Perspective: Validating Surrogate Markers—Are We Being Naive?

V. De Gruttola, T. Fleming, D. Y. Lin, and R. Coombs

Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; Department of Biostatistics, University of Washington, and Department of Laboratory Medicine, University of Washington, Pacific Medical Center, Seattle

Because of the difficulties in conducting studies of clinical efficacy of new therapies for human immunodeficiency virus infection and other diseases, there is increasing interest in using measures of biologic activity as surrogates for clinical end points. A widely used criterion for evaluating whether such measures are reliable as surrogates requires that the putative surrogate fully captures the "net effect"—the effect aggregated over all mechanisms of action—of the treatment on the clinical end point. The variety of proposed metrics for evaluating the degree to which this criterion is met are subject to misinterpretation because of the multiplicity of mechanisms by which drugs operate. Without detailed understanding of these mechanisms, metrics of "surrogacy" are not directly interpretable. Even when all of the mechanisms are understood, these metrics are associated with a high degree of uncertainty unless either treatment effects are large in moderate-size studies or sample sizes are large in studies of moderately effective treatments.

A clinical trials program evaluating interventions for patients infected with human immunodeficiency virus (HIV) must provide a reliable assessment of safety and clinical efficacy. By definition, the measures of clinical efficacy should be outcomes "which unequivocally reflect tangible benefit to the patient" [1]. For example, does a new intervention prolong survival, increase the length of time that an infected person can continue regular recreation or work-related activities, reduce pain or symptoms of AIDS-related dementia, or prevent loss of vision?

Obtaining reliable evaluations of treatment effects on these clinical efficacy measures often requires conducting randomized phase III clinical trials involving >1000 patients and several years for recruitment and follow-up. It also can be difficult to develop instruments that capture effects on quality of life parameters such as patient symptoms or functional status. For these reasons, considerable attention has been given to use of surrogate end points in replacement of the proper clinical efficacy measures in phase III trials. Surrogate end points tend to be measures of biologic activity, such as changes in CD4 cell count or in viral RNA in the peripheral blood, in part because such changes usually are large in magnitude and early in occurrence. In addition, natural history information usually is readily available to establish that these measures of biologic activity are strongly correlated with clinical efficacy measures.

Unfortunately, as stated by Fleming and DeMets [2], "a correlate does not a surrogate make." While it is clear that dropping CD4 cell counts and rising levels of virus load occur with advancing disease and increasing risk of symptomatic complications of HIV infection and risk of death, it does not follow that treatment-induced improvements in these biologic activity measures will reliably predict treatment induced changes in clinical efficacy outcomes. Broad experience from HIV/AIDS clinical trials over the past decade has clearly shown that statistically significant changes in CD4 cell counts induced by nucleoside analogues have been unreliable predictors of the longer-term treatment effects on risk of progression to symptomatic AIDS and death [2]. In this way, experience with surrogate markers in AIDS clinical research has matched that of other disease settings. In fact, the history of research in other settings confirms that powerful correlates of clinical efficacy outcomes have been very poor surrogates of true clinical efficacy [1, 2]. Classes of drugs in which surrogate markers have failed to reflect true clinical effect include antiarrhythmics, drugs for improving cardiac output, calcium channel blockers, and many others [2].

In spite of the experiences with CD4 cell count in the HIV/AIDS setting, and the poor record of markers in other disease settings, which collectively establish the unreliability of surrogate end points, there remains considerable support in many HIV/AIDS circles to rely heavily if not exclusively on surrogate end point effects in future phase III clinical trials. For example, in the evaluation of new anti-HIV drugs, the National Task Force for AIDS Drug Development recently endorsed using "more flexible criteria for approval" [3]. In practice, this may mean using reduction in plasma HIV RNA levels as the primary end point in pivotal clinical trials. Indeed, analyses such as the one that demonstrated correlation between changes in virus burden and rates of clinical progression in a clinical study of delavirdine [4] have been used to support the use of measures of virus burden rather than clinical progression as the study end point in evaluations of new therapies. This use of surrogates

Received 13 September 1996.
Grant support: NIH (AI-28076, AI-29168, AI-27664, AI-30731).
Reprints or correspondence: Dr. V. De Gruttola, Dept. of Biostatistics, Harvard School of Public Health, Boston, MA 02115.

The Journal of Infectious Diseases 1997; 175:237–46

0022–1899/97/7502–0001$01.00

is quite reminiscent of using reduction in tumor volume as a surrogate end point in cancer intervention trials, another example of a surrogate that has provided misleading information about treatment effects on survival and quality of life.

Here we demonstrate why surrogate end points can be unreliable, explore some approaches to validation of surrogates, use simulations and data from HIV/AIDS trials to illustrate the importance of variability in evaluating the reliability of surrogates and provide practical recommendations about how information from surrogate end points can be most effectively used in an HIV/AIDS treatment evaluation program.

Some Approaches to Validation of Surrogates

Determining the reliability of a surrogate end point, and especially establishing its validity, is a difficult task. Valuable insights about validity can be provided by empiric evidence from an array of clinical trials documenting treatment effects on both surrogate and clinical efficacy end points, as well as by a thorough biologic understanding about mechanisms of treatment effect.

Treatment interventions are likely to have many mechanisms of action. This is especially true of drugs or biologics that carry risk of adverse effects in settings of life-threatening diseases. An intervention's effects on a true clinical efficacy outcome such as death can be mediated either through intended effects on the primary HIV disease process or through an array of unintended mechanisms of action. The reliability of a surrogate end point will be compromised if the intended effects are not fully captured by the surrogate, either because of "noisy" or missing data or because surrogates such as CD4 cell counts or virus load in the peripheral blood only partially reflect the impact of treatment on the primary HIV disease process. The array of unintended mechanisms of action provide even greater challenges to reliability of surrogate end points, since the surrogate undoubtedly would not be in the causal pathway of the effects of such mechanisms on the true clinical end point. The setting of lipid-lowering agents clearly illustrates existence and impact of these unintended mechanisms. In a comprehensive overview of 50 randomized trials of cholesterol-lowering agents [5], an average reduction in cholesterol of 10% was achieved along with the intended 9% reduction in coronary heart disease mortality. However, overall mortality was unchanged due to an unintended 24% increase in non–coronary heart disease mortality.

Unfortunately, as in this coronary heart disease setting with lipid-lowering agents, the biologic complexity of the disease process and of the array of potential effects of an intervention make it unrealistic to presume that all influential mechanisms of action can be anticipated in advance, much less biologically explained. While basic science can provide some insights into the plausibility of various mechanisms by which an intervention can affect outcome, empiric evidence is also of importance in validating a surrogate.

Prentice [6] provides a definition of a valid surrogate and gives two sufficient conditions that jointly ensure this validity, thereby providing guidance for how one might approach using empiric evidence to assess validation. By his definition, a surrogate is valid if "a test of the null hypothesis of no relationship (of the surrogate end point) to the treatment groups must also be a valid test of the corresponding null hypotheses based on the true end point." Prentice's first condition to ensure this validity is the "correlate" requirement, that is, a valid surrogate end point must be correlated with the true clinical end point. This condition usually holds since, in practice, potential surrogates are often selected by searching for measures that are strongly correlated with clinical efficacy end points. Prentice's derivation adds a very restrictive second condition that requires the surrogate to fully capture the treatment's "net effect" on the true clinical end point, the net effect being the aggregate effect accounting for all mechanisms of action. This means that if one has the appropriate marker values for a patient, knowledge of treatment provides no additional prognostic information. The restrictiveness of this condition provides important insight into why correlates are rarely valid surrogates.

In applications, extensive analyses have been performed to assess surrogacy of CD4 cell count, using data from several large clinical trials evaluating nucleoside analogues in HIV/AIDS patients. While these analyses consistently show that CD4 cell count is a correlate of the "progression to symptomatic AIDS or death" end point, thereby satisfying Prentice's first condition, CD4 cell counts have not been established to be a valid surrogate end point since the second condition of Prentice consistently fails to hold [7–11].

The validity of Prentice's restrictive second condition, requiring a surrogate to capture fully the net effect of an intervention on the clinical efficacy end point, can be directly assessed using such methods as those described by Tsiatis et al. [10]. Freedman et al. [12] also discuss statistical issues related to investigation of markers in an epidemiologic setting.

Choi et al. [11] and O'Brien et al. [13] consider the validity of surrogates in clinical trials of antiretroviral drugs by computing the proportion of the net treatment effect captured by the marker. It should be recognized, however, that it is not possible to determine the proportion, p, of the treatment effect on the primary HIV disease process that is accounted for by effects on a surrogate end point. To demonstrate how this nonidentifiability arises, we consider a simple example, illustrated in figure 1, in which we assume the clinical end point, or "failure," is death. Suppose that, on the standard-care control regimen, the death rate per 100 person-years induced by those mechanisms of HIV targeted by the treatment is $\mu_h = 10$, while the death rate due to other causes (including those influenced by unintended mechanisms of the drug) is $\mu_o = 1$. Suppose further that the experimental intervention reduces the death rate in-

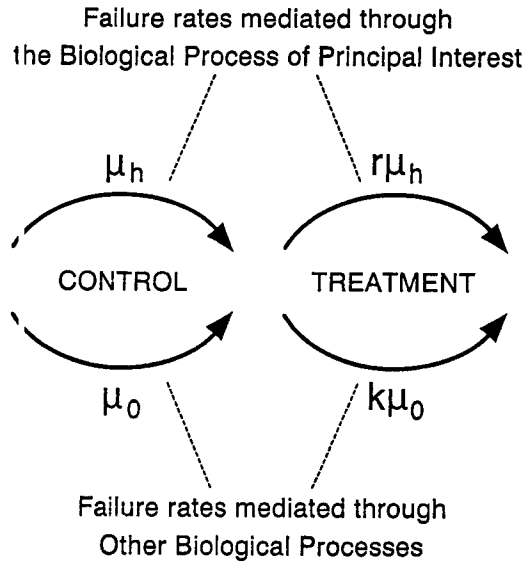

Figure 1. Illustration of the multiple mechanisms of drug action. Failure rates by cause in patients receiving control regimen (μ_h and μ_o) and treatment regimen ($r\mu_h$ and $r\mu_o$).

duced by targeted HIV mechanisms by the multiplicative factor $r = 1/2$, to $r\mu_h = 5$ per 100 years, but increases the death rate due to other causes by the multiplicative factor k, to $k\mu_o = k(1) =$ per 100 years. Consider the case in which the surrogate end point captures only 20% (i.e., $p = .2$) of the intervention's effect on reducing the death rate induced by targeted HIV mechanisms. Then the treatment-induced change in the surrogate end point reduces the death rate by a multiplicative factor

$$r_s = \frac{\mu_h - p(\mu_h - r\mu_h)}{\mu_h} = \frac{10 - .2(5)}{10} = 0.9. \quad (1)$$

To explain the origin of this formula, we note that the change in death rate induced by the intervention's effect on targeted HIV mechanisms is $(\mu_h - r\mu_h)$. Multiplying this quantity by p yields the proportion of this effect captured by the surrogate. Therefore, if the amount of treatment benefit was only that portion captured by the surrogate, the death rate would be the numerator of equation 1. Dividing by μ_h gives us the multiplicative factor, r_s.

The parameter r_s can be measured using data from the study itself or from natural history data bases that allow modeling the association of death with surrogate end points such as CD4 cell count or plasma HIV RNA levels. One can also measure the observed overall net effect of the intervention on death rate,

$$r_o = \frac{r\mu_h + k\mu_o}{\mu_h + \mu_o} \quad (2)$$

and, in turn, can compute the observed portion of the net effect accounted for by the treatment-induced change in the surrogate end point,

$$p_o = \frac{1 - r_s}{1 - r_o}. \quad (3)$$

Equations 1–3 reveal that $p_o = p$ if either $\mu_o = 0$ (i.e., death can be caused only by the targeted HIV mechanisms) or $r = k$ (i.e., the intervention has the same effect on targeted HIV mechanisms and on other causes of death). However, in the more common setting, in which $\mu_o > 0$ and $r < k$, figure 2 provides the surprising insight that even when the true portion of treatment effect on the HIV-related death rate that is accounted for by effects on the surrogate is only $p = .2$, the observed proportion p_o rises to .37 if k is 3 and approaches unity as k approaches 5. This example illustrates the way in which surrogate end points that capture only a small fraction of the change in the death rate induced by treatment effects on targeted HIV mechanisms may appear to capture an observed portion, p_o, near unity, simply due to unanticipated and unrecognized harmful effects on the other causes of death.

A similar phenomenon occurs when a net treatment effect observed in a placebo-controlled study is reduced because patients assigned initially to placebo are offered the active treatment after evidence of clinical deterioration but before death or onset of other clinical end points. Deferred treatment will tend to reduce the difference in treatments actually received by patients in the two study arms and, hence, reduce the net treatment effect. This situation arose in an analysis of surrogacy using results of a study of immediate versus deferred use of zidovudine [12]. The analysis examined the degree to which plasma HIV RNA and CD4 lymphocyte responses after 6 months of treatment captured the effect of zidovudine. Because no more than a few patients on the deferred arm had received zidovudine by month 6, the plasma HIV RNA responses essentially reflect the effects of zidovudine and of placebo. The net treatment effect, r_o, used in the estimation of p_o, however, compares immediate zidovudine to deferred zidovudine. The sooner after 6 months that patients were offered zidovudine, the smaller will be the net treatment effect. In fact, all patients on the deferred arm were offered zidovudine at some point in the follow-up used in this analysis. Once again, the smaller the net treatment effect (this time reduced by both delayed benefit on the deferred arm as well as by cumulative effect of toxicities), the larger will be p_o. Had zidovudine never been offered to patients on the deferred arm, the marker changes would be the same; the net treatment differences, larger; and the p_o, smaller, perhaps considerably so. Thus, p_o can be inflated arbitrarily depending on design features of the study. Similar problems of interpretation arise whenever there is noncompliance

	k			
	1/2	1	3	4.9
Observed intervention effect on risk: $r_o = \frac{r\mu_h + k\mu_o}{\mu_h + \mu_o} = \frac{.5(10) + k}{10 + 1}$	.50	.55	.73	.90
Observed proportion of net treatment effect captured by surrogate: $p_o = \frac{1 - r_a}{1 - r_o} = \frac{1 - 0.9}{1 - r_o}$	.20	.22	.37	1.00

Figure 2. Proportion of net treatment effect captured by surrogate with increasing effects of treatment on death from other causes.

with study medication. If the markers are surrogate, then any modification of study treatment (whether by design or by failure of compliance) must be reflected in the marker if it affects clinical progression.

The Importance of Variability in Validation of Surrogates

Problems of interpretation of p_o are compounded by problems of estimation. Freedman et al. [12] used linear logistic regression models to study the proportion of treatment effect accounted for by the marker, while Choi et al. [11] and O'Brien et al. [13] considered proportional hazards models for failure time data. Here, we focus on the failure time end point. Specifically, as shown in the Appendix, we assume a proportional hazards model for the effects of both treatment and marker. As an approximation of the proportion of net effect on outcome explained by effect of marker, given in equation 9 in the appendix, Freedman et al. [12] propose the following metric

$$p_o^* = 1 - \frac{\beta_a}{\beta}, \tag{4}$$

where β refers to the net treatment effect (natural logarithm of the hazard ratio) and β_a refers to the treatment effects after adjustment for the marker (the unexplained portion of the treatment effect). Because this quantity differs slightly from p_o, we denote it with an asterisk (see Appendix).

As we will demonstrate, the estimation of p_o^* is associated with a high degree of variability in the settings that characterize most clinical trials. Let $\hat{\beta}$ and $\hat{\beta}_a$ denote the estimates of β and β_a. Then p_o^* is estimated by

$$\hat{p}_o^* = 1 - \frac{\hat{\beta}_a}{\hat{\beta}}. \tag{5}$$

Lin et al. [14] developed a closed form solution for the large sample variance of $\hat{p}_o^*$. The factors that determine this variance include the coefficient of variation for β (i.e., the inverse of the unadjusted treatment effect relative to its SE), the value of p_o^* itself, and the values of the variances and covariance of the adjusted and unadjusted treatment effects, β_a and β. When the marker has only a small effect on risk of progression and the correlation between treatment and marker is low (this implies the marker may not be very clinically useful), then the SE is approximately

$$\sigma \approx |p_o^*| \frac{\mathrm{SE}(\hat{\beta})}{|\beta|}, \tag{6}$$

where $|.|$ refers to the absolute value. Suppose that we have a large unadjusted treatment effect that is four times its SE, that is, $\beta/\mathrm{SE}(\hat{\beta}) = 4$, then equation 6 implies that the mean width of the 95% confidence interval (CI) for p_o^* is equal to p_o^* itself. In practice, formula 6 tends to underestimate the true variability of $\hat{p}_o^*$, so this estimate should be seen as a lower bound (see Appendix). In fact, the estimate $\hat{\beta}_a$ becomes increasingly unstable as the correlation between treatment and marker increases. (An extreme scenario occurs when, in a placebo-controlled trial of a treatment, all treated and no untreated patients have a marker response.) Thus, an unadjusted treatment effect that is >4 times its SE is a necessary, though insufficient, condition for a precise estimation of p_o^*. Freedman et al. [12] observed that the unadjusted treatment effect of >4 times its SE yields a high probability of reaching a conclusion that the surrogate marker explains at least 50% of the net treatment effect on the clinical outcome under the hypothesis that it explains all of the net treatment effect. Their numerical results, however, were overly optimistic because they implicitly assumed that the variances of the adjusted and unadjusted treatment effects were the same.

To demonstrate the uncertainty in the estimation of p_o^* under the assumption of proportional hazards, we did a simulation in a setting comparable to that found when investigating new AIDS treatments. In our simulation, the data sets were created to have unadjusted treatment effects of ~2 and 4 SEs, which will be referred to as cases a and b, respectively. We considered

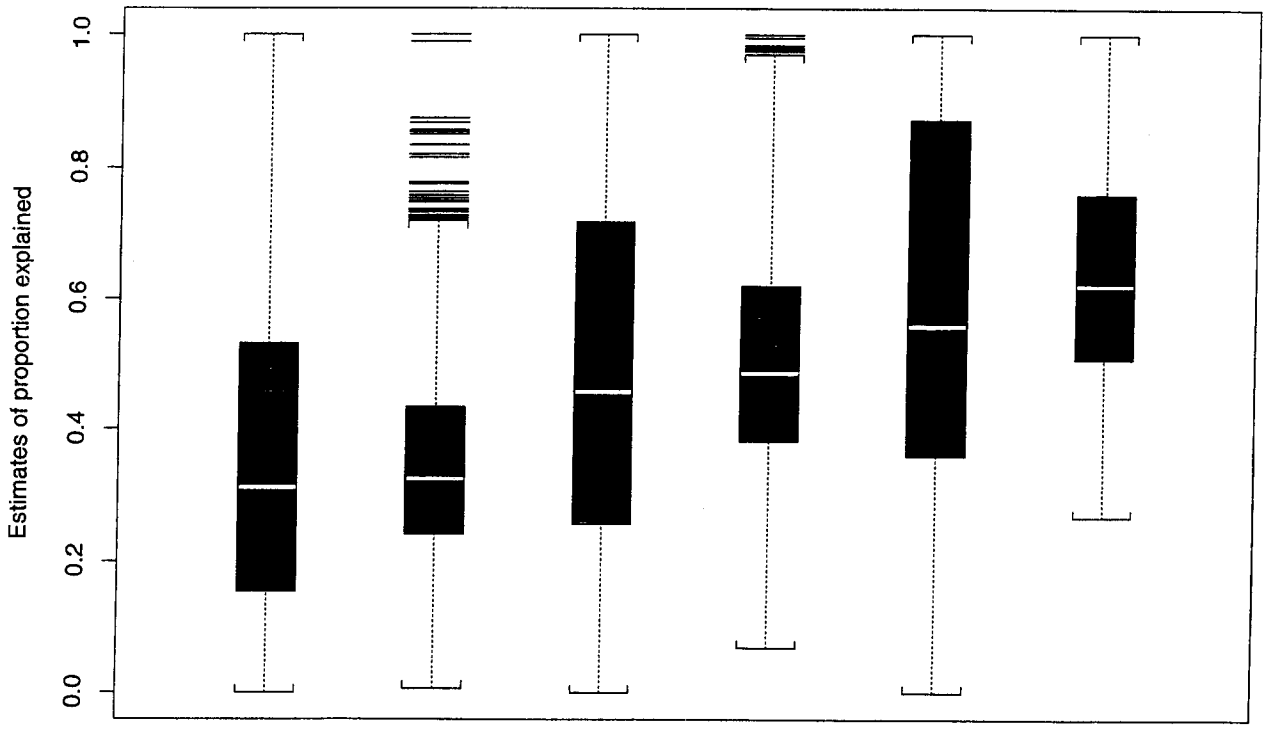

Figure 3. Estimates of proportion of treatment effect explained by markers assuming treatment effects of 2 SEs (bars 1, 3, and 5) and of 4 SEs (bars 2, 4, and 6). Actual proportion of treatment effect explained is $\frac{1}{3}$ (bars 1 and 2), $\frac{1}{2}$ (bars 3 and 4), and $\frac{2}{3}$ (bars 5 and 6). White lines display medians; black rectangles, interquartile spreads; and dotted lines, ranges of values within 1.5 times interquartile spread of median.

p_o^* of $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{2}{3}$. For details of the simulation, see the Appendix. The results based on 1000 simulation samples are displayed in figure 3 (and in the Appendix, table 2). Figure 3 displays the medians (white horizontal line) and interquartile spread (black rectangle) for each of the simulations. The interquartile spread for cases b (second, fourth, and sixth bars) is about half that for cases a (first, third, and fifth bars). The dotted lines extend out either to a distance from the median of 1.5 times the interquartile spread or to values of 0 or 1 (whichever is closer to the median); outliers beyond these values are shown individually if they are between 0 and 1. These plots show the high degree of variability in $\hat{p}_o^*$, especially for cases a. The formula of Lin et al. [14] allows us to calculate a 95% CI for each simulation sample. As shown in the Appendix, the median width of the 95% CI is quite wide (between .5 and .75), even in cases b, and always exceeds 1 for cases a. Additional factors make precise estimation of p_o^* even more difficult, specifically the following: Markers are measured with error and often have missing values, treatment effects may be transient, and the relationship between marker and risk of disease may be complex and vary over time and disease stage, while analyses are often restricted to initial response. Tsiatis et al. [10] consider approaches to handle such problems, but these methods require a model for the process that results in missing marker values.

Choi et al. [11] considered the proportion of treatment effect of zidovudine accounted for by different CD4 cell measures among patients who had not progressed by week 16 in a placebo-controlled trial of zidovudine in asymptomatic patients (AIDS Clinical Trials Group [ACTG] 019, CD4 cell counts $<500/mm^3$). Among such patients, the placebo-zidovudine relative risk for clinical progression was 1.48. After adjusting for baseline CD4 cell count, as did Choi et al., this increased to 1.70 ($P = .11$), since the higher rate of failure among placebo recipients in the first 16 weeks leads to higher average baseline CD4 cell counts among remaining placebo recipients compared with zidovudine recipients. Using imputation of values for patients missing 16-week CD4 cell measurements, Choi et al. estimated that the proportion of treatment effect explained, p_o^*, was .46 for CD4 cell count and .74 for net CD4 cell percent, defined to be the product of the percentage of white blood cells that are lymphocytes and the percentage of lymphocytes that are CD4-positive. However, these estimates of p^* are slightly biased toward larger values, since Choi et al. did not include baseline CD4 cell counts in the regression models that yield the marker-adjusted treatment effects, β_a. When the relevant baseline marker values are included in estimation of β_a, as they were in estimation of β, estimates of p^* are reduced to .42 and .69. To examine the effect of the imputation, we repeated the Choi et al. analy-

sis using only the data available at week 16. This analysis produced a p^* of .50 for CD4 cell count and 1.0 for net CD4 cell percent—identical to the results of Choi et al. when they excluded patients with missing week 16 values. The analysis that consistently adjusts for baseline marker values in estimation of β_a as well as β produced p^* values of .43 and .94. In the other studies we consider below, analyses will be done without imputation and will properly adjust for the same baseline covariates when estimating β and β_a.

If it were reliably established that net CD4 cell percent accounted for more than half of the treatment benefit of zidovudine, this result could be a useful contribution to understanding mechanisms of pathogenesis and drug action. Unfortunately, as the preceding simulation has shown, the typically large variability in estimation of p_o^* seriously undermines this reliability. Direct insight into the variability of these estimates of p_o^* is provided by using the formulas for the variance of $\hat{p}_o^*$ and 95% CIs recently derived by Lin et al. [14]. The Choi et al. [11] estimate of p^* for CD4 cell count, .46, has an SE of 0.31 (95% CI: − 15, 1.07); their estimate of p^* for net CD4 cell percent, .74, has an SE of 0.47 (95% CI, −.18, 1.66). CIs for the estimates that consistently adjust for baseline marker values are similarly broad.

Further insight into the variability and unreliability of results by Choi et al. is provided by repeating their analyses in two other zidovudine trials, ACTG 016 and 116b/117. ACTG 016 was a placebo-controlled study of zidovudine among 711 patients with mildly symptomatic disease [15]. This study was chosen because of its similarity to ACTG 019; although the ACTG 016 population had mild symptoms, there was no constraint on the CD4 cell counts, so the median baseline CD4 cell counts for this study and the lower CD4 cell count stratum ($<500/mm^3$) of ACTG 019 were fairly similar. Our analyses examined the amount of treatment effect accounted for by CD4 cell count and net CD4 cell percent among 605 patients who had not progressed clinically by week 16 and had a CD4 cell measurement at week 16. Among these patients, 45 reached a clinical end point by the end of the study. There was a larger treatment effect among such patients in ACTG 016 than in ACTG 019—the placebo-zidovudine relative risk was 3.33 (P = .0004)—but the p^* was only .08 for CD4 cell count and .19 for net CD4 cell percent. These values are low even though, for both markers, baseline values and changes from baseline to week 16 are all predictive of the clinical end point.

The other ACTG study we analyzed was ACTG 116b/117, which compared zidovudine to didanosine in 913 patients who had at least 4 months of experience with zidovudine [16]. Of these patients, 105 had an end point before study week 16; 641 of the remaining 808 had the necessary week 16 CD4 cell counts. Among these patients, 185 reached a clinical end point after week 16 (providing greater power for our analyses than the previously mentioned studies). The unadjusted treatment effect was 1.53 (P = .005). The estimated values of p_o^* were .35 for CD4 cell count (95% CI: 0.035, 0.674) and .08 for net CD4 cell percent (95% CI: −0.174, 0.327). Both markers, baseline values and changes between baseline and week 16, are highly predictive of clinical progression, at a significance level of $P < .0001$! Thus, the high value of p_o^* estimated by Choi et al. is not consistently seen in other studies, even when marker values are strongly predictive of clinical progression. This high value may well reflect the multiplicity of mechanisms of action of zidovudine (both favorable and unfavorable) and random variation, rather than the actual mechanism of drug action.

Analyses of surrogacy of virus load measured by plasma HIV RNA have become a major focus of interest. To examine this issue, we grouped all patients from two studies comparing zidovudine with didanosine who had any prior zidovudine experience. These studies were ACTG 116b/117, described above, and ACTG 116a, a similar study for patients with ≤16 weeks of prior zidovudine experience. This data pooling from two studies was important for acquiring an adequate amount of information for analyses, since virus load was available for only a subset of patients. Results of analyses of virus load, performed separately, are presented elsewhere [17, 18]. Table 1 shows the surrogacy analyses for the Chiron branched DNA and Roche RNA polymerase chain reaction assays. The first row of the left portion of the table gives the risk ratios and probability values associated with didanosine treatment, adjusted for baseline value of plasma HIV RNA and CD4 lymphocyte counts. The risk ratios for the baseline marker values, displayed in the second and third row of the left portion of the table, show the effect on risk of every 10-fold increase in RNA or 2-fold increase in CD4 cell count. The right portion of the table shows the risk ratios associated with didanosine treatment and baseline marker values, after adjustment for the change in RNA at week 8. The risk ratio associated with the RNA change is given in the last row of the table. The latter displays the effect on risk of a 10-fold increase in RNA change between baseline and week 8. Note that all baseline values and changes from baseline are significantly associated with clinical progression. From these analyses, p_o^* is estimated to be .36 for the Chiron branched DNA assay and .26 for the Roche RNA polymerase chain reaction. The corresponding 95% CIs are −0.26, 0.99 and −0.24, 0.75. There is no evidence that any of these markers is superior to any other or provides a valid surrogate for the clinical end point.

O'Brien et al. [13] provide estimates of and CIs for the p_o^* relating to 6-month changes in plasma HIV RNA and CD4 lymphocyte count observed in a study of immediate versus deferred zidovudine. They do not describe their method for obtaining CIs, other than alluding to a "bootstrap" procedure. Their CIs, like ours, have widths of ~1 or larger; the treatment effect in their study has a somewhat smaller, though still marginally significant, P of .03.

Compared with the analyses of ACTG 116/117, their point estimates referring to plasma HIV RNA alone, 0.59 (95% CI:

Table 1. Analysis of patients from AIDS Clinical Trials Group 116/117 with baseline and one follow-up RNA measurement before week 8 by Chiron branched DNA assay (n = 126) or Roche RNA polymerase chain reaction (n = 120).

	Baseline		Week 8	
Assay, parameter	RR (CI)	P	RR (CI)	P
Chiron branched DNA				
Didanosine treatment	0.69 (0.42, 1.15)	.15	0.79 (0.48, 1.32)	.37
RNA*	1.77 (1.18, 2.67)	.006	2.47 (1.57, 3.89)	.0001
CD4 cell count*	0.76 (0.67, 0.86)	.0001	0.75 (0.66, 0.86)	.0001
ΔRNA†	—		2.96 (1.53, 5.73)	.001
Roche RNA PCR				
Didanosine treatment	0.69 (0.41, 1.17)	.17	0.76 (0.45, 1.3)	.32
RNA*	1.48 (0.99, 2.21)	.006	1.65 (1.09, 2.51)	.02
CD4 cell count*	0.76 (0.65, 0.84)	.0001	0.75 (0.66, 0.86)	.0001
ΔRNA†	—		2.96 (0.92, 3.39)	.08

NOTE. RR, risk ratio; CI, 95% confidence interval; PCR, polymerase chain reaction.
* Effect on risk of every 10-fold increase in RNA or 2-fold increase in CD4 cell count.
† Effect on risk of 10-fold increase in RNA change (ΔRNA) between baseline and week 8.

0.13, 1.12), and to RNA combined with CD4, 0.79 (95% CI: 0.27, .45) are somewhat higher. Nonetheless, the relative magnitudes of these quantities are rendered uninterpretable by the use of deferred zidovudine as described above (Approaches to Validation of Surrogates). In addition, the wide CIs imply too much uncertainty in the estimates of p_o^* for meaningful comparison.

Conclusions

These analyses of data from ACTG and other studies demonstrate what the underlying theory implies: It is very difficult to quantify precisely the amount of a treatment's beneficial effects that are captured by a marker. A direct consequence of this difficulty is that one cannot reliably infer that treatments that have a beneficial effect on markers will also have a commensurate clinical effect—especially if the effect of treatment on markers is only observed in the short term. In order for a marker to be a valid surrogate by the Prentice definition [6], it must capture all of a treatment's beneficial and harmful effects. This means, for example, that if mutations arise that affect the virulence or susceptibility to treatment of a virus over time, the clinical effects of these mutations must be reflected in the patient's marker level. Similarly, effects of noncompliance or treatment discontinuation must also be reflected in the marker. Markers that truly capture all of a treatment's effects have never been found. While "partial surrogate markers" that capture some of a treatment's effect may provide insight into biologic mechanisms, analyses of the degree of surrogacy must be regarded with caution. First, the quantity "proportion of treatment effect explained by a marker" is not strictly identifiable unless all of the mechanisms of drug action are understood and can be accurately measured. Even in this unlikely case, a large number of events are required to measure this quantity accurately.

Given these difficulties, it is not realistic to expect that a statistical analysis can be used to prove that a marker is a valid surrogate according to the Prentice definition. For a marker to be a reliable predictor of clinical benefit from a treatment, it is required that the biology of disease and all important effects of this intervention (including adverse effects) be fundamentally understood. For example, to determine, without waiting for onset of clinical symptoms, that a treatment is effective against an infectious disease when it causes the etiologic agent to become undetectable in blood, it must be known that (1) a quantity of agent sufficient to cause disease (threshold for disease) will always be detectable, (2) the effect of treatment is not transient, and (3) the treatment does not have unintended mechanisms of action that could adversely affect patient well-being or survival.

When disease mechanisms are well understood, laboratory assays are useful both in clinical research and in patient management. An example is the treatment of culture-positive bacterial endocarditis. Clinicians believe that successful eradication of the etiologic agent is essentially demonstrated when cultures that were positive before treatment become negative after a prolonged course of antibiotic therapy; this belief implies that conditions (1) and (2) are met. However, neither drug toxicities nor clinical consequences of infection to the heart values are reflected by the culture results per se. The undeniable value of performing cultures for patient management in this setting comes not from a belief that the cultures are perfect surrogates for (or perfect predictors of) clinical outcomes but from an understanding of the pathogenesis of the disease and the mechanisms of antibiotic action. If a negative culture assures that the etiologic agent is eradicated, then the benefit of two different

drugs can be compared by assessing the proportion of patients receiving these drugs whose cultures become negative. Nonetheless, this comparison may not completely characterize the usefulness of different antibiotics if they differ in toxicity or can have unintended effects on clinical status (e.g., aminoglycoside nephrotoxicity). Furthermore, if cultures are not always reliably performed, or if residual infection could reside in compartments that are not reflected by standard blood culture technique, then cultures might not provide adequate information about the ability of comparative treatments to eradicate the etiologic agent. Thus, appropriate use of markers in patient management and clinical research must be based on an understanding of the underlying biology; estimation of metrics such as p_o are useful for testing hypotheses about mechanism, not simply for replacing clinical end points.

Although drugs that eradicate HIV may not arise soon, analyses of markers are nonetheless important. Markers, singly or in combination, can tell us about the biologic activity of drugs and biologics and thereby help to screen useful interventions for further research. Although high values of p_o^* are not proof that a marker is a valid surrogate, reliably estimated low values are evidence that the marker is not and therefore may not be the best screening tool. Equation 6 provides insight about the size of studies that are required for such reliable estimation; the studies described in this report were generally large enough to show significant treatment differences but not to estimate p_o^* with adequate precision.

An important use of markers is the guidance of clinical care of individual patients. While markers of HIV-1 infection undoubtedly provide information about disease progression and drug effects, the best way to use these markers is not apparent. If it were true that a marker captured all of the effects of a treatment (current and future), the pursuit of the optimal drug use strategy might be simplified to identifying the approach that maintains plasma HIV-1 RNA at the lowest possible levels. In fact, the uncertainty in the proportion of treatment effect explained and the reasons for believing that it differs from 100% suggest that it is important to conduct randomized trials to evaluate the best use of markers in making treatment decisions. In addition to statistical arguments, there is scientific evidence that a number of factors may play important roles in disease progression independently of RNA. The current clinical practice guidelines for the use of plasma HIV-1 RNA measurements to initiate, monitor, and switch antiretroviral therapy fail to account for such factors as the syncytium-inducing phenotype, which is a strong independent predictor of disease progression and is also influenced by therapy [17–21]. As a result, patients who are individually managed by HIV-1 RNA levels and CD4 cell count alone may be subjected to excessive drug toxicities by the quest to drive virus load to minimum levels without consideration of the other factors that are implicated in pathogenesis. In fact, efforts such as the recently designed randomized Community Program for Clinical Research on AIDS study, evaluating the clinical usefulness of providing periodic RNA measurements to physicians and patients, are required to test theories about optimal treatment. Controlled clinical studies of different strategies for managing patients based on virus burden are urgently needed. Analyses of proportion of treatment effect captured by a marker can provide guidance for design of such studies, by indicating the markers and times of measurement that best capture a treatment's net effects.

Analyses of markers have the potential to provide benefits to medical research if the methods of analyses and their limitations are fully understood. They also have the potential to do great harm if they bring about a premature and naive consensus about the effects of inadequately studied drugs and biologics.

Acknowledgment

We are grateful to Laura Smeaton for her efforts on this study.

References

1. Fleming TR. Surrogate markers in AIDS and cancer trials. Stat Med **1994**; 13:1423–35.
2. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? Ann Intern Med **1996**; 125:605–13.
3. National Task Force on AIDS Drug Development: Recommendations of June 29 and 30, **1995**.
4. Wathen LK, Nickens DJ, Chuang-Stein CJ, et al. HIV-1 RNA viral burden at baseline or its reduction following antiretroviral therapy is highly correlated with reduced HIV-1 disease progression [abstract LB8c]. In: Program and abstracts of the 3rd Conference on Retroviruses and Opportunistic Infections. Washington, DC: Infectious Disease Society for the Foundation for Retrovirology and Human Health, **1996**.
5. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute infarction. N Engl J Med **1993**; 329:673–82.
6. Prentice RL. Surrogate end points in clinical trials: definition and operational criteria. Stat Med **1989**; 8:431–40.
7. Lagakos SW, Hoth DF. Surrogate markers in AIDS: where are we? Ann Intern Med **1992**; 116:599–601.
8. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the role of CD4-lymphocyte counts as surrogate end points in human immunodeficiency virus clinical trials. Stat Med **1993**; 12:835–42.
9. De Gruttola V, Wulfsohn M, Fischl M, Tsiatis A. Modeling the relationship between survival and $CD4^+$ lymphocytes in patients with AIDS and AIDS-related complex. J Acquir Immune Defic Syndr **1993**; 6:359–65.
10. Tsiatis AA, De Gruttola 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(429):27–37.
11. Choi S, Lagakos SW, Schooley TT, Volberding PA. $CD4^+$ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. Ann Intern Med **1993**; 118:674–80.
12. Freedman LS, Graubard BL, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med **1992**; 11:167–78.
13. O'Brien W, Hartigan P, Martin D. Changes in plasma HIV-1 RNA and $CD4^+$ lymphocyte counts and the risk of progression to AIDS. N Engl J Med **1996**; 334:426–31.

14. Lin DY, Fleming TR, De Gruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. Stat Med (in press).
15. Fischl MA, Richman DD, Hansen N, et al. The safety and efficacy of zidovudine (AZT) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type I (HIV) infection: a double-blind, placebo-controlled trial. Ann Intern Med **1990**;112:727–37.
16. Kahn JO, Lagakos SW, Richman DD, et al. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. N Engl J Med **1992**;327:581–7.
17. Coombs RW, Welles SL, Hooper C, et al. Association of plasma human immunodeficiency virus type 1 RNA level with risk of clinical progression in patients with advanced infection. J Infect Dis **1996**;174:704–12.
18. Welles SL, Jackson JB, Yen-Lieberman B, et al. Prognostic value of plasma human immunodeficiency virus type 1 (HIV-1) RNA levels in patients with advanced HIV-1 disease and with little or no zidovudine therapy. J Infect Dis **1996**;174:696–703.
19. Zheng NN, McQueen PW, Hurren L, et al. Changes in biologic phenotype of human immunodeficiency virus during treatment of patients with didanosine. J Infect Dis **1996**;173:1092–6.
20. Karlsson A, Parsmyr K, Sandstrom E, Fenyo EM, Albert J. MT-2 cell tropism as prognostic marker for disease progression in human immunodeficiency virus type 1 infection. J Clin Microbiol **1994**;32:364–70.
21. Delforge ML, Liesnard C, Debaisieux L, Tchetcheroff M, Farger CM, Van Vooren JP. In vivo inhibition of syncytium-inducing variants of HIV in patients treated with didanosine. AIDS **1995**;9:89–90.

Appendix

We assume that the failure rate at time t in treatment group z is

$$\lambda(t|z) = \lambda_o(t)e^{\beta z}, \qquad \text{(A1)}$$

where $z = 0$ for control and $z = 1$ for experimental treatment, β is an unknown constant, and $\lambda_o(t)$ is an arbitrary positive function. From equations 2 and A1, the net treatment effect is $r_o = e^{\beta}$. In turn, incorporating the effect of the surrogate $X(t)$ on the failure rate at time t, we have

$$\lambda(t|z, X(t)) = \tilde{\lambda}_o(t)e^{\beta_a z}e^{\alpha X(t)}, \qquad \text{(A2)}$$

where β_a and α are unknown constants, and $\tilde{\lambda}_o$ is an arbitrary positive function which may differ from λ_o.

Strictly speaking, models A1 and A2 cannot hold simultaneously; however, they may hold approximately when either α or λ is small. HIV/AIDS clinical trials are generally terminated well before the majority of the patients have experienced the event of interest; therefore, for most studies, it is reasonable to assume that the event is rare enough for models A1 and A2 to hold approximately during the follow-up period. For simplicity, we will assume that the effects of model misspecification are negligible (see Lin et al. [14] for a rigorous discussion of this issue).

By equations 1, 2, and A2,

$$r_s = e^{(\beta - \beta_a)}. \qquad \text{(A3a)}$$

Thus,

$$p_o = \frac{1 - e^{\beta - \beta_a}}{1 - e^{\beta}}. \qquad \text{(A3b)}$$

Freedman et al. propose as an approximation of the proportion of net effect on outcome explained by effect of marker,

$$p_o^* = 1 - \frac{\beta_a}{\beta}. \qquad \text{(A3c)}$$

The two quantities, p_o and p_o^*, are equivalent when $\beta_a = \beta$ or $\beta_a = 0$ and differ only slightly for intermediate values. Of course, as shown above, the quantities are equal to p only under very special cases. Note that while p, the proportion of the intended effect on targeted HIV processes captured by the marker, is always a proportion in the mathematical sense of lying in the interval [0,1], p_o, the "proportion" of the net

Table 2. Summary of simulation results.

	$p_o^* = 1/3$		$p_o^* = 1/2$		$p_o^* = 2/3$	
	Case a	Case b	Case a	Case b	Case a	Case b
Mean of $\hat{\beta}$	0.77	0.75	1.03	0.98	1.43	1.38
Mean of $\hat{\beta}_a$	0.52	0.50	0.55	0.50	0.55	0.52
Variance of $\hat{\beta}$	0.14	0.04	0.26	0.06	0.43	0.11
Variance of $\hat{\beta}_a$	0.17	0.05	0.32	0.07	0.54	0.13
Covariance of $\hat{\beta}$ and $\hat{\beta}_a$	0.14	0.04	0.26	0.06	0.39	0.11
Mean of $\hat{p}_o$	0.38	0.35	0.65	0.52	0.74	0.66
Median of $\hat{p}_o^*$	0.31	0.32	0.46	0.49	0.58	0.62
SE of $\hat{p}_o^*$	1.19	0.17	1.64	0.20	1.89	0.22
Interquartile range of $\hat{p}_o^*$	0.38	0.19	0.46	0.24	0.52	0.25
95% confidence interval for $\hat{p}_o^*$						
Mean width	7.10	0.64	11.11	0.74	11.36	0.83
Median width	1.11	0.55	1.21	0.64	1.40	0.74

NOTE. Unadjusted treatment effects (β) are 2 and 4 times SEs of $\hat{\beta}$ in cases a and b, respectively.

treatment effect captured by the marker, need not lie in the interval [0,1]. When β_a and β differ in sign (a situation that arises when the marker captures all of the benefit so that only the harmful effect is reflected in β_a), p exceeds 1; when $\beta_a > \beta > 0$, p_o is negative.

This formulation of the problem allows us to develop a large-sample variance for $\hat{p}_o^*$. Let $\hat{\beta}$ and $\hat{\beta}_a$ denote the estimates of β and β_a, obtained by the usual method of maximum partial likelihood. Then p_o^* is estimated by

$$\hat{p}_o^* = 1 - \frac{\hat{\beta}_a}{\hat{\beta}}. \quad \text{(A4)}$$

Lin et al. [14] developed a closed form solution for the large sample variance that showed that, for large samples, $\hat{p}_o^*$ is approximately normal with mean p_o^* and with variance

$$\sigma^2 = \frac{V_\beta}{\beta^2}\left\{\frac{V_{\beta_a}}{V_\beta} + (1 - p_o^*)^2 - 2(1 - p_o^*)\frac{V_{\beta\beta_a}}{V_\beta}\right\}, \quad \text{(A5)}$$

where V_β and V_{β_a} are the variances of $\hat{\beta}$ and $\hat{\beta}_a$, and $V_{\beta\beta_a}$ is their covariance.

The uncertainty in the estimation of p_o^* under models A1 and A2 was explored in the simulation mentioned above (Importance of Variability in Validation of Surrogates). Cases a and b refer to data sets with unadjusted treatment effects of ~2 and 4 SEs, respectively. The marker values were selected randomly from a normal distribution with mean 0 and variance 1 for untreated patients and a normal distribution with mean 1 and variance 1 for the treated. To create the three values of p_o^* (⅓, ½, and ⅔), we generated the data from model A2 with $\beta_a = 0.5$ and $\alpha = 0.25$, 0.5, and 1. The corresponding values of β under model A1 were ~0.75, 1, and 1.5. We set the baseline hazard $\lambda_o(t)$ to be constant and simulated staggered study entry with censoring times that were uniformly distributed from time 0 until the 25th percentile of the (uncensored) failure times, corresponding to 85% censorship. This corresponds roughly to the amount of censorship in studies that have been used as a basis for surrogacy analyses, such as ACTG 019 and ACTG 175, which compared combination nucleoside therapy with monotherapy. The sample sizes were 250, 160, and 90 for $p_o^* = ⅓$, ½, and ⅔, respectively, in cases a and were quadrupled for cases b. The results based on 1000 simulation samples are displayed in figure 3 (described in the text) and in table 2.

Table 2 shows that the biases of $\hat{\beta}$ and $\hat{\beta}_a$ are small. The variance of $\hat{\beta}_a$ is greater than that of $\hat{\beta}$, and the covariance between $\hat{\beta}$ and $\hat{\beta}_a$ is roughly equal to the variance of $\hat{\beta}$. The means and medians of $\hat{p}_o^*$ are fairly close to the hypothetical values of ⅓, ½, and ⅔, especially in cases b. In cases a, $\hat{\beta}$ may take values close to 0, which causes $\hat{p}_o^*$ and the associated CIs to take extremely large values. Consequently, the SEs of $\hat{p}_o^*$ and the mean widths of the CIs are quite large in cases a. The estimates $\hat{p}_o^*$ are much more stable in cases b. For the same value of p_o^*, the interquartile range of $\hat{p}_o^*$ and the median width of the CI decrease by about half from case a to case b. The variability of p_o^* and the width of the CI increase with increasing value of p_o^*. The estimates $\hat{p}_o^*$ are highly variable and the 95% CI are quite wide, always exceeding 1 for cases a and .5 for cases b.