ELSEVIER

Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institutes of Health Workshop*

Victor G. De Gruttola, DSc, Pamela Clax, DPM, David L. DeMets, PhD, Gregory J. Downing, DO, PhD, Susan S. Ellenberg, PhD, Lawrence Friedman, PhD, Mitchell H. Gail, MD, Ross Prentice, PhD, Janet Wittes, PhD, and Scott L. Zeger, PhD

Harvard University School of Public Health, Boston, Massachusetts (V.G.D.G.); National Institute of Allergy and Infectious Diseases, Bethesda, Maryland (P.C.); University of Wisconsin-Madison, Madison, Wisconsin (D.L.D.); Office of the Director, National Institutes of Health, Bethesda, Maryland (G.J.D.); Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, Maryland (S.S.E.); National Heart, Lung, and Blood Institute, Bethesda, Maryland (L.F.); National Cancer Institute, Bethesda, Maryland (M.H.G.); Fred Hutchinson Cancer Research Center, Seattle, Washington (R.P.); Statistics Collaborative, Washington, DC (J.W.); and Johns Hopkins University School of Public Health, Baltimore, Maryland (S.L.Z.)

ABSTRACT: We report on recommendations from a National Institutes of Health Workshop on methods for evaluating the use of surrogate endpoints in clinical trials, which was attended by experts in biostatistics and clinical trials from a broad array of disease areas. Recent advances in biosciences and technology have increased the ability to understand, measure, and model biological mechanisms; appropriate application of these advances in clinical research settings requires collaboration of quantitative and laboratory scientists. Biomarkers, new examples of which arise rapidly from new technologies, are used frequently in such areas as early detection of disease and identification of patients most likely to benefit from new therapies. There is also scientific interest in exploring whether, and under what conditions, biomarkers may substitute for clinical endpoints of phase III trials, although workshop participants agreed that these considerations apply primarily to situations where trials using clinical endpoints are not feasible. Evaluating candidate biomarkers in the exploratory phases of drug development and investigating surrogate endpoints in confirmatory trials require the establishment of a statistical and in-

Address reprint requests to: Victor G. De Gruttola, D.Sc., Director, Adult Statistical and Data Analysis Center, Harvard School of Public Health, Bldg 2-439, 655 Huntington Avenue, Boston, MA 02115 (victor@sdac.harvard.edu).

Received October 16, 2000; accepted May 7, 2001.

**NIH Workshop: Research Needs for the Design and Analysis of Surrogate Endpoints in Clinical Trials, December 1–2, 1998 (<<http:\www.od.nih.gov/osp/ospp.biomarkers/coverpage.htm>>).*

Controlled Clinical Trials 22:485–502 (2001)

655 Avenue of the Americas, New York, NY 10010

0197-2456/01/$–see front matter
PII S0197-2456(01)00153-2

ferential framework. As a first step, participants reviewed methods for investigating the degree to which biomarkers can explain or predict the effect of treatments on clinical endpoints measured in clinical trials. They also suggested new approaches appropriate in settings where biomarkers reflect only indirectly the important processes on the causal path to clinical disease and where biomarker measurement errors are of concern. Participants emphasized the need for further research on development of such models, whether they are empirical in nature or attempt to describe mechanisms in mathematical terms. Of special interest were meta-analytic models for combining information from multiple studies involving interventions for the same condition. Recommendations also included considerations for design and conduct of trials and for assemblage of databases needed for such research. Finally, there was a strong recommendation for increased training of quantitative scientists in biologic research as well as in statistical methods and modeling to ensure that there will be an adequate workforce to meet future research needs. *Control Clin Trials* 2001; 22:485–502

KEY WORDS: *Surrogate endpoints, biomarkers, meta-analysis*

INTRODUCTION AND BACKGROUND

Design and Analysis of Clinical Trials amid Rapid Advances in Biotechnology and Genomics

Research in biosciences and technology is yielding promising new ways of understanding and measuring human disease processes. Genome sequencing, DNA microarrays, proteomics, and magnetic resonance imaging are giving rise to new tools of biostatistics and epidemiology that are making their way into clinical investigation and are producing vastly more information than was obtained through previous methods. This emerging field of bioinformatics contends with the explosion of data in molecular biology and genetics. Statisticians are just beginning to develop formal methods of estimation and hypothesis testing using DNA and protein sequence data. These new technologies and sources of information will enable investigators to pose scientific questions and approach problems in ways that, until recently, were barely conceivable.

The revolution in biotechnology is generating myriad biomarkers, some of which may serve as useful early indicators of either therapeutic benefit or harm. Medical scientists are interested in exploring the use of these markers in clinical practice. In the following discussions, we use the term "treatment" generically to apply to all interventions whether for prevention, diagnostic, or therapeutic purposes. Some fields already use biomarkers to identify subgroups of patients who respond to therapies in different ways. Biomarkers are also used to aid in early detection of disease and in the investigation of interventions aimed at reducing the risk of disease. In addition to these uses, there is scientific interest in exploring whether, and under what conditions, a biomarker may be used for screening candidate interventions in a phase II trial and substitute for a primary endpoint of phase III trials. Such substitution could allow more rapid and less costly evaluation of a new treatment.

Workshop participants emphasized, however, that many clinical trials have been, and can be, conducted successfully and efficiently using clinical outcomes. Such trials rely on biomarkers or surrogate endpoints only to help explain the mechanisms of action of the interventions and to provide better understanding of the biology of the conditions being studied. Because these tri-

als are larger and usually last longer than those using biomarker endpoints, they can directly address the issue of potential adverse effects of the interventions without assuming that biomarkers are adequate to assess adverse effects. This discussion applies primarily to situations where trials using clinical endpoints are not feasible or cannot be carried out efficiently. In addition to the extent that new biostatistical and mathematical approaches are developed in response to the needs for biomarkers and surrogate endpoints, studying biomarkers will advance the methodology of clinical trials.

Biomarkers and Surrogate Endpoints

Over the last 50 years, biostatistics has provided a framework for designing and analyzing clinical investigations to determine the clinical benefits of a treatment as well as to determine its effects on biomarkers of health or disease status. Since about 1989, biostatisticians have investigated approaches to evaluating whether a biological parameter might serve as a substitute or "surrogate" for a clinical endpoint in the study of a particular therapy for a particular disease. There has not been a consistent use of terminology in the scientific and medical literature describing the substitution of biological parameters for clinical endpoints. Recently, a National Institutes of Health (NIH) working group recommended preferred terms and definitions that have broad applications [1]:

Biological Marker (Biomarker): A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

Clinical Endpoint: A characteristic or variable that reflects how a patient feels or functions, or how long a patient survives.

Surrogate Endpoint: A biomarker intended to substitute for a clinical endpoint. A clinical investigator uses epidemiologic, therapeutic, pathophysiologic, or other scientific evidence to select a surrogate endpoint that is expected to predict clinical benefit, harm, or lack of benefit or harm.

Prentice defined the surrogate endpoint in statistical terms [2]. Many investigators have proposed statistical models for evaluating surrogates. Freedman et al. [3], Tsiatis et al. [4], and Daniels and Hughes [5] have addressed the following questions: What is a surrogate endpoint? What evidence is necessary to establish that a biomarker is a useful surrogate for a particular treatment-clinical endpoint combination? Is it useful to investigate the degree of surrogacy that a biomarker captures in the treatment effect on a clinical endpoint? What information about surrogacy is available from a single study in comparison with many studies on similar treatments? An empirical evaluation of a biomarker as a surrogate requires large studies that compare treatments of interest. Such data will allow exploration of the relationship among treatments, surrogate endpoints, and clinical endpoints. These analyses lead to the proposal of models that predict the effect of a new treatment on clinical endpoints, given an observed treatment effect on the surrogate.

For a biomarker to serve as a surrogate for the effect of an intervention on a clinical endpoint at the population level, more is required than just the ability of the marker measured on an individual to predict that individual's clinical endpoint. The extent to which a biomarker is appropriate for use as a surrogate endpoint in evaluating a new treatment depends on the degree to which the

biomarker can reliably predict the clinical benefit of that therapy, as compared to a standard therapy. Such use generally requires extrapolation from data generated for different treatments than the one under investigation. Substituting a surrogate requires that it not only predicts the clinical outcome of interest but also fully captures all the major effects of the new treatment. Surrogate endpoints might also be used to advise patients about modifications of treatment after they have reached a surrogate endpoint but not yet reached the true clinical endpoint.

AN EVOLVING FRAMEWORK: BRIDGING EMPIRICAL AND MECHANISM-BASED KNOWLEDGE

Biostatistics uses quantitative data to estimate biologic parameters and to test hypotheses. The use of a surrogate endpoint would represent a prediction or extrapolation from past information about the relationship of treatment, surrogate, and clinical endpoint to a new similar treatment. The more that is known about the biological mechanism underlying the disease and mechanism of action of the treatment, the more accurate the prediction is likely to be. The current revolution in bioscience promises improved understanding of mechanisms. The challenge for new statistical models of surrogacy is to use mechanistic knowledge to build appropriate statistical models. In some settings, one can exploit knowledge about mechanisms to make reasonable assumptions about the distributions of the parameters that characterize the relationships under study. Both classical and Bayesian statistical methods may be appropriate for utilizing information collected in a variety of settings. New techniques are needed for testing the validity of presumed mechanisms and for updating the evaluation of surrogates with new mechanistic evidence and clinical data. Biostatistical methods for evaluating biomarkers as surrogate endpoints will be needed to combine knowledge based on mechanisms with empirical observations.

Emerging Needs for Analytical Tools in Clinical Research

The workshop focused on the role of biomarkers in clinical research, especially as potential surrogate endpoints in clinical trials. The discussions addressed the uses of biomarkers for screening candidate interventions in phase II trials and as substitute endpoints in phase III confirmatory trials. The role of biomarkers in phase II trials was not controversial; however, some participants expressed considerable concern regarding the use of surrogate endpoints in phase III trials. Presentations at the workshop included case studies and analyses from past clinical studies that demonstrated the use of biomarkers and surrogate endpoints, for example, in the evaluation of Human Immunodeficiency Virus (HIV) antiviral agents, vaccines, and cancer chemotherapy. Other case studies provided substantial evidence of the risk to patients that arises when treatment policy is based on surrogate endpoints that do not fully explain the effects of disease interventions, such as in the case of certain antiarrhythmic therapies [6].

The participants also considered evaluating the usefulness of biomarkers as surrogate endpoints by combining evidence from across studies—a meta-analytical approach. Other discussants presented approaches to predicting the ef-

fect of a new treatment on a clinical endpoint based on the effect of the treatment on a biomarker and on the relationship between the biomarker and the clinical endpoint using data from previously conducted clinical trials.

Participants were also interested in the design and analytic issues related to the use of newer biomarkers, particularly those arising from genomic and proteomic array technologies. Data arising from these technologies have important implications for the design and conduct of drug screening activities and related phase II trials. Statistical aspects of such usage may involve methodologies (e.g., pattern recognition methods and pharmacokinetic models) that have not yet seen much use in clinical trial design or analysis. The need for such new methodologies points to the necessity for statistically and mathematically trained scientists to become better informed about array technologies and other novel screening methods, and about the properties and features of the corresponding measurements. This area of research is particularly important because quantitative intermediate endpoints that are highly sensitive, specific, and reproducibly measured have the potential to improve the efficiency of treatment and prevention trials. The value of this research was recently demonstrated in the use of viral load for development of therapeutics for the treatment of HIV infection; using viral load in this way permits more rapid evaluation of treatments. As in any case of rapid development and evaluation, the longer-term clinical consequences of a treatment require long-term follow-up and the assessment of clinical endpoints. The key to developing useful surrogate endpoints is to identify biomarkers that reflect fundamental aspects of the treatment to be extended on disease pathogenesis. Circumstances in which disease pathogenesis pathways are well understood greatly expand the potential to identify useful biomarkers.

State of the Art: Statistical Approaches for Evaluation of Surrogate Endpoints

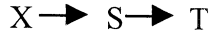
Discussions at this workshop began with a critical review of current methods of evaluating biomarkers as surrogate endpoints. Participants addressed both the utility and limitations of surrogates. The following sections summarize major concepts in the evaluation of surrogates. Surrogate endpoints have been investigated in many disease areas; examples include blood cholesterol level as a surrogate for cardiovascular heart disease morbidity and mortality in studies of cholesterol-lowering drugs, and bone density as a surrogate for fractures in studies of treatments for osteoporosis.

The following sections summarize major concepts in the evaluation of surrogates and use examples from research on treatment of HIV infection.

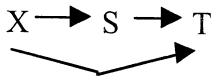
Current Analytical Approaches and Adequacy of Existing Methods

Investigation of biomarkers and their relationship to clinical endpoints may be based on data from a single study, or, more reliably, from multiple studies. The workshop participants reviewed methods for both single and multiple studies, but strongly supported the need for multiple studies to make reliable inferences. Either approach to analysis must begin with a simple model characterizing the relationship between treatment, biomarker, and clinical endpoint.

A "perfect" surrogate endpoint, as described by Prentice, can be represented as



where *X* is treatment, *S* is the surrogate endpoint, and *T* is a clinical endpoint or outcome (often, a time to event) [2]. In this representation, the biomarker *S*, which can be precisely determined, mediates all of the effect of *X* on *T*. A more complex, but more likely, situation arises when *X* has a direct effect on *T* that is not mediated through *S*.



Statisticians have used these models to quantify the extent to which treatment effects are mediated through surrogates. One straightforward approach is to examine the effect of *X* on *T* using a simple regression model, and then test whether including *S* in the model modifies the estimated effect. The ratio of the amount by which a treatment effect on the clinical endpoint is changed after including a surrogate endpoint in the model and the unadjusted treatment effect on the clinical endpoint is sometimes referred to as the proportion of treatment effect (PTE) explained by the surrogate [3, 7]. Participants in the workshop were skeptical of PTE analyses, especially when applied to individual studies, because of the unreliability of the statistical properties of the estimators and the uncertainty of their interpretation [8]. Unless one can perfectly model treatment effects on surrogate endpoints and clinical events, high values of PTE do not necessarily imply that the surrogate endpoint is an important part of the causal pathway leading from treatment to disease.

In analyses that decompose these estimators into their component parts, Buyse et al. demonstrated the difficulty of interpreting PTE [9]. They proposed a model that permits separate evaluation of two components: the effect of treatment on the surrogate endpoint and the effect of surrogate endpoint on the clinical endpoint under the influence of the treatment. The fact that the PTE conflate these two components contributes to the difficulty in its interpretation. More research is required to investigate the usefulness of metrics that, like the PTE, are intended to provide information about whether biomarkers actually do explain the effect of treatment in some settings. In particular, participants challenged statisticians to develop metrics that are reasonably robust to some types of model misspecification and to study whether resampling methods can provide insight into the reliability of these metrics.

In addition to using biomarkers as surrogate endpoints, some researchers have considered using such markers as "auxiliary endpoints" in clinical trials. Fleming et al. defined auxiliary endpoints as "response variables or covariates that can strengthen true endpoint analysis. Specifically, such response variables provide some additional information on true endpoint occurrence times for study subjects having censored times" [10]. Such use of biomarkers requires weaker assumptions than necessary for substituting surrogates for clinical endpoints.

New Methods for Using Biomarkers in Clinical Research: Mathematical Models

Understanding mechanisms of action may allow the use of mathematical models to explicate the relationship among surrogate endpoints, treatment ef-

fects, and course of clinical disease. While such models may never be able to characterize all factors that affect the relationship between surrogate and clinical endpoints, they may capture enough of the underlying mechanisms to help in the selection of treatment and doses, design of studies, and choice of study endpoint. In addition, they may add to the evidence supporting the biological plausibility of substituting a surrogate for a clinical endpoint.

The development of HIV viral dynamic models by Perelson et al. provides an example of mathematical models of a disease process that had a major impact on clinical research [11, 12]. The earliest and simplest of these models was based on an assumption that HIV virions infect target T lymphocytes and turn them into productively infected cells (i.e., cells capable of producing virus) at a rate proportional to the product of the number of virions and of target cells in a body compartment. Each of these cells was assumed to produce a constant number of virions in its lifetime. The administration of potent therapy was assumed to cause all newly produced virions to be noninfectious. Finally, the models assumed a constant rate of clearance of virions and cells. The model, though it oversimplifies the dynamics of virus in the body, had a major impact both on the scientific understanding of HIV infection and on the effect of treatment on infection. For example, the model implied that there are different reservoirs of virus and that clearance of virus happens rapidly. The first phase of viral decay, believed to result from short-lived productively infected cells, has an estimated half-life of 1.1 days; the second phase, also believed to result from long-lived infected cells, has an estimated half-life of 14.1 days.

Recently, several investigators have proposed using viral dynamics to evaluate antiviral therapies [12, 13]. This application of these models may help investigators to learn about the antiviral potency of a new compound or combination of drugs from very short courses of antibiotics. For example, Perelson et al. proposed a method for assessing new antiviral agents based on a parameter called relative efficiency (RE) [12]. The method approximates decay in HIV-1 RNA copies in plasma with a single exponential curve and assumes 100% viral inhibition in patients treated with a three-drug combination regimen. The RE compares the effect of a new agent on viral decay to that of the three-drug regimen. A study of six doses of nelfinavir showed a high correlation between dose and RE (0.97); the average RE was 93% for the highest and 57% for the lowest dose.

Ding and Wu have established a more formal relationship between viral decay rates and treatment potency that has permitted the development of statistical methods for assessing the potency of antiviral therapies using viral decay rates [13]. Such methods allow the use of viral dynamic models in rapid and efficient evaluation of antiviral therapies. This approach is also useful for selecting the therapeutic dose of a new agent and for deciding which agents are potent enough for investigation of long-term durability of effects.

Limitations of Analyses of Proportion of Treatment Effect Explained—A Case Study: Use of Viral Load to Predict Maternal-Child Transmission of HIV-1

One of the most important advances in AIDS research has been reducing the risk of maternal-child transmission of HIV-1 through the use of antiretroviral

drugs. This effect, first demonstrated in an investigation of the benefit of zidovudine monotherapy in over 400 pregnant women and their offspring, is remarkably large—a 70% reduction in the risk of transmission (from 22.6% to 7.6%). This magnitude of the effect is surprising because zidovudine monotherapy has only a modest effect on HIV-1 RNA, the most commonly used biomarker employed as a surrogate endpoint in clinical management of HIV infection.

At the time of delivery, mothers receiving zidovudine had a median HIV-1 RNA value 1.7-fold lower than did mothers receiving placebo. Given the wide range of RNA values, measured either at entry into the study or at delivery in mothers who transmitted HIV-1 to their offspring, one would expect such a modest zidovudine effect to result in little reduction in the rate of transmission. As seen in Figure 1, entry level of HIV-1 RNA in mothers affects transmission rate. The figure also would lead us to expect that the 70% reduction in transmission would require larger reductions in maternal HIV-1 RNA because it is only in the lowest quartile of virus that mothers receiving placebo have a much lower risk of transmission than in the other quartiles. Zidovudine monotherapy is not powerful enough to bring many mothers with higher entry levels of virus down to this lowest quartile.

Sperling et al. calculated that maternal levels of HIV-1 RNA accounted for less than 20% of the effect of zidovudine treatment on transmission [14]. Expressed differently, one would have expected less than a 14% reduction in transmission rate, rather than the observed 70%, if all of the benefit of zidovudine were mediated through level of HIV-1 RNA. According to Sperling et al., "Only a small part of the treatment effect could be explained by the observed RNA measurements, which is further evidence that the protective effect of zidovudine results at least in part from a mechanism other than the plasma viral burden." These analyses demonstrated that while biomarkers like HIV-RNA may be useful for assessing the activity of a drug, they may not provide good predictions of drug effects on clinical endpoints, especially in new clinical settings.

More recent studies have demonstrated benefits of zidovudine and other antiviral drugs in preventing maternal transmission of HIV, but have shown inconsistent PTE accounted for by maternal levels of HIV-1 RNA [15, 16]. For example, a study of the effect of a short course of zidovudine compared to placebo in an international setting had an estimated PTE of about 80% [95% CI: 36–336 (0.36–3.36)] [16]. In this study, mothers receiving zidovudine had a 3.7-fold greater reduction in HIV-1 RNA at the time of delivery than in the AIDS Clinical Trials Group 076 Study, but both studies showed considerable variability in the HIV-1 RNA at delivery, not only among mothers who transmitted HIV to their babies, but also among those who did not. As additional studies are conducted, many with more potent agents than zidovudine alone, the ability to perform meta-analyses on study results will aid in assessing the role of maternal HIV-1 RNA and other important factors in maternal-child transmission.

Meta-Analysis

Several participants discussed meta-analyses results from multiple studies [3–5, 9, 17]. Combining information from multiple studies often provides a

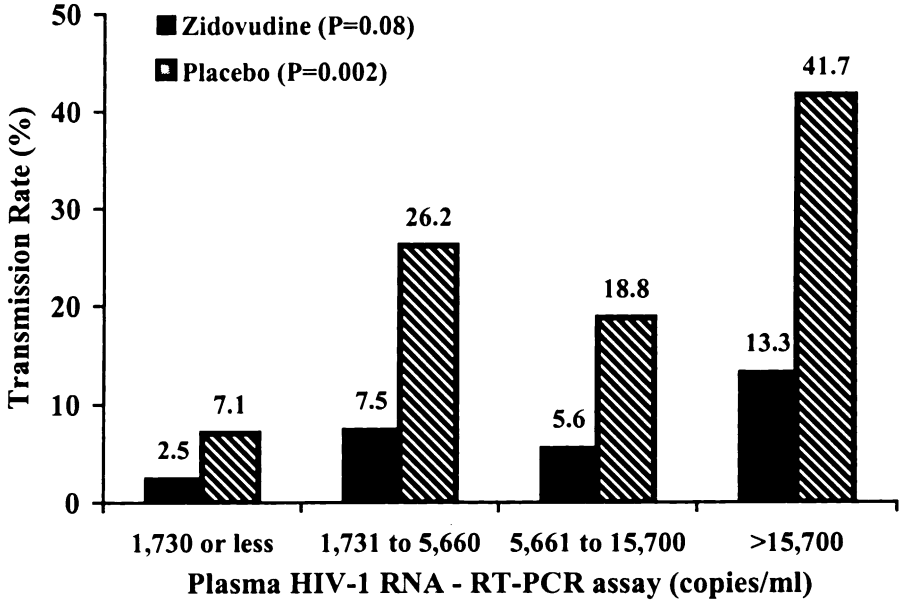


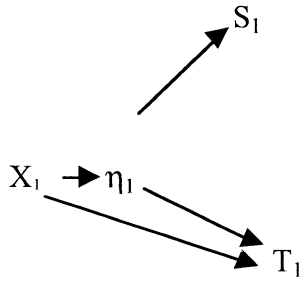
Figure 1. Effect of zidovudine on perinatal HIV transmission as a function of maternal HIV-1 mRNA levels at time of entry in trial (with permission from Sperling et al. [14]).

more accurate basis for extrapolation than using results of individual studies [18]. As in single studies, these meta-analyses investigated the association between treatment effects on the surrogate endpoint S and a true clinical endpoint T. This approach describes the association using results from the model and then assesses the model's reliability for predicting the treatment effect on T, which gives an observed effect on S. Such models assume that the new experimental treatment X_n and its control treatment were drawn from a class of similar studies. Daniels and Hughes assumed the impact of X_n on T and on S is multivariate normal with mean and variance parameters that vary across studies [5]. By "borrowing" information regarding estimates of the effects of X on T and on the relationships between T and S given X in previous studies, they predicted effects of X_n on T from data on the surrogate endpoint. Buyse et al. used a linear mixed model to describe the effects of treatment on S and T [9]. Their methods differed from that of Daniels and Hughes in that they predicted treatment effects on T from data on the separate responses S in treated and untreated groups, rather than from the estimated treatment effect on S alone.

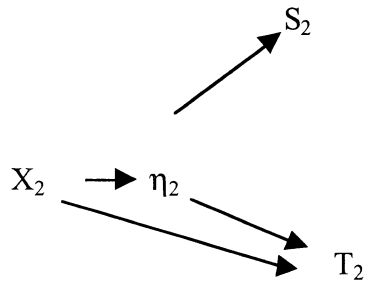
Two important considerations for further development of meta-analytic models are that markers are likely to be measured with error and that marker measurements may only indirectly reflect important processes on the causal path to clinical disease. Meta-analyses may be based on disease models that can accommodate these features. Xu and Zeger proposed a latent disease model for analysis of studies in which both the surrogate endpoint and the true clinical endpoints are mediated through a latent variable η, shown below

[18]. Such models can accommodate situations in which biomarkers *S* are measured with error and involve other factors besides η.

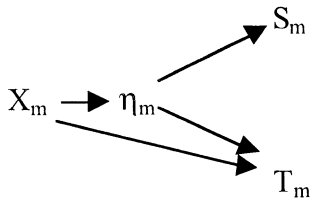
For study 1,



For study 2,



And in the *m*th study:



As with the approaches of Buyse et al. [9] and of Daniels and Hughes [5], one goal of such an analysis is to predict the time to event for a new treatment of the same class. Achieving this goal requires characterizing the distribution of estimates of treatment effects across the studies. Xu and Zeger's model is similar to those described earlier in that it permits prediction of the clinical effect of a new treatment on the basis of its effect on the biomarker, but it differs from those described earlier in that it depends more heavily on patient-specific data [18]. Collection of this information will require registries and databases. Timely analysis of data would allow sharing of prediction models. The development of prediction models presents an opportunity to work with the pharmaceutical industry because prediction of clinical response would help select candidate drugs early in drug development. Models would be useful for such goals as prioritization of candidate therapeutics and for selection of appropriate doses and target populations.

Models such as those described will need extensive exploration, as inadequate models could lead to potential bias in projections. Further research is also required to characterize the statistical properties of estimation procedures. Specifically the models should be examined with respect to:

- Sensitivity analyses to study assumptions regarding effects of treatment on surrogates and markers as well as variability in these effects across studies and to examine assumptions regarding errors in measurement.
- Similarity of mechanisms of action of the new drug or other intervention compared to those previously studied. Investigators must define a class of

similar studies and specify rules for determining which studies to include; inappropriate rules could lead to selection bias. If the effects on the surrogate are at the limits of previous observations, predictions may not be reliable.

- Need for characterizing the uncertainty in parameter estimation and prediction. Uncertainty will depend on the amount of between-study variation in the parameters that describe the relationship between the effects of treatment on surrogate endpoints and on true clinical endpoints and on how precisely the parameters describing between-study variation can be estimated [17].
- Difference in approaches required for prospective and retrospective analyses. Although retrospective meta-analyses have provided much of the information assessed across studies to date, prospective planning will make meta-analyses more reliable. Strategies for collecting information for prospective analyses include establishing and monitoring registries of trials with a common theme. The prospective approach may provide more useful information about different drugs and potential surrogates.
- Appropriateness of the application of meta-analyses and predictive models. Meta-analyses may be appropriate for situations with a broad range of outcomes as well as for those with a single major clinical outcome, but the former situation requires more complex statistical models.
- Consideration of the type of data used in conducting the analysis—clinical trial level or individual patient level. Individual patient data provide more information to serve as a basis for prediction but are more difficult to collect.
- Need to assemble databases, registries, and shared data sets that focus on the information required for meta-analysis.

Prospective design of studies intended for meta-analysis requires defining the class of similar studies as well as the specific measures of biomarkers of interest and presents an opportunity to examine more than one surrogate endpoint. The prospective approach requires agreement on the surrogate endpoints and ways of measuring them as well as on the true clinical endpoints in the various studies.

Example of Meta-Analysis

The meta-analysis conducted by Daniels and Hughes has implications for the design and conduct of trials [5]. This analysis combined results across a number of clinical trials of antiretroviral drugs for patients with HIV infection. The surrogate endpoint under consideration was CD4+ T-lymphocyte count and the true clinical endpoint was onset of AIDS or death. To measure treatment effects on the clinical endpoint, Daniels and Hughes used the log of the ratio of the hazard of developing AIDS or death (whichever came first) for two treatments under study. The goal of their analysis was assessment of the reliability of predicting the treatment effect on clinical disease given an observed effect on CD4+ count. Such analyses require assumptions about the effects of treatments on these outcomes; they assumed the differences between treatment arms in CD4+ count and in the log of the hazard ratio were bivariate normal

with mean and variance parameters that vary across studies. Information from previous studies, on the relationship between the effects of treatment on clinical disease and on the surrogate, allowed them to predict the clinical effect of a new treatment from its effect on CD4+ count.

To fit their model for meta-analysis across a range of clinical trials of antiretroviral drugs, Daniels and Hughes used a Bayesian approach. Such approaches are useful for combining information across studies when it is reasonable to make assumptions about the distributions of the parameters that characterize the relationships under study. When little is known about these distributions, as in this case, one can try to limit the effect of these assumptions by selecting noninformative prior distributions. Alternatively, one could perform several analyses using a range of prior distributions.

In comparing two treatments, a larger CD4+ benefit tends to be associated with a larger clinical benefit, but the magnitude of the benefit cannot be precisely estimated from data on CD4+ counts alone. Furthermore, larger joint CD4+ and clinical benefits are only seen in the placebo-controlled trials, so results may not generalize to active-controlled trials with larger CD4+ effects. Figure 2 plots the association between the log hazard ratios of developing AIDS or dying and the difference in the mean change in CD4+ cell count for placebo-controlled trials (labeled "p") and for active-controlled trials (labeled "a"). The ellipses around the estimates for each study show the 95% confidence regions associated with the observed results from each study. The large size of the confidence regions underscores the need to assemble a considerable amount of information to make reliable inferences. In addition, the larger effects in placebo-controlled studies imply that more information from active-controlled trials is needed to determine whether the association between larger CD4+ and clinical benefits will hold true for this group of trials as well.

General Assessment of Current Analytical Approaches

The participants generally agreed that analysis of surrogate endpoints making use of cross-protocol analyses (or meta-analysis) of clinical trials is likely to be more robust than analyses of single studies, unless the single study is very large. Meta-analyses may be particularly advantageous when available studies evaluate effects of different classes or types of interventions on biomarkers and clinical endpoints. Of particular interest are analyses that use effects on biomarkers to predict effects of treatment on clinical endpoints. Information pertaining to treatment, biomarker response, and clinical endpoint are required to build prediction models.

SUMMARY OF WORKSHOP GOALS AND RECOMMENDATIONS

An overarching aim of the workshop was to develop recommendations to guide NIH and the research community at large in the design and analysis of methods for evaluating the uses of biomarkers as surrogate endpoints in clinical trials. In doing so, participants recognized that many successful clinical trials use neither biomarkers nor surrogate endpoints. Information derived from trials of clinical outcomes, including adverse effects, is extremely important in assessing the clinical utility of an intervention. Often the combined information

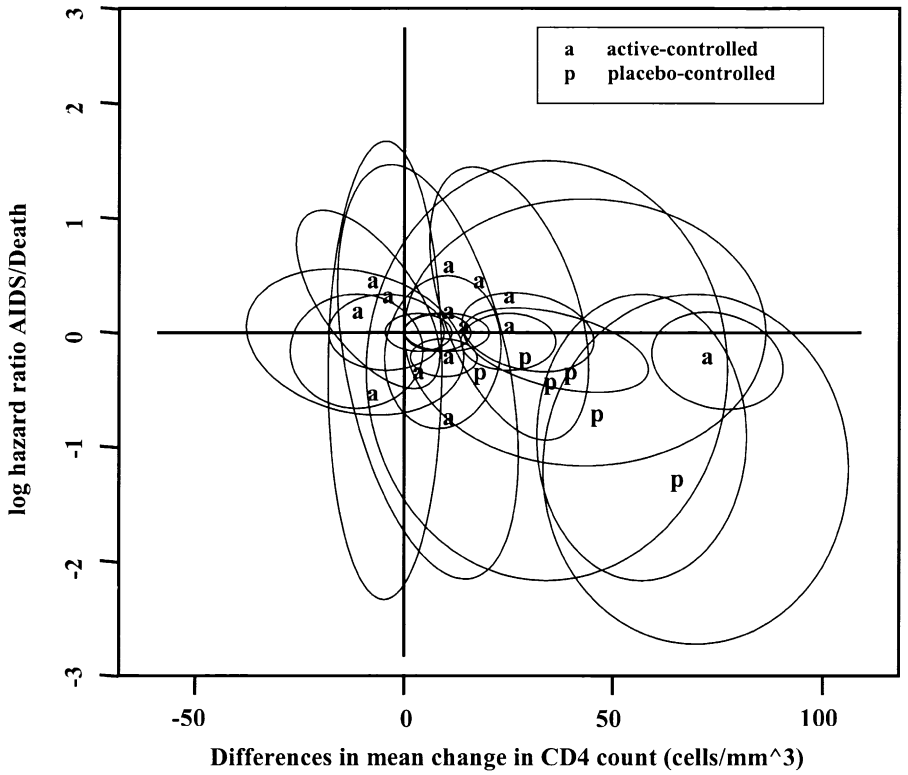


Figure 2. Association of log hazard ratio of developing AIDS or dying and the difference in the mean change in CD4+ cell count for placebo-controlled trials (labeled "p") and for active-controlled trials (labeled "a"). Ellipses represent 95% confidence intervals associated with the observed results from each study (with permission from Daniels and Hughes [5]).

from biomarkers and clinical outcomes provides the strongest rationale for optimal use of interventions. A major incentive for investigation of surrogate endpoints is improving the efficiency of clinical trials.

Workshop participants defined recommendations to address the following five goals related specifically to evaluation of surrogate endpoints:

- development of a statistical and inferential framework,
- design and conduct of trials,
- assemblage of databases,
- assessment of complex statistical and mathematical models, and
- meeting educational needs for the medical and statistical research communities.

Goal 1. Development of a Statistical and Inferential Framework

The workshop participants emphasized the need to establish a statistical and inferential framework for evaluating candidate biomarkers in the exploratory phases of drug development and for investigating surrogate endpoints in confirmatory trials. They emphasized that clinical endpoint trials—random-

ized clinical trials based on clinical endpoints that directly measure how patients feel and function and whether they survive—remain the ultimate priority of clinical investigations of novel interventions. Unless a potential surrogate endpoint adequately conveys necessary information about the relationship of treatment to the relevant clinical endpoint, then clinical outcome trials will still be required.

Recommendations

- A statistical framework must aid in assessing the medical conditions under which one might consider investigating a biomarker as a potential surrogate endpoint in a confirmatory clinical trial. The goals for statistical investigation of biomarkers should be expanded to include investigation of a candidate therapy in the exploratory phase of therapeutic development with an emphasis on the relation of the biomarkers to the mechanism of action of the intervention. This understanding is necessary to identify settings to evaluate biomarkers as potential substitutes for clinical endpoints.
- Establishment of the framework requires consideration of the need for statistical expertise and reasoning at all stages of therapeutic development. Statisticians should be involved in the design of trials that will permit investigation of disease mechanisms using biomarkers. These studies include the exploratory phase of the development of drugs or other interventions as well as phase II and confirmatory trials. Such investigations should consider pharmacokinetics, pharmacodynamics, and disease modeling.

Additional research is needed to develop hierarchical models that are based on scientific understanding of mechanisms (including nonlinear models where appropriate) and can be applied for continuous, discrete, or censored outcomes. Applications of hierarchical models should also be considered for meta-analysis using patient-level data.

Goal 2. Future Development of Methods and Standards for the Design and Conduct of Clinical Trials

Evaluation of biomarkers typically requires information across a variety of studies. Only if studies collect appropriate information in standard ways can cross-protocol investigation produce useful and reliable results. The required type and frequency of measurements of markers depends on the use to which the marker is put. Investigators intending to carry out cross-protocol analyses should set standards for data collection across studies, especially regarding the items most important for evaluation of surrogacy.

Successful evaluation of a biomarker will require refining study designs for gathering appropriate data, such as case-cohort and two-stage designs, and long-term follow-up of patients to assemble sufficient data to assess performance of the biomarker.

Recommendations

- Collect follow-up data of all patients in clinical trials, regardless of treatment status.

- Insofar as consistent with the goals of a specific trial, standardize schedules of efficacy and toxicity measurement, and other study procedures.
- Develop designs to reduce cost, such as cohort sampling.
- Make multiple measurements of biomarkers at baseline and other time points to permit assessment of variability and to allow unbiased estimates of baseline levels.
- Develop other types of designs, such as two-stage designs, to maximize analytical options.

Goal 3. Assemblage of Databases for Evaluating Surrogate Endpoints

Ongoing evaluation of biomarkers, both candidate surrogates and those already in use as surrogates, requires assemblage of pertinent databases. The sources of information for such evaluation include established epidemiological cohorts and randomized clinical trials. Such investigation is most fruitful if investigators have access to a variety of sources of data. Given the importance placed by workshop participants on meta-analysis of clinical trials, it will be increasingly important to provide investigators with information on biomarkers and on disease progression across classes of treatments from randomized studies.

Participants concluded that assessment of the validity of surrogate endpoints requires substantial amounts of data on biomarkers and clinical endpoints across studies. Databases must capture a wide array of studies to ensure adequate information and to protect against possible bias resulting from selectively including studies with a positive outcome (i.e., excluding studies that do not show statistically significant evidence of benefit). Such analyses have increased reliability when they include studies demonstrating a range of effects on markers and on clinical endpoints. The use of datasets from epidemiological studies and trials of treatments on established biomarkers, including the meta-analysis described above, led to the acceptance of HIV-1 RNA (in HIV/AIDS) as a surrogate endpoint for the approval of antiretroviral therapy for patients with HIV infection.

Recommendations

- Mechanisms must be developed to support an infrastructure for assembling and managing such databases. Rapid advances in information technologies can assist data collection, data archiving, and retrieval of clinical trials data to allow statistical analysis.
- Public and private organizations supporting clinical research should encourage sharing of data. Establishment of electronic means for sharing data is important for meta-analysis. Such resources and analyses are valuable when timed in accordance with the emergence of new research questions, provided that the analysis does not compromise any individual study. Efforts should be undertaken to minimize the variability in biomarker measurements across clinical trials, as this is important in the assessment of valid surrogate endpoints. This aim can be aided by collecting information in standard ways as well by specifying the times at which measurements are required.

- Collection of new kinds of data should be encouraged for evaluating a biomarker as surrogate endpoints. Necessary information may include photographic or radiographic documentation, composite data from gene or protein arrays, patient reported behavioral scales, and databases from clinical studies.
- Surveillance systems that collect data on efficacy and/or adverse events should be established. Such systems would provide an added level of protection against erroneous acceptance of biomarkers as surrogates, especially when therapies are licensed on the basis of short-term changes in surrogate endpoints. Active surveillance is preferable to passive systems with optional submission of data to the database.

Goal 4. Further Development of Methods to Use Mechanistic Knowledge in Biomarker Evaluation

Workshop participants foresee biomedical research yielding a rapidly increasing understanding of disease pathogenesis, thereby facilitating selection and application of biomarkers. Selection and evaluation of new biomarkers will require development of new statistical models to evaluate biologic mechanisms. The participants pointed to the value of the HIV viral dynamics and diabetes complication models as examples of how to inform therapeutic development and assessment in other disease settings.

Recommendations

- Develop models that can accommodate measurement error and missing data/informative censoring for investigating biomarkers in different disease areas.
- Evaluate latent variable and other models using patient-specific data for prediction.
- Establish prediction models that accommodate multiple surrogate endpoints and/or multiple clinical outcomes.
- Develop methods to integrate patient noncompliance on assessment of surrogate endpoints.
- Build models to incorporate longitudinal measurement of biomarkers and sequential treatments (in some cases treatments that may be influenced by biomarkers).
- Consider a variety of estimation procedures (e.g., classical methods, empirical Bayes, or Markov Chain Monte Carlo techniques).
- Develop methods for assessing reliability of prediction.

Goal 5. Meet Educational Needs for the Medical and Statistical Research Communities

Participants emphasized the need for training to augment the medical and biostatistical workforce currently engaged in research on design and analysis methods to meet future needs. Rapid advances in discovery tools and high throughput technologies for biological and clinical data have presented new challenges and strained the capacity of current statistical approaches. Competi-

tion for highly trained computer scientists and mathematicians among other fields of science and the marketplace has hampered recruitment efforts of new, quantitatively sophisticated investigators to biomedical research.

Recommendations

- Training of statisticians in specific areas of biologic research (e.g., genetics, molecular biology, and computational biology) and continuing education in biostatistics, particularly in the areas of methodological development and mathematical modeling, is needed. Statisticians must receive training in approaches to complex models that will be used more frequently as biomarker research evolves. Sabbatical years may be considered for intensive training terms with experts from other disciplines or clinical scientists to enhance integration of knowledge and insights from multiple disciplines.
- Biostatistical programs should expand the breadth and depth of their training in biological science, especially in molecular biology and genetics. As the challenges in science and medicine become more complex, successful solutions will require increasing cross-fertilization among researchers of different disciplines. Clinical researchers, basic scientists, and mathematical and analytical scientists should discuss innovative approaches to career enhancement and training.
- Medical researchers will benefit by participating in clinical research training programs that address complex issues on clinical trials arising in the investigation of surrogate endpoints.
- Training of statisticians in the principles of pharmacokinetics and pharmacodynamics is necessary to facilitate the development of models to investigate toxicity or safety markers (or predict toxicity or safety outcomes).
- Continuing education programs to provide updates on new methodology, new model development, and state-of-the-art information about disease areas are needed to advance the field of statistics.

REFERENCES

1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacol Ther* 2001;69:89–95.
2. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med* 1998;8:431–440.
3. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992;11:167–178.
4. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *J Am Stat Assoc* 1995;90:27–37.
5. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med* 1997;16:1965–1982.
6. Fleming TR, DeMets DL. Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med* 1996;125:605–613.
7. Lin DY, Fleming TR, DeGruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat Med* 1997;16:1515–1527.

8. DeGruttola V, Fleming T, Lin DY, Coombs R. Perspective: Validating surrogate markers—Are we being naive? *J Infec Dis* 1997;175:237–246.
9. Buyse M, Molenberghs G, Buzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000;1:49–67.
10. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* 1994;13:955–968.
11. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1-dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* 1996;271:1582–1586.
12. Perelson AS, Essunger P, Cao Y, et al. Decay characteristics of HIV-1 infected compartments during combination therapy. *Nature* 1997;387:188–191.
13. Ding AA, Wu H. Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics. *Math Biosci* 1999;160:63–82.
14. Sperling RS, Shapiro DE, Coombs RW, et al. Maternal viral load, zidovudine treatment, and the risk of transmission of human immunodeficiency virus type 1 from mother to infant. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *N Engl J Med* 1996;335:1621–1629.
15. Shaffer N, Chuachoowong R, Mock PA, et al. Short-course zidovudine for perinatal HIV-1 transmission in Bangkok, Thailand: A randomized controlled trial. Bangkok Collaborative Perinatal HIV Transmission Study Group. *Lancet* 1999;6;353:773–780.
16. Executive Summary of *HIVNET* 012. <<http://www.niaid.nih.gov>>
17. Gail MH, Pfeiffer R, Van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000;1:231–246.
18. Xu J, Zeger SL. The joint analysis of longitudinal data comprising repeated measures and times-to-events. *J Royal Stat Soc*, Series C (Applied Statistics), 2001;50:375–387.