Statistical Issues in Evaluation of Surrogate Endpoints

Victor De Gruttola

Mitchell Gail

Scott Zeger

Among the most important issues in the design of a clinical trial is the selection of the primary endpoint. Fleming et al. (1998) provided two criteria to govern its selection: the endpoint should be (i) sensitive to treatment effects, and (ii) clinically relevant. Endpoints that directly reflect how a patient feels, functions, or survives have clear clinical relevance; we refer to these as true clinical endpoints. The cost in time and resources associated with the choice of such endpoints has led to consideration of whether measures of biologic activity if of a treatment, or biomarkers, can be appropriate as surrogates for clinical endpoints. We use the term "surrogate endpoint" to describe biomarkers intended to substitute for a clinical endpoint in a clinical trial. While use of such endpoints in early phase trials is well accepted, their use in Phase III clinical trials--trials intended to define the role of a therapy in standard clinical practice is more controversial. Fleming and DeMets (1996) reviewed previous experience with use of surrogate endpoints in a variety of disease settings to underscore the difficulties in developing and using such endpoints.

Evaluation of the utility of a surrogate endpoint requires consideration of the extent to which treatment effects on the surrogate assure comparable treatment effects on endpoints with more direct clinical relevance. A major goal of this workshop is to consider different ways of formulating this evaluation as a statistical problem, and to consider analytic approaches for its solution. Other important goals include defining the information needed for such analyses and exploring ways to investigate their reliability. What follows is a brief discussion of some published articles regarding surrogate endpoints, and of some new work to be presented at the Workshop. This statement is by no means intended to be a thorough review, but simply to provide some focus for Workshop discussion by describing connections among different approaches to this problem.

Prentice (1989) developed a statistical definition of a surrogate as "a response variable 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". His approach to validation involved two criteria that provide guidance about analyses for assessment of markers:

1. the surrogate endpoint must be correlated with the true clinical endpoint and;
2. the surrogate should fully capture the treatments "net effect" on the clinical endpoint, where the net effect is the aggregate effect resulting from all mechanisms of action.

This restrictive second condition implies that if a marker is a good surrogate for *T*, a true time-to-failure, clinical endpoint, then the hazard of *T* should be independent of treatment, conditional on the surrogate marker i.e., all the beneficial effect of treatment is mediated through the marker. Much of the investigation of surrogate endpoints in medical research has focused on assessment of whether this condition is met. In Prentice's formulation, the notion that the surrogate *S(t)* could capture the dependence of *T*, a true is expressed as:

$$\lambda_T(t; S(t), x) = \lambda_T(t; S(t)) \quad .$$

Freedman, Graubard, and Schatzkin (1992) implemented Prentice's criterion by examining whether an effect of an intervention on a true clinical endpoint, adjusted for the intermediate or surrogate endpoint, is reduced to zero. For this assessment, they considered the proportion of the treatment effect that can be accounted for by the marker. They discussed the case of a binary endpoint and assumed a linear logistic regression model relating this endpoint to the marker and the surrogate. Using this model, they proposed the following metric, which we refer to as "proportion of treatment effect explained" or PTE:

$$p_o = 1 - \frac{\beta_\alpha}{\beta}$$

where the β refers to the net treatment effect (logistic regression parameter relating treatment to clinical endpoint) and β_a refers to the treatment effect after inclusion of the surrogate in the model (the unexplained portion of the treatment effect). For consideration of a true endpoint that

is a failure-time endpoint, a number of authors, including Choi et al. (1993), O'Brien et al. (1996), and Lin et al. (1997), assumed a proportional hazards model for the effects of both the treatment and the surrogate. These authors considered estimation of a quantity analogous to the PTE of Freedman et al.; Lin et al. also propose an estimator for its variance.

Tsiatis et al. (1995) and De Gruttola et al. (1993) considered an alternative approach to estimating PTE. They use a mixed effect model for the surrogate endpoint process and assume a proportional hazard relationship to the time of true clinical endpoint, T. From this model, they predicted what the benefit of experimental treatment on survival would have been (compared to placebo), had all of the benefit of treatment resulted from improvement in the surrogate. Comparison with the actual effect of treatment allowed for graphical display of the PTE. Buyse and Molenberghs (1998) considered replacing PTE with other measures to assess the quality of a surrogate. The first one, termed relative effect, is the effect of the treatment on the true endpoint relative to that on the surrogate endpoint. The second one is the adjusted association between both endpoints, after accounting for the effect of treatment.

In addition to the approaches described above, some authors have considered estimating the effect of treatment on the true clinical endpoint, T, from data on the surrogate S. We first discuss surrogates that do not vary with time, and consider the case where interest lies in estimating the distributions $[T|X=1]$ and $[T|X=2]$ for the new treatment $X=1$ and the control treatment $X=2$. We assume that available data include complete observations on $[T,S|X]$, as well as incomplete observations on $[S \mid X]$. If we have models for the joint distributions $[T,S|X=1]$ and $[T,S|X=2]$, then established methods for missing data can be used to strengthen inference on $[T|X=1]$ and $[T|X=2]$. Such methods require further assumptions about whether the "complete" and "incomplete" observations are representative samples. Surrogates have been termed "auxiliary outcome data" in this application (Pepe, Reilly and Fleming, 1994). In the usual application of surrogates, however, all data on the new treatment is "incomplete"; that is, only data on S given X are available. In this case, some other information on the relationships between T and S given X are needed to allow estimates of treatment effects on T from data on S. Meta-analysis of results from previous trials is one possible way of developing such information.

Buyse et al., (1998) Molenberghs et al., (1998) and Daniels and Hughes (1997) described meta-analyses of results from clinical trials to investigate the association between treatment effects on the surrogate *S* and treatment effects on the true clinical endpoint *T*. Daniels and Hughes modeled this association using results from trials of antiretroviral agents, and then assessed the model's reliability for predicting the treatment difference on *T*, given an observed difference on *S*. Their model assumed that a new *(N)* experimental treatment and its control treatment were drawn from a class of similar studies, C . In their model, the impact of treatment on *T* and on *S* was assumed to be multivariate normal with mean and variance parameters that vary across studies. By "borrowing" information from previous studies on the relationship between the effect of treatment on *T* and the effect of treatment on *S,* which they assumed to be linear, they could predict the treatment effect on T_N from data on the surrogate $S_{N.}$. To fit their model for meta-analysis across a range of clinical trials of antiretroviral drugs, Daniels and Hughes used a Bayesian approach that assumed non-informative prior distributions on the parameters defining the linear relationship between treatment effects on T and those on S, and they used Markov Chain Monte Carlo techniques.

Buyse, Molenberghs, Buryzynowski, Renard and Geys (BMBRG), in unpublished work to be discussed at the conference, used a linear mixed model to describe the effects of treatment on S and on T. Data from a meta-analysis of previous studies in the class *C* are used to estimate the parameters of this model. Then, given data from the new study N on the surrogate in both treated and untreated groups, the parameter defining the effect of treatment on T in the new study can be estimated. This method differs from that in Daniels and Hughes in several respects, the most important of which is that BMBRG predict 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.

In work to be presented at the workshop, Gail considers a meta-analytic model similar to that of BMBRG and generalizes it to handle more complicated outcomes. Let $(T_{1i,} S_{1i,} T_{2i,} S_{2i})$ be the vector of sample mean responses in a previous "complete" data experiment, i, in the class *C,* where the subscripts 1 and 2 indicate *Z=1* and 2 respectively. Gail then assumes that the corresponding population means for experiment i, $\mu_{1ti}, \mu_{1si}, \mu_{2ti,}$ *and* μ_{2si}, have a joint normal

distribution. Thus, even though (T_{1i},S_{1i}) and (T_{2i},S_{2i}) are conditionally independent given these population means, they are correlated unconditionally. Gail allows (T_{1i},S_{1i}) and (T_{2i},S_{2i}) to have different covariances, given their population means, unlike BMBRG. This generalization may be useful if treatment affects both the mean and variance of the distribution and for applications to non-linear models. Analysis of a series of previous studies with complete data permits estimation of the distribution of μ_{1ti}, μ_{1si}, $\mu_{2ti,}$ and μ_{2si} in samples from the class C. In general, the optimal estimate of the difference in the underlying population means of the main endpoint, μ_{1tN} - μ_{2tN}, in the new experiment is obtained by regressing this quantity on both S_{1N} and $S_{2N,}$ rather than simply on the difference, S_{1N} - $S_{2N,}$ that was used by Daniels and Hughes. In many cases, however, the latter regression is nearly as efficient as the former. This analysis shows why meta-analytic approaches can be very inefficient, however, compared to having data on the main endpoint.

Gail also presents a generalization useful for more complex models, such as piecewise exponential survival models and models for repeated measures. These models need only describe the marginal distributions of $T_{1ij,}$ $S_{1ij,}$ $T_{2ij,}$ and S_{2ij}, , where j indexes the j^{th} individual in study i on the active (1) or control (2) treatment. Gail assumes that the parameters describing these four marginal distributions come from a multivariate normal population over studies in the class *C*, and that conditional on the parameters of a given study, estimates of these parameters are asymptotically jointly normally distributed, as would usually be the case. Complete data from previous studies can be used to estimate the distribution of parameters in repeated sampling of studies from *C*. Hence one can estimate the parameters governing T_{1ij}, and $T_{2ij,}$ in the new study from estimates of the parameters governing $S_{1ij,}$ and S_{2ij} in the new study.

Some general issues that may limit the acceptance of the meta-analytic approach include: defining the class, C, of "similar" experiments; developing realistic models for the joint distributions of *T* and *S* given *Z* and given underlying population parameters; developing realistic distributions that govern the sampling of parameters for a particular experiment; including covariates to control the between experiment variation; and allowing for other factors, such as idiosyncratic toxicities, that are not related to the main endpoint. This last issue affects conventional studies of main endpoints as well.

For settings in where the surrogate, S is measured repeatedly, models developed for analysis of longitudinal data with missing, possibly non-ignorable observations (Little and Rubin, 1987), may be particularly useful for the joint analysis of S and T as functions of covariates X. Three broad classes of models have been discussed: selection models in which $[T, S|X]$ is decomposed into $[T \mid S, X][S \mid X]$ (e.g., Diggle and Kenward, 1994; Baker, 1994; Molenberghs et al, 1996; Fitzmaurice et al, 1996); pattern-mixture models where the decomposition $[S \mid T, X]$ $[T \mid X]$ is used (eg, Lagakos, 1976; Little, 1993; Hogan and Laird,1997); and latent variable models where $[T, S|X] = \int [T, S \mid \eta, x]\, d\,[\eta \mid x]$ and η is an unobserved latent variable that makes the surrogate information S informative about T (eg, Heckman, 1979; Wu and Carroll, 1988; Wulfsohn and Tsiatis, 1997; and Fawcett and Thomas, 1996). Hogan and Laird (1997) provide an excellent overview of recent work.

In the Workshop, Scott Zeger will present approaches based on a latent variable model in which the relationship between T and S given X is assumed to come from a latent process η. This model assumes that: 1) T and S are conditionally independent given η; and 2) X can affect T either through η or directly, but that X only affects S through its influence on η. In this sense, S is an imperfect measure of η. These assumptions lead to:

$$[T, S \mid X] = \int [T, S, \eta \mid x]\, [\eta \mid x]\, d\eta$$
$$= \int [T \mid \eta, X]\, [S \mid \eta]\, [\eta \mid X]\, d\eta \quad .$$

The formulation is completed by assuming: a regression for S given η and one for T given η and X; and a model for the underlying process η.

In the study of a surrogate, the quantity of interest is $[T \mid X] = \int [T \mid \eta, X]\, d\,[\eta \mid X]$ or some functional such as the ratio of cumulative hazards for the two treatment groups. These measures of treatment effect average over the unobserved η and in so doing, make use of the repeated measures S, which are informative about T when it is censored for an individual. This approach has also marked possible calculation of predictive distribution, for a patient with a particular

history *S(t)* or for a population. It can also be extended into hierarchal model for data from a particular study.

The latent variable formulation is conveniently implemented with a Bayesian approach using Markov chain Monte Carlo methods, which facilitate integration over η. See Fawcett and Thomas (1996) for an example.

As mentioned above, this statement highlights only a few of the ways to address the problem of evaluating surrogate endpoints that will be discussed at the Workshop. A major goal of the Workshop will be to consider strengths and weaknesses of different approaches, and to discuss the connections of these approaches among each other and to and to related areas of statistical research.

References

1. Baker, S. "regression analysis of group survival data with incomplete covariates," *Biometrics*, **1994**, 821-826

2. Buyse, M., Molenberghs, G., "Criteria for the Validation of Surrogate Endpoints in Randomized Experiments," unpublished

3. Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., "The validation of surrogate endpoints in meta-analyses of randomized experiments," unpublished

4. Choi S, Lagakos SW, Schooley TT, and Volberding PA, " CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine," *Annals of Internal Medicine*, **1993**. 118: 674--680.

5. Daniels MJ, Hughes MD, Meta-analysis for the evaluation of potential surrogate markers. *Stat Med*. **1997**, Sep 15;16(17):1965-82.

6. De Gruttola, Tsiatis, Wulfsohn and Fischl , "Modelling the relationship between survival and CD4 lymphocytes in patients with AIDs and AIDS-related complex," *J. Acquired Immune Deficiency Syndromes,* **1993**, 6, 359-365.

7. Diggle, P.J. and Kenward, M.G. "Informative dropout in longitudinal data anlaysis (with Discussion). *Applied Statistics*, 1994, 43:49-93

8. Fawcett C and Thomas D, "Simultaneously modelling censored survival data and repeatedly measured covariates," *Statistics in Medicine,* **1996**, 1663-1685

9. Fitzmaurice, G., Heath, A., Glifford, P. "Logistic Regression models for binary panel with attractions, *Journal of Royal Statistical Society*, **1996**, 249-263

10. Fleming , TR and DeMetz DL, "Surrogate endpoints in clinical trials: Are we being misled?" *Annals of Internal Medicine*, **1996**, **125**, 605–613.

11. Fleming, TR, DeGrttola, V, DeMets, DL, "Surrogate Endpoints," *Encyclopedia of Biostatistics*, **1998**, 4425-4431

12. Freedman LS, Graubard BI, Schatzkin A., "Statistical validation of intermediate endpoints for chronic diseases," *Stat Med*. **1992**, Jan 30;11(2):167-78.

13. Heckman JE, Yin S, Alzner-DeWeerd B, Raj Bhandary UL "Mapping and cloning of Neurospora crassa mitochondrial transfer RNA genes," *J Biol Chem,* **1979** Dec 25;254(24):12694-700

14. Hogan JW, Laird NM, "Model-based approaches to analysing incomplete longitudinal and failure time data," *Stat Med*, **1997,** Jan 15-Feb 15;16(1-3):259-72

15. Lagakos SW. "A stochastic model for censored-survival data in the presence of an auxiliary variable," *Biometrics*, **1976**, Sep;32(3):551-9

16. Lin D.Y., Fleming T.R., and De Gruttola V., "Estimating the proportion of treatment effect explained by a surrogate marker," *Stat Med* **1997,** Jul 15;16(13):1515-27.

17. Little, R. and Rubin, D., "Unbalanced ANOVA design; Nonresponse; EM algorithm," *Wiley*, **1987**, 278

18. Little, R., "Statistical analysis of masked data," *Journal of Official Statistics,* **1993**, pp 407-426

19. Molenberghs G, Ritter LL "Methods for analyzing multivariate binary data, with association between outcomes of interest," *Biometrics*, **1996**, Sep;52(3):1121-33 .

20. Molenberghs, G., Geys, H., Buyse, M., "Validation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcome, " unpublished

21. O'Brien W., Hartigan P., Martin D., " Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS," *New England Journal of Medicine*, **1996,** 334: 426-431.

22. Pepe M.S., Reilly, M., and Fleming, T.R.. "Auxiliary outcome data and the mean score method," *J. Stat Planning and Inference,* **1994,** 42:137-160.

23. Prentice, R. L., "Surrogate markers endpoints in clinical trials: definition and operational criteria," *Stat Med.* **1989**, 431-440.

24. Tsiatis A, De Gruttola V, Wulfsohn M., "Modelling the relationship of survival to longitudinal data measured with error Application to survival and CD4 counts in patients with AIDS," *Journal of the American Statistical Association*, **1995**, 90:27--37.

25. Wu, M., Carroll, R., **"**Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process,"*Biometrics,* **1988**, Vol. 44:175-188.

26. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics.* **1997** Mar;53(1):330-9.