoring trials in oncology and in AIDS illustrate the complexities of the decision making process. We will review the recent Cancer Intergroup Study 0035 evaluating the role of 5FU + levamisole as adjuvant therapy for resected colon cancer, and the recent placebo-controlled ACTG Study 019 investigating AZT in asymptomatic patients having the human immunodeficiency virus (HIV).

Each year in the United States alone, 100,000 new cases of colon cancer are diagnosed. In many of these cases, cancer has already spread into regional lymph nodes before surgery is performed, yet the disease is sufficiently confined that it can be clinically completely resected. In this setting, referred to as Duke's Stage C colon cancer, approximately 50% of patients will have recurrence of disease and death within 5 years due to undetected microscopic metastatic residual disease at the time of surgery. Cancer Intergroup 0035 involved randomization of 971 Duke's C patients, within 1 month of their clinically complete surgical resection, to adjuvant treatment with levamisole versus 5-FU + levamisole versus observation alone (Moertel et al., 1990). Final analysis was planned to be performed when 500 deaths had occurred, with up to three interim analyses planned after each group of 125 deaths. The primary intent of the trial was improvement in long-term survival, with a reduction in the rate of disease recurrence providing important supportive information. The O'Brien-Fleming group sequential procedure was used to guide decisions about early termination.

Patient accrual began in March 1984 and was completed October 1987, prior to the first interim analysis of efficacy results, which occurred in Spring 1988. At that analysis, there was quite strong evidence that the 5-FU + levamisole regimen was providing a reduction in the rate of recurrence of disease, confirming the result of an earlier smaller yet otherwise identically designed study (Laurie et al., 1989). Median follow-up for survival was a relatively short 18–24 months, with only small trends for survival improvement apparent at that time. Thus, the DMC decided the study should continue and remain blinded.

In late Summer 1988, some members of the DMC favored sharing current results with a small number of leaders from the FDA and NCI to facilitate the regulatory review process should the trial be terminated after the second interim analysis. Even though these FDA and NCI individuals had promised to maintain confidentiality, on the day after the meeting with the committee, an NCI official publicly shared blinded relative efficacy results. Former NCI Director V. DeVita obtained study results through this improper public revelation and promptly published an editorial in *Science* (Marx, 1989), challenging the ethics of study continuation by the DMC. These events illustrate the harmful risks to study integrity resulting from even very limited early release of relative efficacy results.

At the second interim analysis held in Fall 1989, 301 deaths had occurred rather than the planned 250, due to an unexpected cluster of deaths in late summer and a short delay in scheduling the interim analysis due to logistics of assembling the DMC. As discussed by Lan and DeMets (1983, 1989), among others, group sequential designs are flexible and fully allow such changes in the timing of analyses. The O'Brien-Fleming guideline for trial termination when 60% of information (301 of 500 deaths) has been obtained is 0.01 and was satisfied by the 5-FU + levamisole versus observation comparison relative to the primary survival endpoint ($p = 0.006$) as well as the secondary rate of recurrence endpoint ($p = 0.0001$) (Figure 1). The DMC recommended formal termination of the trial and reporting of results. However, since median survival follow-up was still only 3.5 years, the committee also recommended that study investigators continue to follow patients an additional 2–3 years. This formal termination in Fall 1989 was important since it led to an immediate redesign (replacing the untreated control regimen by 5-FU + levamisole) to the Cancer Intergroup's "next generation" study, which had already begun patient accrual, and led to prompt FDA approval for marketing of levamisole in this indication.

The DMC met every 6 months during the 3.5-year period preceding the first formal interim analysis held in Spring 1988. At these meetings, safety issues were reviewed along with issues important to the quality and integrity of the trial. These included assessing frequency of violation of protocol-specified treatment

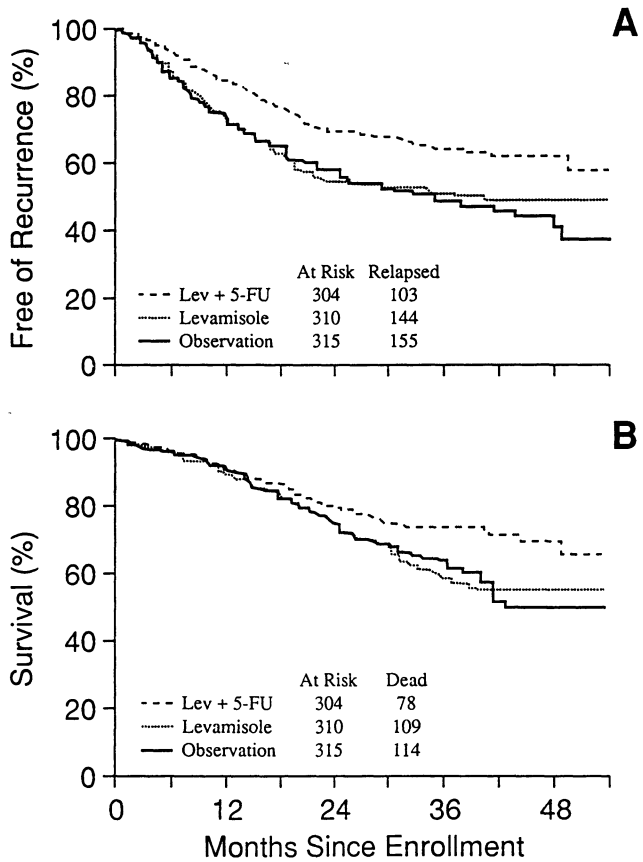

FIG. 1. *Recurrence and survival results, by treatment group, for Cancer Intergroup 0035 in Fall* 1989.

procedures, adherence to eligibility criteria, completeness of follow-up information, accrual rates, treatment balance with respect to key prognostic factors and pooled event rates. Frequent early inspection of these issues allowed prompt corrective measures to be taken with study investigators. Since results on relative efficacy of treatments were not released at these sessions, these analyses did not affect the false positive or false negative error rates of the trial and thus were not part of the formal group sequential design. They can be referred to as "administrative analyses," which formally are analyses conducted without access to data on the relative efficacy of treatments.

The ACTG Study 019, providing a placebo-controlled evaluation of AZT in asymptomatic HIV patients, was the largest and certainly one of the most important randomized trials conducted in patients with HIV infection or AIDS (Volberding et al., 1990). Between July 1987 and March 1989, the ACTG accrued over 3,200 patients to the trial that was designed to have separate analyses in the baseline CD4 < 500 and CD4 ≥ 500 categories, with the latter group requiring substantially longer follow-up to assess treatment effects. We will restrict our discussion to results from the 1,338 patients in the baseline CD4 < 500 category. The patients were randomized to double-blinded administration of high-dose AZT (1,500 mg/day), low-dose AZT (500 mg/day) or placebo. The study's primary objective was to obtain a definitive assessment of the effect of AZT on the clinical outcome: time to advanced AIDS-related complex (ARC), AIDS or death.

The interim analysis at the midpoint of the study in the CD4 < 500 category occurred in August 1989. The key efficacy data presented to the DMC at its meeting on August 2 are shown in Table 1A. At that time, 51 events were reported to the committee. Relative to placebo, the rate of outcome events was substantially reduced in both AZT groups, with the placebo comparison to low-dose AZT ($p = 0.0008$) meeting the O'Brien-Fleming stopping guideline of 0.005.

In order to provide an adequate opportunity to implement quality control procedures to insure completeness and accuracy of reported data, the ACTG study team "froze" the database on May 10, 1989. Thus, at the August 2 meeting, the data provided to the DMC were of high quality and included accurate follow-up information through May 10 on most patients, but did not include any information on outcome events occurring after May 10. Because the committee was seriously considering a recommendation for trial termination and because it expected many more events had occurred during the 3 months between early May and early August, it requested the 019 study team to rap-

TABLE 1
ACTG 019: AZT *in patients with asymptomatic* HIV: *interim results on clinical efficacy outcome*

Treatment	Number of events	Event rate*	P-value vs. placebo
A. 8/2/89 (data freeze on 5/10/89)			
Placebo (428)	31	7.5	–
500 mg (453)	8	2.1	0.0008
1,500 mg (457)	12	3.4	0.015
B. 8/16/89			
Placebo (428)	38 = 31 + 7	7.6	–
500 mg (453)	17 = 8 + 9	3.6	0.0030
1,500 mg (457)	19 = 12 + 7	4.2	0.05

* Failures per 100 person years of follow-up.

idly update only the information on the primary outcome. Through an intensive effort, the study team promptly provided the requested update that allowed the committee to meet again only 2 weeks later. Table 1B contains the updated results reviewed by the DMC on August 16 and shows 23 additional events had been documented. The total 74 events included progression to ARC in 19 patients (with three later having documented progression to AIDS) and progression to AIDS in 55 patients (with eight dying by the mid-August analysis). Even though the additional 23 events were nearly evenly distributed among the three treatment groups, the estimates of the AZT effect remained impressive, with the placebo versus low-dose AZT comparison still meeting the O'Brien-Fleming guideline. The DMC recommended termination of the placebo arm at this meeting on August 16.

The updated data do alter one's impression of the nature of the treatment effect estimated by the Kaplan-Meier time-to-event curves. For the placebo versus low-dose AZT comparison, the curves on August 2, in Figure 2A, reveal an estimated 96% versus 87% difference at 18 months, with a sustained reduction in the hazard rate over time. There is a suggestion that low-dose AZT would provide very substantial improvement in long-term survival. In contrast, the curves presented on August 16, in Figure 2B, reveal an estimated 94% versus 89% difference at 18 months, with the low-dose AZT curve essentially being a 6-month translation of the placebo curve. Several members of the DMC interpreted the update as providing evidence consistent with a delay, rather than cure, in AIDS progression, and the committee indicated there was no evidence as yet that early treatment with AZT would be more effective than delaying administration of the drug until CD4 lymphocyte counts fell to the 200 range.

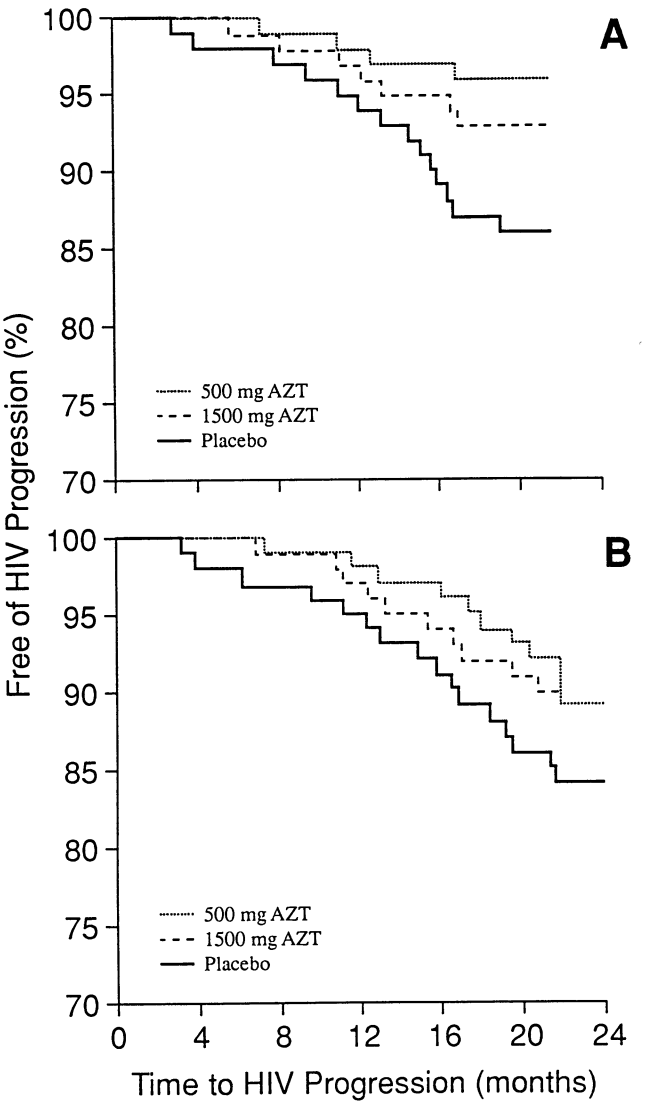

FIG. 2. HIV *progression results, by treatment group for* ACTG 019, *on August* 2, 1989 (*A*) *and on August* 16, 1989 (*B*).

Monitoring Clinical Trials: Some Issues

It is apparent that there are several issues needing greater attention in areas of development and implementation of statistical methods for monitoring trials.

Large rapidly accruing trials requiring long-term follow-up for occurrence of clinical endpoints may well provide definitive evidence of short-term beneficial effects of treatment, even though chronologically one may still be quite early in the planned duration of the study. This issue arose in Cancer Intergroup 0035 and to an even greater extent in ACTG 019, when nearly 90% of patients remained free of clinical events at the time of trial termination. Continued follow-up to obtain longer term results is occurring in 0035, with current results reported at the 1992 ASCO Plenary Session revealing substantial treatment effects on survival after median follow-up of 6 years. However, most patients on the placebo arm in ACTG 019 began taking AZT immediately after August 16, 1989, which precluded obtaining a long-term controlled evaluation of AZT in these asymptomatic patients. New designs and methods that enable one to obtain greater insights into long-term treatment effects in such settings would be very useful. Although not providing the same strength of evidence that is obtained from a long-term randomized trial, these useful methods might be formulated by an appropriate approach to pooling early results from the trial together with longer term results (on control treatment) from historical databases and (on experimental treatment) from data generated during a period of open label treatment that often follows early termination of a positive trial.

Data management procedures should be in place to allow high-quality data to be available to the DMC. However, as illustrated by ACTG 019, it is important

that follow-up data on outcome measures of relative efficacy be nearly current. A guideline for most settings is that follow-up data should be nearly uniformly complete to a date of "data freeze," which is no longer than 2 months prior to the date of the committee's meeting.

Extensive experiences in monitoring trials strongly reinforce the conclusion reached by the scientific review of Green, Fleming and O'Fallon (1987) and by the negative consequences of early data release in Cancer Intergroup 0035. Specifically, members of the DMC should be the only individuals to whom the clinical trial's Data Analysis Center provides interim results on relative effects of treatment interventions. All individuals with access to these interim results must commit to maintain confidentiality. In addition, members of the committee should be free of apparent significant scientific or financial conflict of interest.

Data Monitoring Committees are now routinely established in most government-sponsored Phase III randomized trials in the setting of life-threatening diseases. Due to the even greater potential for these committees to substantially increase the integrity and credibility of studies in industry, broad efforts are needed to increase their use in the industry-sponsored setting. Membership should be independent of industry, although the Open Sessions alluded to earlier can provide a forum for industry to provide information to the committee. When the Data Management and Data Analysis Center are not independent of the sponsoring company, it is preferable that the relative efficacy data be available only to the few individuals from the Data Analysis Center who are responsible for presenting results to the committee. In industry-sponsored studies, company officials should not be unblinded to efficacy results without authorization by the DMC, and any consideration by the company to override a DMC recommendation about trial continuation or termination should involve full consultation jointly with the DMC and the relevant regulatory agency, such as the FDA.

Important monitoring issues also arise in trials having active control designs. These will be discussed in the next section.

ACTIVE CONTROL DESIGNS

When effective standard treatment (STD) exists, it is frequent that one wishes to evaluate an analogue or alternative therapy that promises to have fewer side effects, be less costly or be easier to administer. Rather than proving superiority, one need only establish in such settings that the efficacy of the experimental treatment (EXP) is equivalent to that of STD. Some recent illustrations include evaluation of mitoxantrone (MITX) as an alternative to adriamycin (ADR) in treatment of advanced breast cancer, idarubicin (IDR) to replace daunorubicin (DNR) in first-line combination treatment (with ARA-C) for acute nonlymphocytic lymphoma (ANLL) and trimetrexate with leucovorin rescue (TMTX) to replace bactrim (BAC) in treatment of *Pneumocystis carinii* pneumonia in AIDS patients.

A common misconception among clinical investigators is that a nonsignificant test for equality establishes equivalence. For example, at the March 1986 FDA Oncology Advisory Committee meeting to review MITX in advanced breast cancer, the sponsor provided four studies that revealed the anti-tumor response rate on MITX was only approximately two-thirds that on ADR and survival was about 80% as long. Because the efficacy differences did not reach statistical significance in individual studies, the sponsor reached the misleading conclusion that equivalence had been established.

P-values from tests for equality, of course, fail to indicate how large a difference in treatment effects could still exist with reasonable likelihood. To establish that the EXP is equivalent to the STD, one must obtain definitive evidence against any hypothesized difference in favor of STD that would be judged to be clinically meaningful. This approach is explored in detail in Fleming (1987, 1990) and is formulated in terms of confidence intervals for the efficacy of EXP relative to that of STD. In each of the three illustrations above, it was judged that equivalence in patient survival would be established if one could rule out that survival on STD would be at least 25% longer than that on EXP; statistically, the lower limit of a 95% confidence interval of the STD/EXP hazard ratio in the relevant Cox (1972) proportional hazards regression model should exceed 0.8. It is immediately clear that the advanced breast cancer studies do not establish equivalence, since 0.8 was nearly the point estimate of the ADR/MITX hazard ratio.

Once analyses of active control trials have been formulated in terms of confidence intervals, one can then apply repeated confidence interval methods to guide early termination decisions. See Fleming (1987, 1990) and Emerson and Fleming (1989), and related work by Jennison and Turnbull (1984, 1989). Figure 3 illustrates this approach using the three ANLL studies designed to compare survival and complete response rates, that is, rates of complete eradication of clinically detectable disease, on IDR + ARA-C versus DNR + ARA-C. The final data, as presented at the July 1990 meeting of the FDA Oncology Advisory Committee, are from Southeast Group (SEG) and Adria Labs (ADRIA) trials, which were not terminated early, and from the Memorial Sloan Kettering Cancer Center (MSKCC) trial, which was stopped at the third of four planned analyses using an O'Brien-Fleming guideline. Since significance levels in a four-stage O'Brien-Fleming guideline are given by $\alpha < 0.001$, $\alpha < 0.004$, $\alpha < 0.018$ and

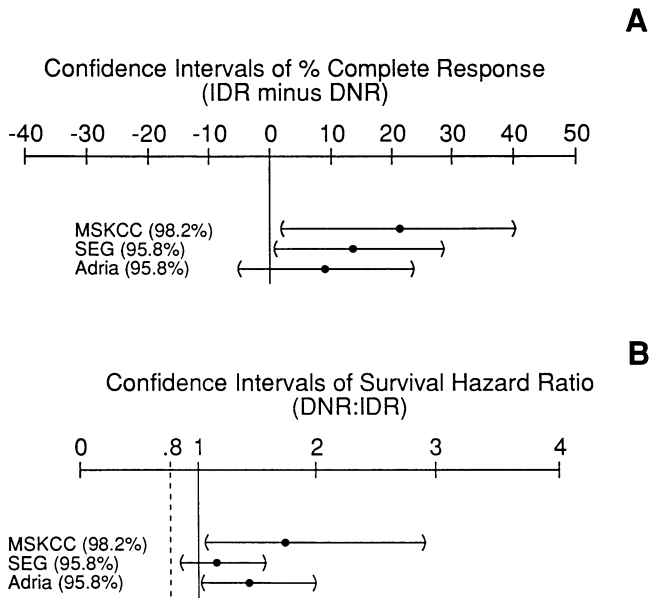

FIG. 3. *Confidence intervals for differences in % complete response and for the survival hazard ratio, for studies comparing* IDR + ARA-C *versus* DNR + ARA-C *in acute nonlymphocytic leukemia. Point estimates are represented by the solid dots.*

$a < 0.042$, the criterion for equivalence is based on the lower limit of the 98.2% confidence interval from the MSKCC study and 95.8% confidence intervals from the SEG and ADRIA trials. Figure 3 reveals that all three studies have hazard ratio lower limits above 0.8, establishing equivalence for survival, with the MSKCC and ADRIA trials also satisfying the stronger criterion for superiority by having lower limits above 1. All three also were judged to establish equivalence for rates of complete response, since lower limits of confidence intervals for the IDR-DNR difference in response rates exceeded −10%.

It is apparent that, unless a study has sufficiently large sample sizes to yield narrow confidence bands, one needs at least slightly favorable point estimates of the efficacy of EXP relative to STD in order for the trial to positively establish equivalence. Point estimates need to be quite favorable in order for equivalence to be definitively established at an interim analysis.

The active control trial comparing TMTX versus BAC in treatment of *Pneumocystis carinii* pneumonia (PCP) in AIDS patients illustrates the use of repeated confidence intervals to guide an early termination decision of a study having negative results. At the third analysis of a five-stage O'Brien-Fleming design, the estimated survival on TMTX was inferior to that on the STD, bactrim. Specifically, the estimate of the BAC/TMTX hazard ratio was 0.57. Noting that the O'Brien-Fleming guideline for the third look of five is 0.01, we computed the 99% confidence interval for the BAC/TMTX hazard ratio, which yielded (0.27, 1.19). The trial was stopped since these early results were sufficiently negative to rule out that the "efficacy-to-toxicity" profile of TMTX could be meaningfully more favorable than that of BAC.

Monitoring Multiple Measures of Treatment Effect: Some Issues

Even though formal guidelines for early stopping of clinical trials usually focus on a single primary outcome measure, often there is substantial interest at the time of data analysis in evaluating treatment effects on some secondary measures. For example, in the MSKCC trial illustrated in Figure 3, the study protocol indicated that the O'Brien-Fleming group sequential procedure would be applied to complete response data to guide decisions about early termination. The FDA was especially interested in the effect of treatment on patient survival and was interested in what adjustment would be required to account for the group sequential data evaluation. It is intuitively clear that the greater the correlation between the statistics assessing treatment effect on a primary measure and its effect on a secondary measure, the larger the adjustment that is necessary for the secondary outcome.

Rigorous solutions to this problem of determining proper adjustment to secondary outcomes after group sequential monitoring of a primary outcome would prove useful. Whitehead (1986) desribes analysis conditioning on the primary outcome. Unconditional analyses would be desirable that do not provide the same risk of adjusting away that part of the treatment effect related to its mechanisms of action that influence the primary outcome.

In a related issue, Lin (1991) has proposed a method for applying group sequential guidelines to multivariate outcomes. This method, which weighs outcome measures proportionally to the frequency of occurrence of each type of endpoint, should be quite efficient. Nevertheless, useful extensions of Lin's approach would be obtained by allowing investigators to prespecify bounds on how outcome measures might be weighed. This would be particularly important, for instance, when the clinically more important outcome might be late in its occurrence, such as patient survival, whereas another outcome might be an early occurring surrogate marker.

SURROGATE MARKERS AND AUXILIARY INFORMATION

Overview

In designing clinical trials to evaluate new interventions, one often must address some difficult and controversial issues when selecting proper measures of treatment effect. These "measures" or "endpoints" should not only be sensitive to the effect of treatment,

but also should be clinically relevant. In smaller screening studies that occur in the earlier stages of clinical experimentation, one usually focuses on measures of *biological activity*. Examples include evidence of tumor shrinkage in oncology, increase in ejection fraction or lowering of blood cholesterol in patients recovering from a myocardial infarction or measures of immune function (such as CD4 lymphocyte counts) or of viral load (such as P-24 antigen levels) in HIV-positive patients. Effects of treatment on these biological measures usually can be established quite rapidly. In contrast, in larger trials intended to allow a definitive evaluation of the role of a new treatment in clinical practice, one should focus on measures that unequivocally reflect tangible benefit to the patient; we refer to these as measures of *clinical efficacy*. These include length of survival and various measures of quality of life such as pain relief, cognitive ability, sense of well being, days spent in a hospital and ability to move about or to carry out normal activities. One could argue that there are additional measures of clinical efficacy. For example, in HIV-infected patients, these include occurrence of ARC symptoms such as wasting syndrome, thrush, hairy leukoplakia, persistent fever or fatigue or occurrence of AIDS-defining events such as opportunistic infections, Kaposi's sarcoma or AIDS-related dementia.

Obtaining high-quality data on each of these measures of clinical efficacy often is quite difficult. One may need several years of follow-up in order to address the effect of treatment on patient survival or on AIDS-defining events, and the interpretability of long-term endpoint data might be compromised by risks of loss-to-follow-up or of noncompliance. Measures of quality of life usually are very subjective, and often it is very difficult to obtain agreement on proper procedures to obtain reliable and meaningful quality of life information that can be collected uniformly across study centers. With a sense of urgency to identify effective therapies, especially for patients with life-threatening diseases such as cancer or AIDS, there is a strong reluctance to conduct randomized trials that are likely to require many years to definitively establish the effects of new interventions on measures of clinical efficacy. There is considerable interest in the use of surrogate or replacement endpoints in order to reduce the size, duration and cost of clinical trials. These surrogate endpoints usually are those measures of biological activity that can be evaluated after a short period of follow-up. For example, *Statistics in Medicine* papers by Ellenberg and Hamilton (1989), by Wittes, Lagakos and Probsfield (1989), and by Hillis and Siegel (1989) considered potential biological markers as surrogate endpoints in cancer treatment trials, cardiovascular disease prevention and treatment trials and ophthalmologic studies, respectively. Extensive recent

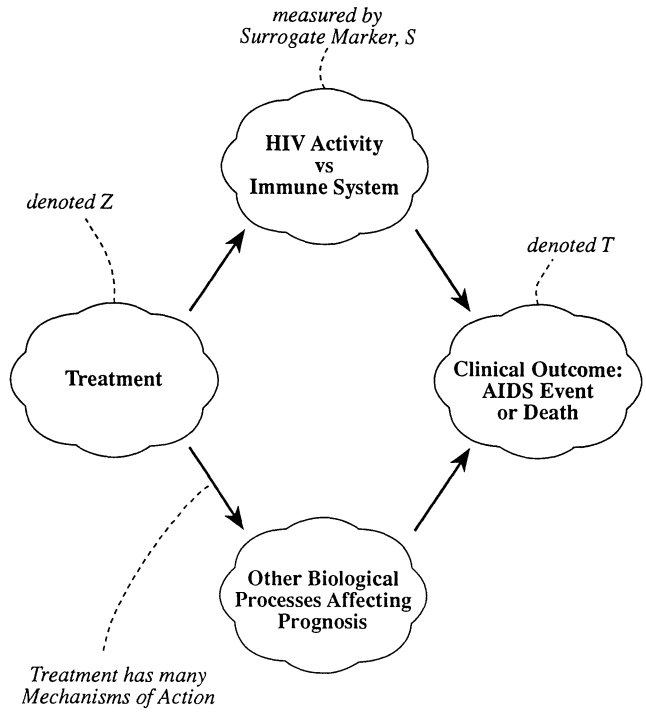

FIG. 4. *Relationships between treatment (Z), a surrogate marker (S), and the clinical outcome (T).*

discussion has focused on the use of CD4 lymphocyte counts and other biological markers as surrogates for progression to AIDS or death in studies of HIV-infected patients (e.g., IOM Conference Summary, 1990; Machado, Gail and Ellenberg, 1990; Jacobson, Bacchetti, Kolokathis et al., 1991; Ellenberg, 1991; Lagakos and Hoth, 1992).

Unfortunately, even though one might restrict attention to biological markers that are known predictors of the clinical outcome in natural history data, one can obtain highly misleading false positive or false negative conclusions when using treatment effects on biological markers to assess effects on longer term clinical outcomes.

Figure 4 provides insight into this important issue. As an illustration, consider the setting of HIV-positive patients. There is a principle underlying biological process, termed "HIV activity versus the immune system," that is of focal interest because it has a direct clinical effect of leading to AIDS-defining events and an increased risk of death. The status of this underlying biological process might be imperfectly measured by a surrogate marker, such as the patient's CD4 lymphocyte count. Treatments often are chosen based on their anticipated or documented effect on the surrogate marker, the CD4 lymphocyte count, with the expectation that such a treatment would provide a beneficial effect on the clinical outcome that would be mediated through its anticipated effect on the principal underlying biological process. This expectation may not be

fulfilled, even when natural history data clearly establish that the surrogate marker is strongly predictive of the risk of occurrence of the clinical outcome. Specifically, since there exist many other biological processes that also affect prognosis and since most treatments have many mechanisms of action, the actual treatment effect on the clinical outcome might be substantially altered by the treatment's effect on these other biological processes.

Suppose one does use treatment effects on biological markers to assess effects on longer term clinical outcomes. If the outcome effects mediated through other biological processes are quite negative, then beneficial effects of treatment on the surrogate marker could lead to false positive conclusions. Conversely, suppose treatment does not provide beneficial effects on the surrogate marker. False negative conclusions would occur if treatment has a beneficial effect on clinical outcome that either is mediated through beneficial effects on the principle underlying biological process that are not captured by the imperfect surrogate marker variable or is mediated through beneficial effects on the other biological processes affecting prognosis.

Misleading Use of Surrogate Markers: Some Illustrations

Many examples could be provided to illustrate the risks involved in the reliance on surrogate markers when attempting to evaluate clinical efficacy of treatment interventions. We will focus on two recent studies, in patients with ventricular arrhythmias after myocardial infarction (Cardiac Arrhythmia Pilot Study, 1988; Cardiac Arrhythmia Suppression Trial, 1989) and in those with chronic granulomatous disease (International Chronic Granulomatous Disease Cooperative Study Group, 1991), that clearly illustrate how effects on markers can lead to either false positive or false negative conclusions.

In patients having had a recent myocardial infarction, it is known that ventricular arrhythmias are a risk factor for subsequent sudden death. As a result, the antiarrhythmic drugs encainide and flecainide had become widely accepted, with an estimated 500,000 new patients per year receiving the drugs in the U.S. alone. Indeed, many argued a placebo-controlled trial to establish their effect on survival would not be ethical. Nevertheless, the Cardiac Arrhythmia Suppression Trial (CAST) involving over 2,000 randomized patients was conducted and established the startling result that the drugs nearly tripled the death rate relative to placebo. Interestingly, in April 1991, the U.S. Congress chided the FDA for its false positive evaluation and premature release of the antiarrhythmic drugs, and yet they continue to pressure the agency to use surrogate marker data to hasten approval of AIDS drugs.

Reliance on surrogate markers can also lead to false negative conclusions about the effects of treatments on clinical efficacy endpoints. Chronic granulomatous disease (CGD) is a serious childhood disease involving rare disorders of the immune system. Phagocytes from CGD patients ingest microorganisms normally but fail to kill them due to an inability to generate a respiratory burst dependent on the production of superoxide and other toxic oxygen metabolites. This in turn leads to a significant risk of recurrent serious and sometimes life-threatening infections.

There was evidence establishing a role for gamma interferon as an important macrophage-activating factor that could restore superoxide anion production and bacterial killing by phagocytes in CGD patients. The International CGD Cooperative Study Group planned a double-blinded placebo-controlled clinical trial to evaluate the role of gamma interferon in this setting. It was estimated that study treatments would need to be continued for 1 year to determine their effect on the rate of serious infections. Because this involved providing three weekly placebo injections for 1 year to one-half of the children on the study, many argued that the study should be limited to a 1-month duration, which should be adequate to assess the effect of gamma interferon on the surrogate markers, superoxide production and bacterial killing. After considerable debate, the study team did decide that surrogate markers too frequently provide misleading information about the clinical effect of treatment. Hence, the 1-year study was conducted with an interim analysis, guided by the O'Brien-Fleming design, to be performed at 6 months.

Patients were accrued between October 1988 and March 1989. At the interim analysis performed in July 1989, gamma interferon was found to provide better than a threefold reduction in the rate of recurrent serious infections, with the significance of the association being sufficiently strong to meet the O'Brien-Fleming guideline for early termination of the trial. The study was stopped with open-label gamma interferon then being made available to all study patients. The FDA approval for use of gamma interferon in CGD patients rapidly followed. This study illustrates that clinical trials with group sequential guidelines and clinical efficacy endpoints enable reliable evaluation of treatment interventions and can lead to rapid availability to patients of those therapies that are truly effective. Interestingly, when the surrogate biological marker data were analyzed in the CGD study, there was no detectable effect of treatment on either measurements of superoxide production or bacterial killing. Reliance on surrogate markers can lead to false negative conclusions and in this setting would have deprived children with CGD of an effective intervention.

It is apparent that statistical conditions that rigorously establish when surrogate markers are valid

would be very useful in clinical applications. We will consider conditions formulated by Prentice (1989) and their use in AIDS data.

Prentice Criteria for Valid Surrogates

As stated by Prentice (1989), corresponding to any true clinical endpoint, T, the label *surrogate endpoint* should be reserved for those variables, S, "for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." In symbols, S must satisfy

$$(1) \qquad P(S|Z) = P(S) \Leftrightarrow P(T|Z) = P(T),$$

where Z denotes treatment. Prentice then identified two conditions that essentially are sufficient to establish the validity of (1):

(i) $P(T|S, Z) = P(T|S)$; that is, S fully captures the effect of Z on T; and
(ii) $P(T|S) \neq P(T)$; that is, S is informative about T.

To obtain ($\Rightarrow$) in (1), note

$$\begin{aligned} P(T|Z) &= \int P(T, S|Z)\, dS \\ &= \int P(T|S, Z)P(S|Z)\, dS \\ &= \int P(T|S)P(S)\, dS \\ &= \int P(T, S)\, dS \\ &= P(T), \end{aligned}$$

where the third equality follows by $P(S|Z) = P(S)$ and by (i). Conversely, note $P(T|Z) = P(T)$ implies

$$\int P(T|S, Z)P(S|Z)\, dS = \int P(T|S)P(S)\, dS$$

and so, by (i),

$$\int P(T|S)P(S|Z)\, dS = \int P(T|S)P(S)\, dS.$$

Thus, if we restrict attention to the important setting in which, for each t over the support of the distribution, $P(T \leq t|S)$ is strictly monotone in S and in which stochastic ordering holds in Z for the distribution $P(S|Z)$, then (ii) would establish ($\Leftarrow$).

Unfortunately, condition (i), requiring the surrogate endpoint to fully capture the effect of treatment on the true clinical endpoint, is so restrictive that it would rarely hold in clinical applications. For example, in evaluating the Burroughs-Wellcome (BW) 02 and the ACTG 016 controlled trials of AZT in HIV-infected individuals, Lin, Fischl and Schoenfeld (1992) found that the time-varying covariate representing CD4-lymphocyte change does not fulfill the Prentice criteria for being a valid surrogate for the clinical endpoints. Specifically, they found that the overall effect of AZT on AIDS-defining events and death exceeds the effect of AZT on these clinical outcomes, which is mediated through its effect on the surrogate marker. When data were analyzed from other controlled trials of AZT, similar conclusions were reached by Tsiatis and DeGruttola (personal communication) using ACTG 002 and BW 02, and by Choi and Lagakos (personal communication) using ACTG 019.

One rarely can establish that surrogate endpoints are valid. Even in that rare setting in which data on treatment Z would allow one to view S as a valid surrogate for T, one cannot extrapolate this surrogacy to any new treatment Z^* that could have mechanisms of action that differ from those of Z. As an illustration, gp160 and gp120 vaccines (Z^*) are now being evaluated as therapies for the early stages of HIV infection. These treatments provide novel mechanisms of action relative to currently used anti-retrovirals (Z) such as nucleoside analogues, AZT, ddI or ddC. Thus, the FDA Vaccine Advisory Committee, at its meeting on November 12, 1991, unanimously agreed that vaccine trials must allow direct evaluation of the effect of Z^* on clinical outcomes (T), whether or not CD4-lymphocyte changes (S) would be judged to provide a valid surrogate for T in the setting of the nucleoside analogues.

Whenever attempting to provide a definitive evaluation of the effect of treatment on a true clinical endpoint, one should be extremely cautious about the amount of emphasis placed on the association of treatment with surrogate markers. Because one rarely can establish that surrogate endpoints are valid, we pursue in the next section some approaches that use the information in S as an auxiliary variable to strengthen standard analyses of the association of treatment with the true endpoint, T.

Auxiliary Variables

Rather than serving as surrogates to replace true endpoints, response variables, such as measures of biological activity discussed earlier in this section, can be used to strengthen clinical efficacy analyses. These variables, S, then should be called auxiliary. Suppose one's interest is in the effect of treatment on time to a true clinical endpoint, T. Suppose further that the auxiliary information, S, is readily available, whereas T is censored in a substantial fraction of those patients having relatively late occurring clinical endpoints. If S and T are strongly correlated, one can expect that S will provide useful additional information about the clinical endpoint in those patients in which T is censored.

Three approaches that have been proposed for using auxiliary variables have been referred to as "variance reduction," "augmented score" and "estimated likelihood." The variance reduction approach, explored by Kosorok (1991), is applicable when S is a time-to-event

endpoint and when the treatment relationship with S yields a zero mean statistic X such that $\text{cor}(X,Y) \equiv \rho$ is positive, where Y is a standard statistic used to assess the effect of treatment on T. The statistic $Y-\rho X$ proposed by Kosorok makes use of auxiliary information to provide a "variance reduced" alternative to using Y.

The "augmented score" and "estimated likelihood" approaches, which we will review in somewhat greater detail, were explored by Fleming et al. (1992). For either approach, we assume we have n independent cases, where case i provides data (X_i, δ_i, Z_i). If T_i and U_i are independent latent failure and censoring variables for case i, then $X_i = \min\{T_i, U_i\} \equiv T_i \wedge U_i$, $\delta_i = I_{\{X_i = T_i\}}$ and the vector Z_i provides covariate information. Here, $I_{\{A\}}$ denotes an indicator for A. Assume the usual proportional hazards model

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta' Z) \tag{2}$$

for the relationship between the covariate vector Z and the hazard function for the clinical outcome T. Denote the cumulative hazard for λ_0 by

$$\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds.$$

To motivate the augmented score approach, recall that, when λ_0 is assumed to be known, a maximum likelihood estimate for β can be obtained by solving the score estimating equation

$$\sum_{i=1}^{n} Z_i M_i(X_i|\beta) = 0, \tag{3}$$

where, for any $t \geq 0$,

$$M_i(t|\beta) = I_{\{T_i \leq t\}} - e^{\beta' Z_i}\Lambda_0(t \wedge T_i), \tag{4}$$

with $M_i(t|\beta)$ being a case-specific martingale in t in uncensored data (i.e., $U_i \equiv \infty$), essentially representing "observed" minus "model predicted" events over $[0, t]$ for case i.

In turn, in the semi-parametric setting where λ_0 is unspecified, the Cox (1975) maximum partial likelihood estimate of β is obtained by solving the estimating equation

$$\sum_{i=1}^{n} Z_i \hat{M}_i(X_i|\beta) = 0, \tag{5}$$

where $\hat{M}_i$ is obtained from (4) by estimating $\Lambda_0(t)$ using the semi-parametric Breslow (1972) estimator evaluated at β,

$$\hat{\Lambda}_0(t) = \sum_{\{j: X_j \leq t, \delta_j = 1\}} \left\{ \sum_{k=1}^{n} I_{\{X_k \geq X_j\}} e^{\beta' Z_k} \right\}^{-1}.$$

Censorship reduces the information available in (3) or (5) that is used for the estimation of β. Specifically, $M_i(t|\beta)$ is only known over $t \in [0, X_i]$ rather than over $t \in [0, T_i]$, and, in (5), less information is available to formulate $\hat{\Lambda}_0$.

Fortunately, the surrogate information, S_i, does allow recovery of some of this lost information. Suppose τ denotes some arbitrary large time, and temporarily assume λ_0 is known. To recover some information over $(X_i, \tau]$ for a censored case (i.e., with $\delta_i = 0$), we consider

$$e_{M_i}(\beta) \equiv E\{M_i(\tau|\beta) - M_i(X_i|\beta)|X_i, \delta_i = 0, S_i\}, \tag{6}$$

which essentially is the conditional expectation of the lost information over $(X_i, \tau]$, given available information on case i to X_i. It is straightforward to show that the right-hand side of (6) reduces to zero when one conditions only on $(X_i, \delta_i = 0)$, verifying that the recovery of information on β is made possible only through using the available information on S_i.

The expectation $e_{M_i}(\beta)$ involves the unknown joint distributions for S_i with T_i and U_i. Fleming et al. (1992) formulate an estimator $\hat{e}_{M_i}(\beta)$ in the special case in which S_i is a censored time-to-event endpoint. Then, for the setting in which λ_0 is unspecified, they propose estimation of β based on solving the "augmented score equation:"

$$\sum_{i=1}^{n} Z_i \hat{M}_i(X_i|\beta) + \sum_{i=1}^{n} (1 - \delta_i) I_{\{X_i < \tau\}} Z_i \hat{e}_{\hat{M}_i}(\beta) = 0.$$

Exploration needs to be done to determine a proper choice of the arbitrary τ that would allow the capture of as much information as possible, while still allowing the estimates $\hat{e}_{\hat{M}_i}(\beta)$ to be stable.

The "estimated likelihood" approach grew out of earlier work by Pepe and Fleming (1991) and Carroll and Wand (1991). These authors independently studied the estimated likelihood for inference with mismeasured covariate data using models that are nonparametric with respect to the mismeasurement process. Pepe (1992) extended this work to the setting of missing outcome data, where she assumed one had a validation set of patients having complete uncensored information on both the biological marker, S, and clinical outcome, T, and a nonvalidation set of patients having only the marker data, S.

Fleming et al. (1992) extended Pepe's approach to be applicable in the usual setting where one has censored data on both T and S, rather than validation and nonvalidation sets. Continue to assume the usual proportional hazards model in (2), and temporarily assume λ_0 to be known. Following Pepe's (1992) semi-parametric approach, which involves nonparametric estimation of $P(S|T, Z)$ to obtain greater robustness, the corresponding estimated likelihood is

$$\hat{L}(\beta) = \prod_{\delta_i = 1} P_\beta(T_i|Z_i) \prod_{\delta_i = 0} P_\beta(T > X_i|Z_i) \cdot \prod_{\delta_i = 0} \hat{P}_\beta(S_i|T > X_i, Z_i), \tag{7}$$

where S_i can be an arbitrary right-censored vector-valued process providing auxiliary information. The first two terms on the right-hand side of (7) represent the usual likelihood when the auxiliary information, S, is not taken into account. Under (2), these two terms reduce to the usual Cox partial likelihood when λ_0 is considered to be unspecified and, in turn, is estimated by the piecewise linear approach presented in Breslow (1974).

Turning to the third term in the estimated likelihood in (7), we have

$$(8) \qquad \hat{P}_\beta(S_i|T>X_i,Z_i) = \int_{X_i}^{\infty} P_\beta(t|T>X_i,Z_i)\hat{P}_\beta(S_i|t,Z_i)\,dt,$$

where $\hat{P}_\beta(S_i|t, Z_i)$ is a nonparametric estimate, as explored in Fleming et al. (1992). From (8), it is clear that the amount of improvement provided by the estimated likelihood relative to the usual partial likelihood depends on the degree of dependence of $P_\beta(S|t, Z_i)$ on t. In fact, with nonparametric estimation of $P_\beta(S|t, Z_i)$, the information about β reduces to that provided by the usual partial likelihood when

$$P(S|T,Z) = P(S|Z)$$

or, equivalently, when $P(T|S, Z) = P(T|Z)$. Thus, it is only necessary that S relate to the true endpoint T, given Z, in order that it serve as a potentially useful auxiliary variable.

This estimated likelihood approach to using auxiliary information is quite flexible since it allows S to be a multivariate censored stochastic process. For example, components could be a patient's evolving CD4 lymphocyte count or blood pressure over time. However, important work remains before this estimated likelihood would provide an efficient method to using auxiliary information to strengthen standard analyses of the association of treatment with the true clinical endpoint. For example, some improvements in the nonparametric estimates of the term $P_\beta(S_i|t, Z_i)$ in (8) should be possible by using smoothing techniques to increase available information and in turn to reduce variability. This smoothing could be done both relative to T (allowing information from neighboring failure times to be used) and relative to S (allowing matching to individuals with similar rather than identical auxiliary information). In addition, for each of the three approaches we have described for using auxiliary information, investigations should be performed to determine the degree of correlation between S and T and the nature of censoring required for these approaches to be useful.

Improvements in efficiency with these approaches using auxiliary information are likely to be small unless S and T are highly correlated and unless there is one pool of patients having longer term follow-up and another pool of patients with auxiliary information but with relatively short-term follow-up on the clinical endpoint. In spite of these limitations, approaches using auxiliary information are of interest since they avoid the substantial risks for false positive or false negative conclusions that arise when surrogate markers are used to replace measures of clinical efficacy.

DISCUSSION AND SUMMARY

There is an urgent need for rapid development and evaluation of promising new interventions, especially in the setting of life-threatening diseases such as cancer and AIDS. Meanwhile, spiraling costs for health care are limiting the access of an increasing number of patients to effective treatments that currently are available. Toxic and ineffective treatments that become widely available contribute to these spiraling costs without providing the desired therapeutic benefit. Strong leadership is needed from statistical scientists in the effort to confront these critical public health issues.

In clinical research, this leadership includes close collaboration with medical researchers, on an ongoing basis, to assure that the design, conduct, analysis and interpretation of results of clinical trials are properly performed, and it includes serving on Data Monitoring Committees for government- or industry-sponsored trials. In methodologic research, this includes the development and evaluation of new scientific approaches to medical research that will allow more rapid and efficient evaluation of treatments without compromising the reliability of conclusions. Finally, in public service, this includes playing an active role on Review Committees and Advisory Committees for NIH and for the FDA or other regulatory agencies that have considerable influence in defining the scientific standards for clinical research and in establishing what evidence should be required before releasing new drugs and biologics.

In this paper, we have discussed some of the designs, methods and important issues in evaluating therapeutic interventions and some areas of future research. Independent Data Monitoring Committees should be established in government- and industry-sponsored randomized trials designed to definitively establish treatment efficacy and safety, particularly in the setting of diseases that are life-threatening or produce irreversible morbidity. These committees should have multidisciplinary representation and membership free of apparent significant conflict of interest, and, ideally, their members should be the only individuals to whom the trial's Data Analysis Center provides access to interim results on relative efficacy of treatments. These committees should be guided by protocol-specified group sequential designs. Among future research topics, methods are needed that enable one to obtain greater insights into long-term treatment effects in trials that are reported early due to definitive evidence of short-term benefits.

Active control designs should be used when attempting to establish that a new therapeutic approach is less toxic or costly than standard treatment, yet equally efficacious. Analysis of active control trials should focus on whether confidence intervals allow one to exclude clinically meaningful differences that favor the standard treatment. An illustration is provided by the recent ACTG Study 021, in which bactrim, as secondary prophylaxis of *Pneumocystis carinii* pneumonia in AIDS patients, was found to have efficacy sufficiently favorable to satisfy these criteria relative to the much more expensive standard treatment, aerosolized pentamidine.

In the evaluation of a new intervention, the selection of the measures of treatment effect can be difficult and controversial. To be clinically relevant, trials intended to allow a definitive evaluation of the role of a new treatment in clinical practice should use measures of *clinical efficacy*. Surrogate or replacement endpoints, which usually are measures of *biological activity*, often are considered instead in an attempt to reduce the size, duration and cost of clinical trials. The use of surrogate endpoints currently is one of the most intensely debated issues in AIDS research. Because one rarely can establish the validity of a surrogate endpoint, an alternative approach could be considered in which measures of biological activity are used as auxiliary information to strengthen analyses of the effect of treatment on the clinical endpoint. Much important work remains in developing efficient approaches to using auxiliary information.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institute of Allergy and Infectious Diseases (5 R01 AI29168-03) and from the National Cancer Institute (2 R01 CA39929-08). Special acknowledgment for helpful discussion and comments should also be given to Ross Prentice, Margaret Pepe, Susan Ellenberg, David DeMets, James Zidek and the *Statistical Science* reviewers.

REFERENCES

ARMITAGE, P., MCPHERSON, C. K. and ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132** 235-244.

BRESLOW, N. E. (1972). Discussion of "Regression models and life-tables" by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 216-217.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89-99.

CARDIAC ARRHYTHMIA PILOT STUDY (CAPS) (1988). Effects of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: The CAPS. *American Journal of Cardiology* **61** 501-509.

CARDIAC ARRHYTHMIA SUPPRESSION TRIAL (CAST) (1989). Preliminary Report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine* **312** 406-412.

CARROLL, R. J. and WAND, M. P. (1991). Semi-parametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B* **53** 573-585.

COX, D. R. (1972). Regression models and life-tables [with discussion]. *J. Roy. Statist. Soc. Ser. B* **34** 187-220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269-276.

ELLENBERG, S. S. (1991). Surrogate endpoints in clinical trials (editorial). *British Medical Journal* **302** 63-64.

ELLENBERG, S. S. and HAMILTON, J. M. (1989). Surrogate endpoints in clinical trials: Cancer. *Statistics in Medicine* **8** 405-413.

EMERSON, S. S. and FLEMING, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45** 905-923.

FLEMING, T. R. (1987). Treatment evaluation in active control studies. *Cancer Treatment Reports* **17** 1061-1065.

FLEMING, T. R. (1990). Evaluation of active control trials in AIDS. *Journal of AIDS* **3** S82-S87.

FLEMING, T. R. and WATELET, L. F. (1989). Approaches to monitoring clinical trials. *Journal of the National Cancer Institute* **81** 188-193.

FLEMING, T. R., GREEN, S. J. and HARRINGTON, D. P. (1984). Considerations for monitoring and evaluating treatment effects in clinical trials. *Controlled Clinical Trials* **5** 55-66.

FLEMING, T. R., PRENTICE, R. L., PEPE, M. S. and GLIDDEN, D. (1992). Surrogate and auxiliary endpoints in clinical trials with potential application in cancer and AIDS research. Technical Report, Dept. Biostatistics, Univ. Washington.

GREEN, S. J., FLEMING, T. R. and O'FALLON, J. R. (1987). Policies for study monitoring and interim reporting of results. *Journal of Clinical Oncology* **5** 1477-1484.

GREENBERG REPORT (1988). Organization, review, and administration of cooperative studies. *Controlled Clinical Trials* **9** 137-148.

HILLIS, A. and SEIGEL, D. (1989). Surrogate endpoints in clinical trials: Ophthalmologic disorders. *Statistics in Medicine* **8** 427-430.

INTERNATIONAL CHRONIC GRANULOMATOUS DISEASE COOPERATIVE STUDY GROUP (1991). A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. *New England Journal of Medicine* **324** 509-516.

IOM CONFERENCE SUMMARY (1990). Surrogate endpoints in evaluating the effectiveness of drugs against HIV infection and AIDS.

JACOBSON, M. A., BACCHETTI, P., KOLOKATHIS, A., CHAISSON, R. E., SZABO, S., POLSKY, B., VALAINIS, G. T., MILDVAN, D., ABRAMS, D., WILBER, J., WINGER, E., SACKS, H. S., HENDRICKSEN, C. and MOSS, A. (1991). Surrogate markers for survival in patients with AIDS and AIDS-related complex treated with zidovudine. *British Medical Journal* **302** 73-78.

JENNISON, C. and TURNBULL, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5** 33-45.

JENNISON, C. and TURNBULL, B. W. (1989). Interim analyses: The repeated confidence interval approach. *J. Roy. Statist. Soc. Ser. B* **51** 305-361.

KOSOROK, M. R. (1991). A variance reduction method for combining multivariate failure times in order to increase power in clinical trials. Ph.D. dissertation, Univ. Washington.

LAGAKOS, S. W. and HOTH, D. F. (1992). Surrogate markers in AIDS: Where are we? *Annals of Internal Medicine* **116** 599-601.

LAN, K. K. G. and DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659-663.

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-