7
Validation of Biomarkers as Surrogates for Clinical Endpoints

Marc Buyse
International Drug Development Institute, Cambridge, Massachusetts, U.S.A.

Tony Vangeneugden, and Luc Bijnens
Janssen Research Foundation, Beerse, Belgium

Didier Renard, Tomasz Burzykowski, Helena Geys, and Geert Molenberghs
Limburgs Universitair Centrum, Diepenbeek, Belgium

I. INTRODUCTION

Biomarkers will become important in the clinic over the years to come, for several reasons. First, an increasing number of new drugs will have a well-defined mechanism of action at the molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers. Second, there will be increasing public pressure for new, promising drugs to be approved for marketing as rapidly as possible, and such approval will have to be based on biomarkers rather than on some long-term clinical endpoint. Finally, if the approval process is shortened, there will be a corresponding need for earlier detection of safety signals that could point to toxic problems with new drugs. It is a safe bet, therefore, that the evaluation of tomorrow's drugs will be based primarily on biomarkers, rather than on the longer-term, harder clinical endpoints that have dominated the development of new drugs until now.

Yet, for biomarkers to be acceptable surrogates for clinical endpoints, a number of conditions must be fulfilled. In this chapter, we review these conditions and we discuss some statistical methods that are useful to address the problem of surrogate marker validation. Much of the work laid out here is still in progress. The statistical approach proposed has been developed using data from a range of clinically diverse situations, including age-related macular degeneration [1–3], cardiovascular disease [2], advanced ovarian cancer [3,4], chronic schizophrenia [5,6], and advanced colorectal cancer [1,4,7,8]. It is currently being validated in other situations, including advanced prostate cancer, advanced breast cancer, early colorectal cancer, early breast cancer, and autoimmune deficiency syndrome (AIDS).

In this chapter, we concentrate on one clinical situation, the hormonal treatment of advanced (metastatic) prostate cancer, to illustrate the statistical methods used for, and the difficulties encountered in, the validation of a biomarker (the prostate-specific antigen, PSA, measured over time) as a surrogate for a clinical endpoint (the patient's death). We will avoid, insofar as possible, technical developments that have been published in full detail elsewhere [1–8]. Although some of our observations are specific to this situation, many of our conclusions are of general relevance to the validation of biomarkers as surrogates for clinical endpoints.

II. CONCEPTUAL FOUNDATION

A. Statistical Definitions and Models

Let us first introduce the problem in general terms, and define some notations that will be used throughout this chapter. We are interested in the effect of some experimental treatment on a clinical or "true" endpoint of interest, as well as on a biomarker that could potentially be used as a "surrogate" endpoint (Fig. 1). In general, the experimental treatment is compared to an appropriate control group in randomized clinical trials.

Statistically, interest will focus on the following parameters (Fig. 1): the effect of the experimental treatment upon the biomarker, called α; the effect of the experimental treatment upon the clinical endpoint, called β; and the effect of the surrogate biomarker on the clinical endpoint, called γ. It will be useful to denote the randomized treatment by Z, the potential surrogate biomarker by S, and the true clinical endpoint by T. Strictly speaking, the biomarker can be used as a surrogate for the clinical endpoint for the purposes of evaluating the experimental treatment if, and only if, a treatment effect on S ($\alpha \neq 0$) predicts a treatment effect on T ($\beta \neq 0$), and no treatment effect on S ($\alpha = 0$) predicts no treatment effect on T ($\beta = 0$). This view of surrogacy, which is rooted in the paradigm of hypothesis testing, had led to a formal statistical definition of surrogacy, but not to useful validation criteria [9–12]. Alternatively, the biomarker can be used as a surrogate for the clinical endpoint for the purposes of evaluating the experimental treatment if, and only if, the estimated treatment

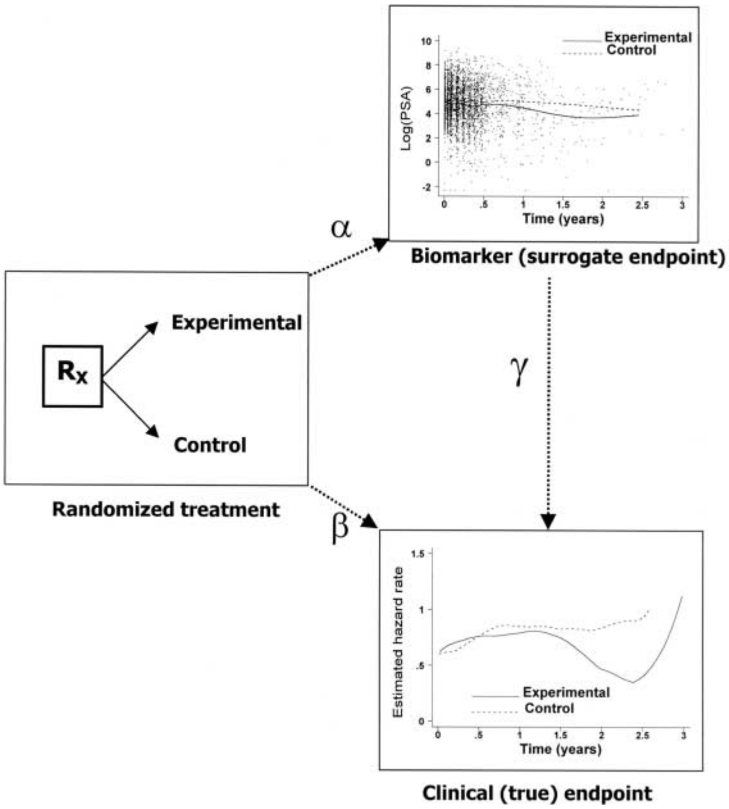

Figure 1 The validation of a biomarker (or intermediate endpoint) as a surrogate for a clinical endpoint (or true endpoint) with respect to the effect of a randomized treatment involves estimating parameters α, β, and γ. In advanced prostate cancer, the biomarker could be the level of PSA over time, and the clinical endpoint could be the time to death. Shown are individual PSA values and their mean over time by treatment group, and the hazard rate over time by treatment group.

effect on S (parameter α) can be used to predict the treatment effect on T (parameter β) with sufficient accuracy [1,3]. This view of surrogacy, which is rooted in the paradigms of estimation and prediction, will be adopted in our analyses of the data in advanced prostate cancer.

Let us first assume the simple, but rare, situation in which the biomarker S and the clinical endpoint T have a bivariate standardized normal distribution. The bivariate normal distribution has been extensively studied, and the statistical techniques required in this situation are straightforward. In reality, the situations will be more complex and will call for less standard models, but the underlying ideas will remain unchanged. If we had data from a single randomized clinical trial with *n* subjects, the relationships between Z, S, and T could be modeled through simple linear regressions:

$$\mathrm{S_i} = \mu_\mathrm{S} + \alpha \mathrm{Z_i} + \varepsilon_\mathrm{Si} \tag{1}$$

$$\mathrm{T_i} = \mu_\mathrm{T} + \beta \mathrm{Z_i} + \varepsilon_\mathrm{Ti} \tag{2}$$

$$\mathrm{T_i} = \mu + \gamma\, \mathrm{S_i} + \varepsilon_\mathrm{i} \tag{3}$$

where μ_S, μ_T, and μ are intercepts; α, β, and γ are the slopes of the regression lines, and also the parameters of interest (Fig. 1); and ε_Si, ε_Ti, and ε_i are normally distributed error terms. The dependence of T upon both Z and S could be modeled through a multiple linear regression:

$$\mathrm{T_i} = \mu' + \beta_\mathrm{S} \mathrm{Z_i} + \gamma_\mathrm{Z} \mathrm{S_i} + \varepsilon_\mathrm{i}' \tag{4}$$

If we had data from several trials, the relationships between Z, S, and T would become:

$$\mathrm{S_{ij}} = \mu_\mathrm{Si} + \alpha_\mathrm{i} \mathrm{Z_{ij}} + \varepsilon_\mathrm{Sij} \tag{1$'$}$$

$$\mathrm{T_{ij}} = \mu_\mathrm{Ti} + \beta_\mathrm{i} \mathrm{Z_{ij}} + \varepsilon_\mathrm{Tij} \tag{2$'$}$$

with notations analogous to those used above, the subscript i now referring to trial and the subscript j to individual patients. In the most general case, a linear mixed model approach could be used, where the intercepts μ_Si and μ_Ti, as well as the slopes α_i and β_i, can be decomposed into fixed and random components [13]. We shall need these models to discuss validation criteria.

B. Types of Biomarkers and Endpoints

Statistically speaking, the biomarker and the clinical endpoint are realizations of random variables. Interest focuses on the joint distribution of these variables, which was assumed bivariate normal in the preceding models. This is, however, seldom the case, because the biomarker and/or the clinical endpoint is often a realization of nonnormally distributed random variables, which can be:

Binary (dichotomous): biomarker value below or above a certain threshold (e.g., CD4+ counts over 500/mm^3) or clinical “success” (e.g., tumor shrinkage)

Categorical (polychotomous): biomarker value falling in successive classes (e.g., cholesterol levels <200 mg/dL, 200–299 mg/dL, 300+ mg/dL) or clinical response (e.g., complete response, partial response, stable disease, progressive disease)

Continuous (normally distributed): biomarker (e.g., log PSA level) or clinical measurement (e.g., diastolic blood pressure)

Censored continuous: time to biomarker below or above a certain threshold (e.g., time to undetectable viral load) or time to clinical event (e.g., time to cardiovascular death)

Longitudinal (repeated measures): biomarker (e.g., CD4+ counts over time) or clinical outcome (e.g., blood pressure over time)

Multivariate longitudinal: several biomarkers (e.g., CD4+ and viral load over time) or several clinical measurements (e.g., dimensions of quality of life over time)

The models used to validate a biomarker as a surrogate for a clinical endpoint will depend on the type of variables observed in the problem at hand. In the example below, we will illustrate this by analyzing the same data in three different ways. The clinical endpoint will be survival in all cases, but the biomarker will consist, respectively, of PSA response (binary variable), time to PSA progression (censored continuous variable), and the PSA pattern over time (longitudinal).

C. Types of Data

To validate the use of biomarkers as surrogates for clinical endpoints, the following information must be available on some series of patients:

Surrogate biomarker or endpoint: most commonly a vector of repeated measurements of the biomarker during the patient's treatment course or follow-up thereafter

Clinical endpoint: most commonly a time (possibly censored) to the clinical event of interest

Treatment: a categorical variable indicating what treatment the patient received (often through randomization)

Unit of analysis: typically a categorical variable indicating the "unit" in which the patient was treated (physician, center, country, trial, meta-analysis, or any other unit defining groups of patients in whom the effect of treatment can meaningfully be estimated)

III. STATISTICAL CRITERIA FOR SURROGACY

The purpose of this section is to provide a brief overview of various statistical ideas that have been proposed for the validation of markers as surrogates for clinical endpoints. In the next section, we will show through an actual example that some of these ideas lead to useful operational criteria.

A. Measures of Association Between the Biomarker and the Clinical Endpoint

Several authors have argued that if a biomarker is to serve as a surrogate for a clinical endpoint, there should be a causal relationship between them [14,15]. If there was a causal pathway from the surrogate marker to the clinical endpoint, then any change in the marker (e.g., as a result of treatment) would translate into a corresponding change in the clinical endpoint. Causality, unfortunately, cannot be tested, and the statistical criteria developed to validate a surrogate marker provide only indirect evidence about the causality of the relationship between the marker and the endpoint.

A first source of evidence is provided by the association, *at the level of the individual patient*, between the marker and the clinical endpoint. One would expect a good surrogate marker to have a strong association with the clinical endpoint at the individual level, reflecting some biological pathway from the biomarker to the clinical endpoint. In that case, the biomarker could be a plausible surrogate on biological grounds, since the clinical endpoint would be largely determined by the biomarker regardless of any treatment effect. This reasoning, although intuitively appealing, has, however, been shown to be potentially misleading, for a good correlate is not automatically a good surrogate [15]. Another source of evidence is needed to quantify the association, *at the level of a trial*, between the effects of a treatment on the marker and on the clinical endpoint. The distinction between these two levels of evidence is essential, but has sometimes been missed in attempts to validate surrogate markers in the past [16]. We return to the trial-level association below.

The individual-level measure of association between the biomarker and the clinical endpoint could be provided by parameter γ_Z in Eq. (4), the slope of the linear regression line between S and T (adjusted for Z), or on a closely related parameter, the squared correlation between S and T (adjusted for Z), which has a more general and intuitive interpretation. The squared correlation (or coefficient of determination) represents the proportion of variance of the clinical endpoint that is explained by the variance of the biomarker, after adjusting for any difference due to treatment. We denote this coefficient $R^2_{\text{individual}}$ to stress that it characterizes the association between the biomarker and the clinical endpoint in individual patients. Just as in linear regression, we will require $R^2_{\text{individual}}$ to be large (close to 1) before we claim that there is a strong association between the biomarker and the clinical endpoint.

For biomarkers and clinical endpoints that are not normally distributed, other measures of association will be used, as will be shown in the analyses below, but the basic idea of a strong association between the biomarker and the clinical endpoint will carry over.

B. Explanation of Clinical Effects from Surrogate Effects

Prentice proposed to define a surrogate marker 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" [9]. As such, this definition is of limited value since direct verification that a triplet {treatment; surrogate biomarker; clinical endpoint} fulfills the definition would require a large number of experiments to be available with information on the triplet. Even if many experiments were available, the equivalence of the statistical tests for the effect of treatment upon the clinical endpoint and the biomarker might not be seen in all of them because of chance fluctuations and/or lack of statistical power. Operational criteria are therefore needed to check if the definition is fulfilled. Prentice proposed four operational criteria:

Treatment must have a significant effect on the biomarker [$\alpha \neq 0$ in Eq. (1)].

Treatment must have a significant effect on the clinical endpoint [$\beta \neq 0$ in Eq. (2)].

The biomarker must have a significant effect on the clinical endpoint [$\gamma \neq 0$ in Eq. (3)].

The *full* effect of treatment on the clinical endpoint must be captured by the biomarker [$\beta_S = 0$ in Eq. (4)].

Even though the prentice criteria were of key importance to help formalize validation approaches, a number of conceptual problems were identified with them. Indeed, it can be shown that Prentice's operational criteria are equivalent to his definition only in the case of binary variables [1]. Moreover, the operational criterion of *full* capture raises a conceptual difficulty in that it requires the statistical test for treatment effect on the true endpoint to be *non*significant after adjustment for the surrogate [11]. Hence this criterion is useful only to reject a poor surrogate biomarker, when the statistical test for treatment effect on the true endpoint remains statistically significant after adjustment for the surrogate. An example of such a situation is given by the effects of zidovudine on clinical endpoints in human-immunodeficiency-virus-positive subjects, which remain significant after CD4+ lymphocytes are taken into account [17,18].

The fourth Prentice criterion cannot be used as such to validate a good surrogate marker, for failing to reject the null hypothesis may be due merely to lack of power. Freedman et al. therefore suggested focusing attention on the proportion of the treatment effect captured by the surrogate, or "proportion explained" [11,19]. In this spirit, a good surrogate is one that explains a large proportion of that effect. Numerically, the proportion explained can be estimated as the ratio $(\beta - \beta_S)/\beta$ from Eqs. (2) and (4). Calculation of its confidence limits requires estimation of the covariance between β and β_S. Several authors have shown that there are fundamental difficulties with the proportion explained, and have proposed alternative approaches [1,12,20].

C. Prediction of Clinical Effects from Surrogate Effects

The reason for using surrogate markers (or surrogate endpoints) is to be able to predict the effect of treatment on the clinical endpoint, having observed its effect on the surrogate marker. This led to consideration of the ratio of the effect of treatment on the clinical endpoint to that on the surrogate marker, or "relative effect" [1]. Numerically, the relative effect can be estimated as the ratio β/α from Eqs. (1) and (2). Calculation of its confidence limits requires estimation of the covariance between β and α. Note that the relative effect depends on the scales chosen to measure S and T. If the relative effect is estimated precisely, then the predicted effect upon the clinical endpoint will in turn be precise enough to be useful. Such a situation requires large numbers of observations that are typically available in large clinical trials, or in meta-analyses of several clinical trials. When meta-analytical data are available, it is also possible to test the assumption implicit in the estimation of the relative effect, i.e., that the treatment effects on the clinical endpoint are proportional to the treatment effects on the surrogate biomarker.

D. Measures of Association Between Treatment Effects

If data are available from multiple sources, for instance if several clinical trials have been performed on the same therapy, it will be possible to estimate the treatment *effects* upon the marker and upon the clinical endpoint in each of these trials [3,21,22]. These treatment effects were denoted α_i and β_i in Eqs. (1′) and (2′). We focus here on the squared correlation (or coefficient of determination) between these treatment effects, which represents the proportion of variance of the treatment effect on the clinical endpoint that is explained by the variance of the treatment effect on the biomarker. We denote this coefficient R^2_{trial} to stress that it characterizes the association between the effects of treatment on the biomarker and on the clinical endpoint in the various trials available. Here again, we will require R^2_{trial} to be large (close to 1) before we claim that there is a strong association between the effects on the biomarker and on the clinical endpoint.

IV. EXAMPLE: PSA IN ADVANCED PROSTATE CANCER

A. The Two Liarozole Trials

We illustrate the statistical approach based on the individual-level and trial-level associations using two trials in patients with advanced (metastatic) prostate cancer. These trials compared oral liarozole, an experimental retinoic acid metabolism-blocking agent developed by the Janssen Research Foundation, with two antiandrogenic drugs: cyproterone acetate (CPA) in the first trial and flutamide in the second. In both trials, patients were in relapse after first-line endocrine therapy [23]. The trials accrued 312 and 284 patients, respectively. Each trial was multinational and multicentric. Since our analyses require the estimation of the effect of treatment in multiple trials or other meaningful groups of patients, we grouped the patients by trial and by country. This allowed us to define 19 groups containing between four and 69 patients per group.

The primary endpoint of the trials was overall survival from the start of treatment. Assessments were undertaken before the start of treatment, at 2 weeks, monthly for 6 months, at 3-month intervals until the second year, and at 6-month intervals until treatment discontinuation or death. The assessments included measurement of the prostate-specific antigen (PSA) level. PSA is a glycoprotein that is found almost exclusively in normal and neoplastic prostate cells. Changes in PSA often antedate changes in bone scan, and they have been used as an indicator of response in patients with androgen-independent prostate cancer [24–26].

We consider, successively, PSA response, time to PSA progression (TPP), and the full longitudinal PSA profile of each patient as potential surrogates for survival in this disease [27].

B. PSA Response as Surrogate for Survival

The best PSA outcome was determined for each patient, and hierarchically ordered as [28]:

Complete response (CR) if the PSA level was at least 20 ng/mL at baseline, returned to normal (<4 ng/mL) at any time, and remained normal for at least 28 days

Partial response (PR) if the PSA level was at least 20 ng/mL at baseline, decreased by at least 50% from the baseline level, and remained under 50% of the baseline level for at least 28 days

No change (NC) if the PSA level was at least 20 ng/mL at baseline, and fluctuated between 50% below and 50% above the baseline level for at least 28 days

Progressive disease (PD) if no other response category applied, and if PSA was at least equal to 10 ng/mL

Not evaluable (NE), if none of the above applied

A patient was defined as having a PSA response if his best PSA outcome was either PR or CR. Hence the biomarker is binary here, and the clinical endpoint is a (possibly censored) survival time.

At the individual level, PSA response was a very strong predictor of survival (Fig. 2a). Because PSA response is binary and survival is censored, the normal theory coefficient of determination (R^2) discussed earlier does not apply, and another measure of association between PSA response and survival is needed. One way to express the impact of PSA response on survival is as follows [8]: consider the odds of surviving beyond time t for PSA responders and for nonresponders; the ratio of these odds is a survival odds ratio. Although the odds of surviving beyond time t decrease in time for both responders and nonresponders, in our model the *ratio* of these odds is assumed constant. This survival odds ratio is equal to 5.5 (95% confidence interval = 2.7–8.2), which means that at any point in time the odds of surviving beyond that time are more than five times higher for patients with a PSA response as compared to patients without such a response. The strong prognostic impact of PSA response can be explained in at least three plausible ways:

PSA response and survival are largely determined by a common set of prognostic factors, so that patients who are likely to have a response are also those who are potentially long survivors.

Patients who survive a long time are more likely to have a PSA response because of length-biased sampling [29].

There is a true causal relationship between the achievement of a PSA response and a prolongation of survival.

The first and second explanations are amenable, at least in part, to statistical investigations, the first through adjustments of the comparison of responders and nonresponders for all known prognostic factors, and the second through a landmark analysis [30]. When these investigations fail to explain a large portion of the prognostic impact of PSA response, then there is indirect evidence that PSA response truly results in a survival improvement [7].

At the group level, the effects of liarozole on PSA response and on survival were poorly correlated, with a coefficient of determination $R^2_{\text{trial}} = 0.05$ (standard error = 0.13) (Fig. 2b).

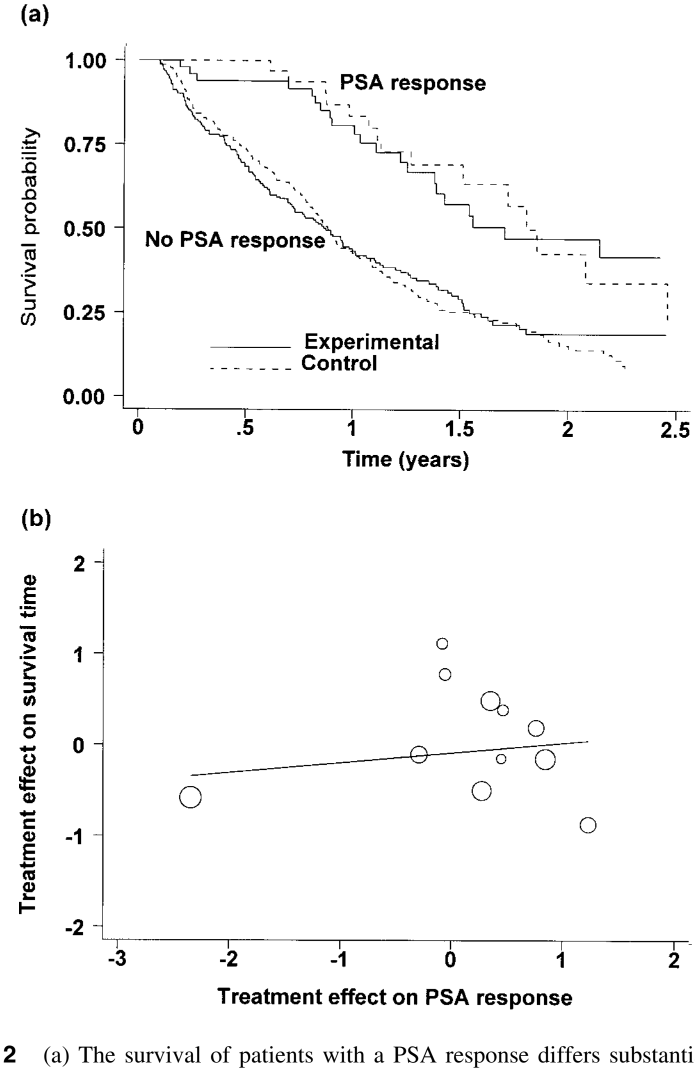

Figure 2 (a) The survival of patients with a PSA response differs substantially from that of patients without a PSA response. At any point in time the odds of surviving beyond that time are more than five times higher for patients with a PSA response as compared to patients without such a response (see text). (b) The treatment effects on survival and on PSA response show no correlation in advanced prostate cancer ($R^2_{\text{trial}} = 0.05$).

There was no overall significant benefit of liarozole over control for either response or survival: the PSA response rate was 16% and 11%, respectively, for liarozole and control ($p = 0.11$), while median survival was 11.3 and 10.9 months, respectively, for liarozole and control ($p = 0.71$).

C. Time to PSA Progression as Surrogate for Survival

The time to PSA progression (TPP) was determined on the basis of a moving average of three consecutive values of PSA. Progression was defined as an increase in PSA equal to, or larger than, 50% above the lowest prior moving average. This increase had to be either the last determination in the patient's follow-up, or maintained for at least 28 days.

At the individual level, PSA progression occurred much earlier than the patients' death. PSA progression occurred within 6 months for half of the patients, while about half of the patients were still alive at 1 year (Fig. 3a). Here again, because TPP and survival may both be censored, the normal theory coefficient of determination (R^2) discussed earlier does not apply, and a possible measure of association between TPP and survival is a generalization of that proposed above [4]: consider the odds of surviving beyond time *t* for patients who have not yet had a PSA progression, and for those who have; the ratio of these odds is a survival odds ratio. Although the odds of surviving beyond time *t* decrease in time for both patients with and without PSA progression, in our model the *ratio* of these odds is assumed constant.

This odds ratio is equal to 6.3 (95% confidence interval = 4.4–8.2), which means that at any point in time the odds of surviving beyond that time are more than six times higher for patients who have not yet had a PSA progression as compared to patients who have already had such a progression. Thus, here again, there is a strong individual-level association between TPP and survival.

At the group level, the effects of liarozole on TPP and on survival were poorly correlated, with a coefficient of determination $R^2_{\text{trial}} = 0.22$ (standard error = 0.18) (Fig. 3b). There was a significant benefit of liarozole over control in terms of time to PSA progression, with a median time of 4.9 months for liarozole and 3.7 months for control ($p = 0.001$).

D. Longitudinal Measurements of PSA as Surrogate for Survival

Since PSA levels were measured repeatedly over time, it seems natural to make use of all these measurements, rather than to define a single PSA response or time to PSA progression for each patient. The Statistical models required to take the longitudinal nature of the measurements into account are more complex, and the analyses potentially more sensitive to model assumptions, than for singly measured

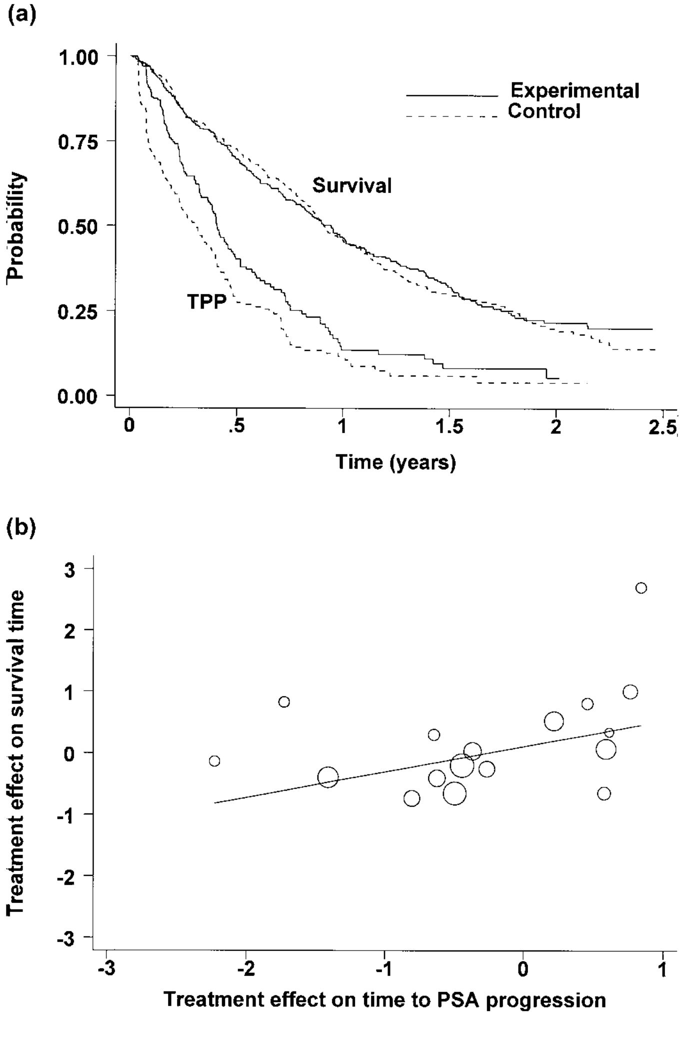

Figure 3 (a) PSA progression is a strong predictor of death in advanced prostate cancer. At any point in time the odds of surviving beyond that time are more than six times higher for patients who have not yet had a PSA progression as compared to patients who have already had such a progression (see text). (b) The treatment effects on survival and on time to PSA progression show very little correlation in advanced prostate cancer ($R^2_{\text{trial}} = 0.22$).

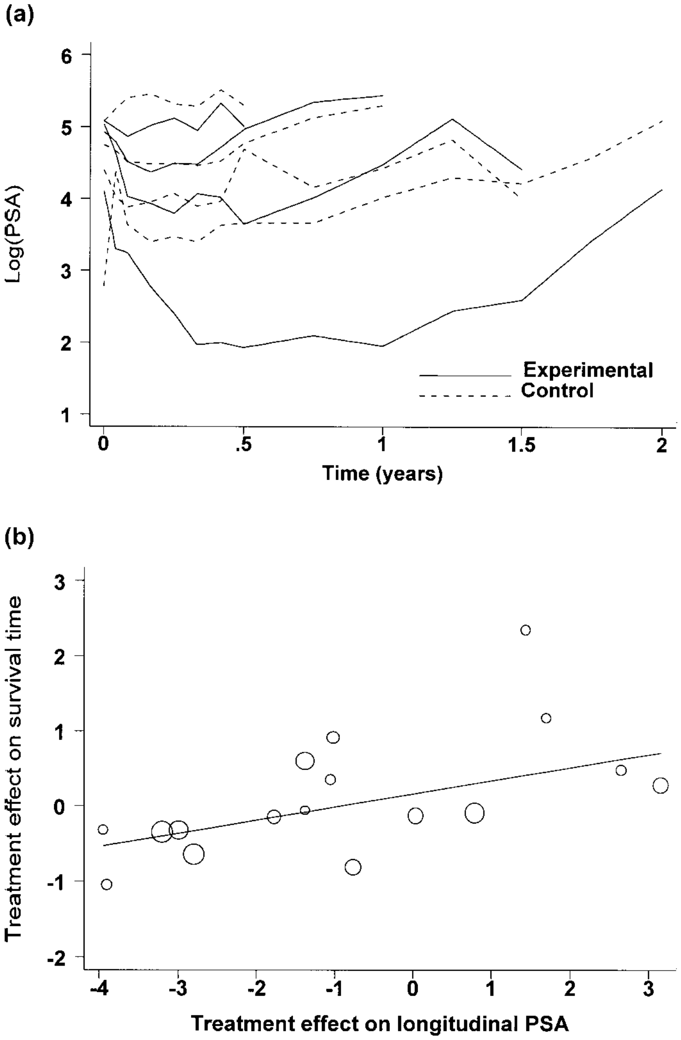

Figure 4 (a) The mean PSA profiles for cohorts of patients with similar follow-up times show a tendency for PSA to go down initially (PSA response), and to come up again after a while (PSA progression). The longitudinal PSA profiles are strongly correlated with the hazard of death ($R^2_{\text{individual}} > 0.84$ at any point in time). (b) The treatment effects on survival and on longitudinal PSA show a weak correlation in advanced prostate cancer ($R^2_{\text{trial}} = 0.42$).

endpoints. Such models have been used extensively to study the relationship between CD4 lymphocytes and survival in patients with AIDS and AIDS-related complex [31–35].

In our example, the mean PSA levels over time shown in the upper-right-hand panel of Fig. 1 are not fully informative, because these means were not calculated on the same patients over time. Indeed, patients who had a PSA progression left the study, and no longer contributed to the mean PSA after that time point, thus creating a selection bias in the calculation of the mean. A more informative way of looking at mean PSA levels over time is to consider cohorts of patients defined by the time they leave the study (for any reason). Figure 4a shows four such cohorts, split by treatment group: patients leaving the study within 6 months, between 6 and 12 months, between 12 and 18 months, and between 18 and 24 months (PSA data became too scarce to calculate meaningful means after 24 months). The patterns exhibited by these cohort-specific means show a tendency for PSA to go down initially (PSA response), and to come up again after a while (PSA progression).

At the individual level, the PSA longitudinal process was correlated with the hazard rate, which is the risk of dying at a certain time for a patient who has survived up until that time. The coefficient of determination between the PSA process and the hazard rate ($R^2_{\text{individual}}$) is here a function of time that cannot be easily summarized into a single measure [5]. Suffice to say that $R^2_{\text{individual}}$ was greater than 0.84 at all times to indicate that there was again a strong association, at the individual patient level, between the evolution of PSA and the hazard of dying.

At the group level, the effects of liarozole on longitudinal PSA and on survival were moderately correlated, with a coefficient of determination $R^2_{\text{trial}} = 0.45$ (standard error = 0.18) (Fig. 4b). There was a significant benefit of liarozole in terms of longitudinal PSA ($p = 0.01$); in other words, the profiles shown on Fig. 4a were significantly different between liarozole and control.

V. DISCUSSION

We have illustrated, through an actual example, statistical approaches that may be useful to study the complex relationships between a biomarker, a clinical endpoint, and the effects of a treatment on both the biomarker and the clinical endpoint. Our analyses emphasize the importance of distinguishing between two types of association: one between the biomarker and the clinical endpoint at the individual level, the other between the effects of treatment on the biomarker and on the clinical endpoint at the trial level. Since only two trials were available for our analyses, we considered country in each trial as the grouping unit of interest. Table 1 summarizes the measures of association between survival and,

Table 1 Individual-Level and Trial-Level Measures of Association Between PSA and Survival in Advanced Prostate Cancer Treated with Either Liarozole or Control[a]

	Individual-level association between PSA and survival [95% confidence interval]	Trial-level association between treatment effects on PSA and survival [standard error]
PSA response	Survival odds ratio = 5.5 (2.7–8.2)	$R^2_{\text{trial}} = 0.05\,(0.13)$
Time to PSA progression	Survival odds ratio = 6.3 (4.4–8.2)	$R^2_{\text{trial}} = 0.22\,(0.18)$
Longitudinal PSA	Coefficient of determination $R^2(t) > 0.84$ at all times t	$R^2_{\text{trial}} = 0.45\,(0.18)$

[a] The individual-level measures show strong associations between PSA and survival, but the trial-level measures show weak associations between the treatment effects on PSA and survival, making PSA a poor surrogate for survival (odds ratio: see text; R^2 = coefficient of determination).

successively, response to PSA, time to PSA progression, and longitudinal PSA (rows in Table 1). It appears clearly that PSA does not qualify as an acceptable surrogate, regardless of how it is analyzed, in spite of its strong associations with survival at the individual level (second column of Table 1). The associations between treatment effects at the trial level are all low (third column of Table 1). Even when the full PSA pattern is taken into account in a longitudinal analysis, R^2_{trial} is still too low to permit reliable prediction of the effect of treatment on the clinical endpoint, having observed the effect of treatment on the biomarker.

It is also clear from Table 1 that the trial-level associations are estimated rather imprecisely, because of the relatively small number of units (centers) available to estimate treatment effects. In general, the individual-level associations can be estimated far more precisely, because of the large number of patients available [1–8].

It should be noted that the methodology we propose is exploratory in nature, and does not purport to classify a biomarker as a "valid" or "invalid" surrogate for a clinical endpoint—although if both $R^2_{\text{individual}}$ and R^2_{trial} were close to 1, we would be in a position to claim the surrogate to be acceptable. Indeed, in such a case, the surrogate would be strongly associated to the clinical endpoint, and any *change* in the surrogate would also translate into a corresponding (and predictable) change in the clinical endpoint. However, caution would still be in order, for neither of these statistical associations would prove a causal impact of the biomarker on the clinical endpoint. Moreover, the trial-level association would have been established only for the treatment comparison at hand, and could be quite different for some new treatment having a different mode of action.

The validation of a biomarker as a surrogate for a clinical endpoint is no easy task. Many authors have expressed an exceedingly negative view on this problem. Theoretical criticisms have borne on problems with overly strict definitions of surrogacy [12,15,20], the validation criteria proposed by Prentice [12,36], the proportion explained [12,20], computation and modeling difficulties [37], and the meta-analytic approach [38]. On the practical side, some supposed surrogates have dramatically failed to predict clinical outcomes [39]. The approval of the antiarrhythmic drugs flecanaide and encanaide, based on their controlling arrhythmias rather than long-term mortality, will long continue to haunt the debates on whether surrogate endpoints can be used to approve new drugs [15,40]. It seems clear that few, if any, biomarkers will ever qualify as "valid" surrogates in a strict sense of the word. Even if we adopt the more liberal view advocated in this chapter, very few, if any, biomarkers will have large enough values of R^2 to qualify as "acceptable" surrogates [41]. In addition, surrogates that are observed very early on in the course of the disease are the most interesting ones, but also those least likely to predict distant clinical endpoints with any acceptable accuracy. In spite of all difficulties, we believe that the search for surrogates should not be abandoned, for the gains might be too important in terms of patients and/or time. For some endpoints, such as delayed toxicities to experimental treatments, the use of surrogates is simply inescapable. In addition, even if biomarkers always turned out to be poor surrogates, it could still be useful to quantify their relationships to the clinical endpoints of interest, because valuable knowledge might well be derived in the process.

A final word on the need for data. The methods presented here require data from several (possibly many) randomized trials to be available. Access to data from randomized trials is difficult, especially for phase III trials carried out by pharmaceutical companies seeking registration of new drugs. We contend that only way to seriously search for valid surrogate biomarkers is to make these data fully accessible for statistical analysis and public scrutiny. Once new drugs are approved, individual patient data from randomized clinical trials upon which the approval was based should be made publicly accessible, as are data from some cooperative groups (the AIDS Clinical Trials Group, for instance). Further analyses of such data in clinical situations of interest may illuminate issues related to surrogate endpoints that, in the absence of detailed statistical analyses, would have remained controversial at best, and ignored at worst.

ACKNOWLEDGMENTS

The authors thank the Janssen Research Foundation for permission to use data from two clinical trials testing liarozole in patients with advanced prostrate cancer.

REFERENCES

1. Buyse, M.; Molenberghs, G. Criteria for the validation of surrogate end-points in randomized experiments. Biometrics **1998**, *54*, 1014–1029.
2. Molenberghs, G.; Geys, H.; Buyse, M. Evaluation of surrogate end-points in randomized experiments with mixed discrete and continuous outcomes. Stat. Med. **2001**, *20*, 3023–3038.
3. Buyse, M.; Molenberghs, G.; Burzykowski, T.; Renard, D.; Geys, H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics **2000**, *1*, 49–68.
4. Burzykowski, T.; Molenberghs, G.; Buyse, M.; Geys, H.; Renard, D. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. Appl. Stat. **2001**, *50*, 405–422.
5. Renard, D.; Geys, H.; Molenberghs, G.; Burzykowski, T.; Buyse, M. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. Biom. J. *in press*.
6. Alonso, A.; Geys, H.; Molenberghs, G.; Vangeneugden, T. Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology, submitted.
7. Buyse, M.; Thirion, P.; Carlson, R.W.; Burzykowski, T.; Molenberghs, G.; Piedbois, P., for the Meta-Analysis Group in Cancer. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Lancet **2000**, *356*, 373–378.
8. Burzykowski, T.; Molenberghs, G.; Buyse, M.; Geys, H.; Renard, D. The validation of surrogate endpoints using data from randomized clinical trials: a case study in advanced colorectal cancer, submitted.
9. Prentice, R.L. Surrogate endpoints in clinical trials: definitions and operational criteria. Stat. Med. **1989**, *8*, 431–440.
10. Schatzkin, A.; Freedman, L.S.; Schiffman, M.H.; Dawsey, S.M. Validation of intermediate end points in cancer research. J. Natl Cancer Inst. **1990**, *82*, 1746–1752.
11. Freedman, L.S.; Graubard, B.I.; Schatzkin, A. Statistical validation of intermediate endpoints for chronic diseases. Stat. Med. **1992**, *11*, 167–178.
12. Molenberghs, G.; Buyse, M.; Geys, H.; Renard, D.; Burzykowski, T. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Control. Clin. Trials. *in press*.
13. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer Series in Statistics; Springer: New York, 2000.
14. Lagakos, S.W.; Hoth, D.F. Surrogate markers in AIDS: where are we? Where are we going? Ann. Intern. Med. **1992**, *116*, 599–601.
15. Fleming, T.R.; DeMets, D.L. Surrogate end points in clinical trials: are we being misled? Ann. Intern. Med. **1996**, *125*, 605–613.
16. Jacobson, M.A.; Bacchetti, P.; Kolokathis, A.; et al. Surrogate markers for survival in patients with AIDS and AIDS related complex treated with zidovudine. Br. Med. J. **1991**, *302*, 73–78.

17. Lin, D.Y.; Fischl, M.A.; Schoenfeld, D.A. Evaluating the role of CD4-lymphocyte change as a surrogate endpoint in HIV clinical trials. Stat. Med. **1993**, *12*, 835–842.
18. Choi, S.; Lagakos, S.; Schooley, R.T.; Volberding, P.A. CD4$^+$ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. Ann. Intern. Med. **1993**, *118*, 674–680.
19. Lin, D.Y.; Fleming, T.R.; De Gruttola, V. Estimating the proportion of treatment effect explained by a surrogate marker. Stat. Med. **1997**, *16*, 1515–1527.
20. Flandre, P.; Saidi, Y. Letter to the editor: estimating the proportion of treatment effect explained by a surrogate marker. Stat. Med. **1999**, *18*, 107–115.
21. A'Hern, R.P.; Ebbs, S.R.; Baum, M.B. Does chemotherapy improve survival in advanced breast cancer? A statistical overview. Br. J. Cancer **1988**, *57*, 615–618.
22. Daniels, M.J.; Hughes, M.D. Meta-analysis for the evaluation of potential surrogate markers. Stat. Med. **1997**, *16*, 1515–1527.
23. Debruyne, F.J.M.; Murray, R.; Fradet, Y.; Johansson, J.E.; Tyrrell, C.; Boccardo, F.; Denis, L.; Marberger, J.M.; Brune, D.; Rassweiler, J.; Vangeneugden, T.; Bruynseels, J.; Janssens, M.; de Porre, P., for the Liarozole Study group. Liarozole—a novel treatment approach for advanced prostate cancer: results of a large randomized trial versus cyproterone acetate. Urology **1998**, *52*, 72–81 .
24. Sridhara, R.; Eisenberger, M.A.; Sinibaldi, V.J.; et al. Evaluation of prostate-specific antigen as a surrogate marker for response of hormone-refractory prostate cancer to suramin therapy. J. Clin. Oncol. **1995**, *13*, 2944–2953.
25. Smith, D.C.; Dunn, R.L.; Stawderman, M.S.; et al. Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer. J. Clin. Oncol. **1998**, *16*, 1835–1843.
26. Kelly, W.K.; Scher, H.I.; Mazumdar, M.; et al. Prostate-specific antigen as a measure of disease outcome in metastatic hormone-refractory prostate cancer. J. Clin. Oncol. **1993**, *11*, 607–615.
27. Scher, H.I.; Kelly, W.K.; Zhang, Z.F.; et al. Post-therapy serum prostate-specific antigen level and survival in patients with androgen-independent prostate cancer. J. Natl Cancer Inst. **1999**, *91*, 244–251.
28. Bubley, G.J.; Carducci, M.; Dahut, W.; Dawson, N.; Daliani, D.; Eisenberger, M.; Fidd, W.D.; Freidlin, B.; Halabi, S.; Hudes, G.; Hussain, M.; Kaplan, R.; Myers, C.; Oh, W.; Petrylak, D.P.; Reed, E.; Roth, B.; Sartor, O.; Scher, H.; Simons, J.; Sinibaldi, V.; Small, E.J.; Smith, M.R.; Trump, D.L.; Vollmer, R.; Wilding, G. Eligibility and response guidelines for phase II clinical trials in androgen-independent prostate cancer: recommendations from the prostate-specific antigen working group. J. Clin. Oncol. **1999**, *17*, 3461–3467.
29. Buyse, M.; Piedbois, P. On the relationship between response to treatment and survival. Stat. Med. **1996**, *15*, 2797–2812.
30. Anderson, J.R.; Cain, K.C.; Gelber, R.D. Analysis of survival by tumor response. J. Clin. Oncol. **1983**, *1*, 710–719.

31. De Gruttola, V.; Wulfsohn, M.; Fischl, M.A.; Tsiatis, A. Modelling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. J. AIDS **1993**, *6*, 359–365.
32. De Gruttola, V.; Tu, X.M. Modelling progression of CD-4 lymphocyte count and its relationship to survival time. Biometrics **1995**, *50*, 1003–1014.
33. Diagnostic and therapeutic technology assessment (DATTA): surrogate markers of progressive HIV disease. J. Am. Med. Assoc. **1992**, *267*, 2948–2952.
34. Ellenberg, S.S. Surrogate endpoints in clinical trials: getting closer to identifying markers for survival in AIDS. Br. Med. J. **1991**, *302*, 63–64.
35. Machado, S.G.; Gail, M.H.; Ellenberg, S.S. On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection. J. AIDS **1990**, *3*, 1065–1073.
36. Begg, C.; Leung, D. On the use of surrogate endpoints in randomized trials. J. R. Stat. Soc. A **2000**, *163*, 26–27.
37. Tibaldi, F.S.; Abrahantes, J.C.; Molenberghs, G.; Renard, D.; Burzykowski, T.; Buyse, M.; Parmar, M.; Stijnen, T.; Wolfinger, R. Computational approaches to the evaluation of surrogate endpoints, submitted.
38. Gail, M.H.; Pfeiffer, R.; van Houwelingen, H.C.; Carroll, R.J. On meta-analytic assessment of surrogate outcomes. Biostatistics **2000**, *1*, 231–246.
39. Temple, R.J. A regulatory authority's opinion about surrogate endpoints. In *Clinical Measurement in Drug Evaluation*; Nimmo, W., Ticker, G., Eds.; Chichester: Wiley, 1995.
40. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators; Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infraction. N. Engl. J. Med. **1989**, *321*, 406–412.
41. Buyse, M.; Molenberghs, G.; Burzykowski, T.; Renard, D.; Geys, H. Statistical validation of surrogate endpoints: problems and proposals. Drug Inf. J. **2000**, *34*, 447–454.