STATISTICS IN MEDICINE, VOL. 8, 431–440 (1989)

SURROGATE ENDPOINTS IN CLINICAL TRIALS: DEFINITION AND OPERATIONAL CRITERIA

ROSS L. PRENTICE

Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, U.S.A.

SUMMARY

I discuss the idea of using surrogate endpoints in the context of clinical trials to compare two or more treatments or interventions in respect to some 'true' endpoint, typically a disease occurrence. In order that treatment comparison based on a surrogate response variable have a meaningful implication for the corresponding true endpoint treatment comparison, a rather restrictive criterion is proposed for use of the adjective 'surrogate'. Specifically, I propose that a surrogate for a true endpoint yield a valid test of the null hypothesis of no association between treatment and the true response. This criterion essentially requires the surrogate variable to 'capture' any relationship between the treatment and the true endpoint, a notion that can be operationalized by requiring the true endpoint rate at any follow-up time to be independent of treatment, given the preceding history of the surrogate variable. I then discuss this operational criterion in the examples of the accompanying papers[1-3] and in the setting of trials aimed at the primary and secondary prevention of cancer.

KEY WORDS Clinical trials Disease prevention trials Hazard rates Surrogate endpoints Therapeutic trials

1. INTRODUCTION

The accompanying papers discuss potential surrogate endpoints and observations in the contents of cancer treatment trials (Ellenberg and Hamilton)[1], cardiovascular disease prevention and treatment trials (Wittes, Lakatos and Probstfield)[2], and ophthalmologic studies (Hillis and Seigel)[3]. Section 2 examines these papers in respect to surrogate endpoint definition, and a more restrictive criterion for the use of the label 'surrogate' is proposed. In Section 3 an attempt is made to operationalize this criterion, and an application to mammographic screening to reduce breast cancer mortality is discussed in Section 4. In Section 5 several of the examples of the papers[1-3] are then revisited in terms of the proposed surrogate response variable criterion, and further applications in the primary prevention of cancer are mentioned. Section 6 provides generalizations and discussion.

2. DEFINITION OF SURROGATE ENDPOINTS

A primary motivation for the use of a surrogate endpoint, emphasized in each of the accompanying papers,[1-3] concerns the possible reduction in sample size or trial duration that we can expect when a rare or distal endpoint is replaced by a more frequent or proximate endpoint. Such reductions have important cost implications, and in some cases may be pivotal in regard to trial feasibility. Other motivations, provided in the preceding papers, include the possibility that true

0277–6715/89/040431–10$05.00

Received December 1987
Revised October 1988

endpoint measurement may be unduly invasive, uncomfortable or expensive, so that we may intentionally permit a degree of measurement error regarding the timing or even the occurrence of a true endpoint event. Furthermore, both Ellenberg and Hamilton[1] and Wittes, Lakatos and Probstfield[2] point out that endpoint events close in time to the treatment or intervention activities under study may be more readily interpreted than are more distal endpoints, such as study subject death, which may be confounded by secondary treatments or competing risks. This motivation, however, seems to relate more to the choice of principal endpoint than to the issue of defining a replacement, or surrogate, for a selected endpoint.

Concerning the definition of a surrogate endpoint, Ellenberg and Hamilton[1] write that 'investigators use surrogate endpoints when the endpoint of interest is too difficult and/or expensive to measure routinely and when they can define some other, more readily measurable, endpoint, which is sufficiently well correlated with the first to justify its use as a substitute.' Wittes, Lakatos and Probstfield[2] simply define a surrogate endpoint as 'an endpoint measured in lieu of some other so-called "true" endpoint'. Hillis and Seigel[3] define a surrogate observation as 'an observed variable that relates in some way to the variable of primary interest, which we cannot conveniently observe directly'.

Each group of authors goes on to expand on the notion of close relationship between the true endpoint and the potential surrogate. For example, Ellenberg and Hamilton comment that 'surrogate endpoints are generally proposed on the basis of a biological rationale' and that 'when this rationale is less than compelling one must establish statistically that the correlation with the "true" endpoint is sufficiently great to justify the surrogate endpoint as a basis for inference.' Similarly, Wittes, Lakatos and Probstfield write 'one can consider a variable to be a surrogate if it bears a clear statistical relationship with the true endpoint', but they caution that 'a convincing surrogate should have both biological relevance and face validity; statistical relationship alone is not sufficient.' Hillis and Seigel argue that the suitability of using elevated intraocular pressure as a surrogate for long-term visual function loss in glaucoma trials depends on 'a generally accepted set of assumptions as to the etiologic role of intraocular pressure in the progression of glaucoma'.

In considering criteria for use of the term 'surrogate' it is natural to ask what we require of a treatment comparison based on a surrogate response variable. Even though a range of endpoint comparisons may have relevance to an understanding of the effects of treatments under study, it seems logical to restrict the use of surrogate to response variables that can substitute for a true response variable for certain purposes. Equivalently, it seems reasonable to require a surrogate for some true endpoint to have potential to yield unambiguous information about differential treatment effects on the true endpoint. While one could attempt to require a surrogate response to provide some quantitative information on the comparison of true endpoint rates among treatments, a criterion involving only a qualitative link will be much more readily applied. Hence, I define a surrogate endpoint to be *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.*

Note that this prescription defines a surrogate for a given endpoint in a manner that depends on the treatments or interventions under comparison. While it may seem attractive to seek a response variable that can universally serve as a substitute for a certain true response variable, such a desire does not seem practical. For example, some response variables constructed from a series of blood pressure measurements may serve as a surrogate for coronary heart disease incidence for the purpose of comparing anti-hypertensive treatments, but such a response variable is unlikely to be suitable for the comparison of cholesterol lowering drugs. Cigarette smoking histories may provide the basis for a surrogate response to compare educational programmes to prevent lung cancer, but such a response would evidently be inappropriate for the comparison of chemopreven-

tive approaches to lung cancer prevention. These examples cause one to think in terms of disease pathways in which the surrogate response is sensitive to treatment differences that may affect the true endpoint event rates, a notion that relates closely to those of 'biological relevance', 'face validity' and 'etiologic role' mentioned by the authors of the preceding papers. The next section attempts to formalize the idea that a surrogate response captures treatment differences as they affect the true endpoint, in the course of operationalizing the above surrogate endpoint definition.

3. OPERATIONAL CRITERION FOR THE IDENTIFICATION OF SURROGATE ENDPOINTS

Often the true endpoint in a clinical trial will be a time-to-failure variate, as is illustrated by many of the examples of the preceding papers. To be specific, let t denote time from enrolment in a clinical trial and let T denote a true time-to-failure endpoint. Typically our interest in defining a surrogate for T will be motivated by the possibility of an earlier treatment comparison or a smaller sample size. Let $x=(x_1, \ldots, x_p)$ consist of indicator variates for $p(\geqslant 1)$ of the $p+1$ treatments to be compared in respect to the corresponding (instantaneous) failure rates $\lambda_T(t; x)$. A whole range of measurements, taken during the course of trial follow-up, could be used in the specification of a surrogate for T. Let $S(t)=\{Z(u); 0\leqslant u<t\}$ denote the history prior to time t of a possibly vector-valued stochastic process $Z(u)=\{Z_1(u), Z_2(u), \ldots\}$ which may be used to specify a surrogate for T. For example, $S(t)$ may capture aspects of the history (prior to time t) of a time-to-response variable $S\leqslant T$, in which case we might define

$$Z(t)=\begin{cases}0 & t<S\\ 1 & t\geqslant S.\end{cases}$$

Alternatively, $S(t)$ may consist of quantitative measurements $Z(u)$ (for example, blood pressure or cholesterol measurements) at specified times $u_1, u_2, \ldots, u_k$ after trial enrolment.

A link between the true endpoint failure rate $\lambda_T(t; x)$ and a potential surrogate response variable $S(t)$ can be obtained as follows:

$$\begin{aligned}\lambda_T(t; x) &= E[\lambda_T\{t; S(t), x\}]\\ &= \int \lambda_T\{t; S(t), x\}\, \mathrm{d}\, pr\{S(t); x, F(t)\},\end{aligned} \tag{1}$$

where E denotes expectation over the distribution of $S(t)$ given $\{x, F(t)\}$; $F(t)$ consists of the failure and censoring histories prior to t for the true endpoint T (note that by definition λ_T is conditional on $F(t)$); and $pr\{\cdot;\cdot\}$ denotes the conditional probability distribution of $S(t)$. The notion, mentioned above, that a surrogate for T should be able to capture the dependence of T on treatment x can now be expressed as

$$\lambda_T\{t; S(t), x\} \equiv \lambda_T\{t; S(t)\}, \tag{2}$$

where here and subsequently $\equiv$ implies equality for all t up to sets of probability zero. The criterion (2) requires the failure rate for T to be independent of treatment, conditional on the surrogate process. It establishes a correspondence between the surrogate variable probability distribution $pr\{S(t); x, F(t)\}$ and the true endpoint probability distribution, as expressed in the hazard function $\lambda_T(t; x)$.

Specifically, if the surrogate response is independent of treatment in the sense that

$$pr\{S(t); x; F(t)\} \equiv pr\{S(t); F(t)\} \tag{3}$$

then (1), under (2), implies

$$\lambda_T(t; x) \equiv \lambda_T(t). \tag{4}$$

Hence any departure from the null hypothesis (4) will be reflected in a departure from the null hypothesis (3) under the criterion (2), so that (2) requires the surrogate variable to be fully sensitive to any treatment difference in true endpoint rates.

For departure from the surrogate response null hypothesis (3) to imply departure from (4) it is necessary for the surrogate response to have some prognostic implication for the true endpoint; that is,

$$\lambda_T\{t; S(t)\} \not\equiv \lambda_T(t), \tag{5}$$

since otherwise (4) would hold under (2), regardless of the extent of departure from (3).

We will take (2) and (5) as our operational criteria for surrogate endpoint definition. In using a test of (3) in place of a test of (4) it will, however, be necessary to restrict the class of alternatives to (3) to those for which the treatment effects on the surrogate response distribution have some impact on average true endpoint risk; that is, to alternatives for which

$$E[\lambda_T\{t; S(t)\} | x, F(t)] \not\equiv E[\lambda_T\{t; S(t)\} | F(t)]. \tag{6}$$

Essentially we are excluding treatment differences on a surrogate endpoint distribution that are exactly compensatory in respect to average true endpoint risk, at almost all follow-up times. Such restriction will typically be innocuous but will require justification using knowledge of the dependence of $\lambda_T\{t; S(t)\}$ on $S(t)$ whenever rejection of (3) is used to argue rejection of (4). Note that the power of a test of (3) under a specified departure from (4) will, under (2), evidently depend in a complicated fashion on $\lambda_T\{t; S(t)\}$, as well as on aspects of the associated dependence of $pr\{S(t); x, F(t)\}$ on x.

Another issue, emphasized in the preceding paper by Wittes, Lakatos and Probstfield, concerns the possibility that the surrogate response $S(t)$ may be missing at t, even though the subject is being actively followed for a true endpoint response. Clearly a test of (3) will continue to be valid as a replacement for a test of (4) provided that, at almost all t, $pr\{S(t); x, F(t)\}$ is common for subjects with and without missing data. More generally, if the rate of missing data differs among treatments then additional analytic work would be required to construct a null hypothesis test of the measured surrogate variable that is equivalent to a test of (3).

4. SURROGATE ENDPOINTS FOR THE EVALUATION OF BREAST CANCER SCREENING TRIALS

Recently the National Cancer Institute requested proposals to identify surrogate endpoints that may allow cancer screening programmes to be evaluated in a more timely fashion than is possible using site-specific cancer mortality rates as endpoints. As a result, some of my colleagues at Group Health Cooperative of Puget Sound (Drs. Robert Thompson, Ed Wagner, and Lou Grothaus) are conducting a study to evaluate possible definitions of 'advanced' breast cancer incidence as a surrogate for breast cancer mortality in the comparison of mammographic screening programmes.

Suppose that a time to 'advanced' breast cancer occurrence variate S is defined as a potential surrogate for breast cancer death. For concreteness, suppose we say that S occurs for a study subject if a breast cancer diagnosis is made based on primary tumour of a specified minimum size (for example, 2.0 cm diameter) and a certain minimum extent of regional or distant spread (such as

some nodal involvement). Let us define the corresponding surrogate response variable by

$$S(t)=\begin{cases}1 & \text{if } S\leqslant t\\ 0 & \text{if } S>t\end{cases}.$$

In this special case expression (1) becomes

$$\lambda_T(t;x)=\lambda_T\{t|S(t)=0,x\}\,pr\{S>t;x,F(t)\}+\lambda_T\{t|S(t)=1,x\}\,pr\{S<t;x,F(t)\}.$$

Hence, under the surrogate endpoint criterion (2), namely

$$\lambda_T\{t|S(t),x\}\equiv\lambda_T\{t|S(t)\},$$

one has

$$\lambda_T(t;x)=\lambda_T(t|S\leqslant t)-\{\lambda_T(t|S\leqslant t)-\lambda_T(t|S>t)\}\,pr\{S>t;x,F(t)\}.$$

Thus, there is a one-to-one correspondence between the dependence of $\lambda_T(t;x)$ on x and the dependence of $pr\{S>t;x,F(t)\}$, provided

$$\lambda_T(t|S\leqslant t)>\lambda_T(t|S>t),\quad \text{all } t; \tag{7}$$

that is, provided the prior occurrence of advanced cancer increases the risk of breast cancer death at all follow-up times. Also $pr\{S>t;x,F(t)\}$ is simply the probability that an advanced cancer diagnosis has not been made by time t, given the mammographic screening pattern x, and given that neither breast cancer death nor censoring of the breast cancer mortality variate have occurred prior to time t. If the advanced breast cancer incidence variate S is subject only to the same sources of (independent) censorship as the breast cancer mortality variable T, then

$$pr\{S>t;x,F(t)\}=\exp\left\{-\int_0^t\lambda_S(s;x)\,ds\right\},$$

where $\lambda_S(t;x)$ is the hazard function for S. More generally, S may be subject to additional independent censorship with censoring rate $\phi_S(t;x)$, in which case

$$pr\{S>t;x,F(t)\}=\exp\left\{-\int_0^t\lambda_S(s;x)\,ds\right\}\exp\left\{-\int_0^t\phi_S(s;x)\,ds\right\}.$$

Hence, if we assume in addition that

$$\phi_S(t;x)\equiv\phi_S(t), \tag{8}$$

then $\lambda_T(t;x)\equiv\lambda_T(t)$ if, and only if, $\lambda_S(t;x)\equiv\lambda_S(t)$. Conditions (2), (7) and (8) are enough to ensure that a test of $pr\{S>t;x,F(t)\}\equiv pr\{S>t;F(t)\}$ is also a valid test of $\lambda_T(t;x)\equiv\lambda_T(t)$. Conditions (7) and (8), both of which seem quite natural in this context, allow one to assert that standard failure times methods (for example, logrank test) for testing $\lambda_S(t;x)\equiv\lambda_S(t)$ using the potential surrogate S will provide a valid test under the corresponding null hypothesis on T. The power of such a test is evidently enhanced if the occcurrence of an advanced cancer markedly increases breast cancer mortality risk (that is, $\lambda_T(t|S\leqslant t)\gg\lambda_T(t|S>t)$) and if the additional censoring rates, $\phi_S(t)$, for the surrogate endpoint are small.

Let us consider further the surrogate endpoint criterion (2) in this breast screening setting. Condition (2) with $S(t)=0$ implies that the breast cancer death rate in the absence of an advanced breast cancer diagnosis must be independent of mammographic screening assignment x. In practical terms, this presumably means that the advanced breast cancer category must be inclusive enough so that breast cancers not included convey minimal breast cancer mortality risk, either because of their expected natural history, or because curative treatments are available.

Expression (2) with $S(t)=1$ is concerned with breast cancer mortality rates following an 'advanced' breast cancer diagnosis. Such mortality rates are to be independent of exposure to mammographic screening. Hence the advanced disease definition needs to be stringent enough that the subsequent breast cancer mortality risk is common to screened and unscreened women, thereby competing with the all-but-minimal risk requirement of the preceding paragraph. In other words, for (2) to hold, the screening programmes being compared may affect the frequency of advanced cancer diagnosis, but may not affect breast cancer mortality risk given such diagnosis or the absence thereof. If such an advanced disease endpoint could be identified, however, it would presumably be able to serve as endpoint for a range of other studies using modifications of the same screening modalities.

Biological considerations may guide one to an advanced breast cancer definition for which (2) is plausible. A database that includes both true and potential surrogate endpoints may allow a useful empirical test of (2). For example, we may specify

$$\lambda_T\{t|S(t), x\}=\lambda_{0T}(t)\exp[xS(t)\beta_1+x\{1-S(t)\}\beta_2+S(t)\beta_3]$$

and test $\beta_1=0$ and $\beta_2=0$ using standard partial likelihood procedures (for example, Kalbfleisch and Prentice[4]). Note that testing $\beta_1=0$ examines whether there is evidence of dependence of breast cancer mortality rates on prior screening among women who have experienced an advanced breast cancer diagnosis, while testing $\beta_2=0$ examines such evidence among women without an advanced breast cancer diagnosis. One could also test assumptions (7) and (8) in a straightforward manner with use of standard censored failure time methods.

Endpoints more complicated than the above advanced breast cancer indicator variate may be required to meet the surrogate response variable criteria (2). For example, the process $S(t)$ could be defined to be a vector that includes an indicator variable for any breast cancer diagnosis along with a second indicator variable for the presence of residual or recurrent disease following a full course of treatment. Innovative statistical methods would then be required to test the surrogate variate null hypothesis (3), and hence the true endpoint null hypothesis (4).

5. OTHER ILLUSTRATIONS

Let us now consider some of the examples of the accompanying papers[1-3] in respect to the surrogate endpoint criterion (2), and the null hypothesis test based on (3). Ellenberg and Hamilton consider tumour response as a surrogate for death in cancer therapy trials in which cancer patients have measurable tumour mass. Suppose that $S(t)$ takes value zero prior to a complete response and value one thereafter, for each patient. Criterion (2), $\lambda_T\{t; S(t), x\}=\lambda_T\{t; S(t)\}$, then requires death rate to be independent of the treatments under study, given the patients tumour response status. In particular, the treatments under study are not allowed to differentially influence mortality via pathways that bypass the primary tumour mass, as may arise if there are treatment toxicities that are life-threatening, or if the treatments have different regional or systemic effects. This criterion also requires the mortality risk to be independent of treatments among patients not experiencing a complete response. Such independence may not hold if the treatments differ in their ability to engender a partial response, in which case a tumour response variate having additional response categories may be considered, Undoubtedly some measure of tumour response is a valuable element of treatment comparison, regardless of whether or not (2) holds. However, if we wish to assert that a test of the null hypothesis of equality of tumour response rates yields a valid null hypothesis test for patient survival, then we must be prepared to defend the stringent criterion (2).

Ellenberg and Hamilton also consider disease recurrence as a surrogate for death in adjuvant cancer treatment trials in which the disease has apparently been eradicated. Condition (2), with

$S(t)$ defined as an indicator variable that takes value zero prior to disease recurrence and value one thereafter, requires mortality rates to be independent of treatment among patients without recurrent disease ($S(t)=0$), thereby not allowing differential treatment effects on competing risks. Mortality rates must also be independent of treatment among patients who have experienced disease recurrence – an assumption that may be suspect, for example, if disease recurrences have a different anatomical distribution under the study treatments. This circumstance could likely be overcome by extending the surrogate specification to include measurements that characterize the nature of the recurrence and the associated mortality risk. The idea of using disease-free survival over a definite time period (for example, 2 years) as a surrogate for cancer cure is rather difficult to formulate as it involves a comparison of death rates among such patients with corresponding rates for comparable persons without prior cancer diagnosis. The related idea of using mortality during the first years (for example, 2 years) of a study as a surrogate for overall mortality requires only that mortality rates after such an initial time period be independent of treatment, in order to satisfy (2).

Wittes, Lakatos and Probstfield consider several possible surrogate variables for total or cardiovascular disease mortality, including ejection fraction for trials of thrombolytic agents, blood pressure reduction in trials of anti-hypertensive agents, and blood cholesterol in trials of cholesterol lowering drugs. As formalized by criterion (2) such variables may serve as surrogates for mortality provided the treatments under study do not differentially affect mortality via pathways that bypass the proposed surrogate, and, more generally, provided mortality rates do not depend on the interventions being compared given the pertinent surrogate variable history. The situation envisaged by Wittes, *et al.*, wherein 'a new intervention may reduce the risk factor by some pathway irrelevant to the development of morbid events' can be expected to yield a violation of (2) if applied to the new and previous interventions. The presentation by these authors reinforces the above message that the suitability of a response variable as a surrogate for mortality rate depends very much on the treatments or interventions under comparison.

Hillis and Seigel consider the use of intraocular pressure as a surrogate for long-term visual function in glaucoma trials. Criterion (2) indicates that its suitability requires blindness rates to be independent of the treatments under study given a single, or multiple, measurement of intraocular pressure. This is essentially a restatement of Hillis and Seigel's assumption 3, with restriction to the specific treatments under comparison. Note that the treatments under evaluation must not differ in their effect on blindness risk via pathways that do not involve intraocular pressure. Hillis and Seigel's assumptions 1 and 2 pertain to the relationship between $\lambda_T\{t; S(t)\}$ and $S(t)$. Expressions (2) and (6) evidently capture the ideas of these assumptions upon restricting the evaluation to the treatments under study.

The surrogate variable criterion (2) can also be considered in the context of cancer prevention trials. With cancer incidence as endpoint, trials in this area need to be so large and expensive that only a handful have been mounted to date. A prevalent view of cancer cell formation involves an initial transformation from normal cell to intermediate cell and a second transformation to a cancer cell ((for example, Moolgavkar and Knudson). Surrogates based on the occurrence of new intermediate cells could be valuable for the comparison of agents to suppress initiation. However, the molecular events that may define an intermediate cell for a given cancer have been defined only for a few rare cancers. Furthermore, such intermediate cells may need to proliferate into sizeable benign lesions before detection is possible. Hence any surrogate endpoint opportunity in cancer prevention is more likely to focus on the comparison of promoting agents; that is, agents that affect the division rates of intermediate cells and hence affect the chance that one or more such cells will undergo a second event and become cancerous. A study carried out in our cancer prevention research programme in Seattle by Dr. Joseph Chu provides an illustration of an intermediate

lesion being used to evaluate a potential cancer preventative agent. Women with mild or moderate cervical dysplasia are randomized to oral folate supplementation or control and followed for the regression or progression of dysplasia. Suppose folate supplementation is shown to beneficially affect the severity of dysplasia after some months of follow-up. In order to interpret this result as providing evidence of the ability of folate to prevent cervical carcinoma it is necessary, under condition (2), to argue that a given severity of dysplasia conveys the same cervical carcinoma risk among treated and control study subjects, and that folate supplementation does not affect transition rates from a dysplastic to a cancer cell. In practical terms, not just a single cancerous cell but a clone of malignant cells would be required for a cervical carcinoma diagnosis. Hence, for a test of improvement in dysplasia severity to be equivalent to a reduction in cervical carcinoma incidence, it would also be necessary to assume that folate also does not affect proliferation rates of malignant cells.

In recent years dietary factors have emerged as having potential for preventing a range of cancers, for example, beta carotene and retinol supplementation for the prevention of lung cancer and other epidermoid cell cancers, and dietary fat reduction for the prevention of breast, colon, prostate and other cancers. Experimental studies in animals suggest that these dietary factors, like folate in the preceding example, may play a role in the promotional phase of carcinogenesis. Suitable intermediate marker lesions have yet to be identified for these important cancers, precluding consideration of the applicability of criterion (2).

6. GENERALIZATIONS AND DISCUSSION

The above discussion was restricted to a time-to-response true endpoint. The preceding papers include examples of other types of true endpoints. For example, Ellenberg and Hamilton consider measurements of carcinoembryonic antigen as a surrogate for tumour response. Hillis and Seigel consider short-term (2-year) cumulative vision loss as a surrogate for longer-term (5-year) cumulative vision loss in a diabetes retinopathy study, and vascular changes in an unaffected eye as a surrogate for such changes in the corresponding eye in which haemorrhage has obscured the retina.

The above discussion was also framed in terms of a clinical trial to compare two or more treatments, denoted by a set of indicator variables x. More generally, we may wish to consider a surrogate response that will be able to replace a true response variable in respect to its dependence on some general exposure or covariate history. Advanced breast cancer incidence as a surrogate for breast cancer mortality in the evaluation of subject-selected breast screening exposures (Section 4) provides illustration.

Suppose that the true response variable in a cohort study is a stochastic process $\{T(t), t \geqslant 0\}$, where t may denote time from the beginning of cohort follow-up and

$$T(t) = \{Y_1(u), Y_2(u), \ldots; 0 \leqslant u < t\},$$

while the potential surrogate, as above, is a process $\{S(t), t \geqslant 0\}$, where

$$S(t) = \{Z_1(u), Z_2(u), \ldots; 0 \leqslant u < t\}.$$

Suppose also that the primary endpoint is to be examined in relation to a possibly time-dependent exposure variable (covariate) denoted by $X(t) = \{W_1(u), W_2(u), \ldots; 0 \leqslant u < t\}$. The probability distribution of T can be linked to that of S by

$$pr\{T(t); X(t), F(t)\} \equiv E[pr\{T(t); S(t), X(t), F(t)\}]$$

where the expectation is over the distribution of $S(t)$ given $\{X(t), F(t)\}$, and $F(t)$ is as above.

As an extension of (2) we may take

$$pr\{T(t); S(t), X(t), F(t)\} \equiv pr\{T(t); S(t), F(t)\} \tag{9}$$

as a criterion for S to serve as a surrogate for T. In order that a test of the null hypothesis

$$pr\{S(t); X(t), F(t)\} \equiv pr\{S(t); F(t)\} \tag{10}$$

be equivalent to the corresponding true endpoint test

$$pr\{T(t); X(t), F(t)\} \equiv pr\{T(t); F(t)\},$$

it will be necessary to restrict the class of alternatives to (10) to those for which

$$E[pr\{T(t); S(t), F(t)\} | X(t), F(t)] \not\equiv E[pr\{T(t); S(t), F(t)\} | F(t)].$$

It is perhaps worth reiterating that the relationship between a range of response variables and certain treatments (or covariates) will often be of interest, even though such response variables may not satisfy the surrogate variable criterion (2), or more generally (9). These criteria are proposed here on the assumption that the surrogates for a true endpoint will include only those response variables that permit some direct inference on the relationship between the true endpoint and the treatments under study. Specifically, a surrogate response variable is defined as one for which the null hypothesis of no association with treatment is also a valid test of the null hypothesis of no association of the true endpoint with treatment. Data sets involving both the true and potential surrogate can allow an empirical test of the appropriateness of conditions (2), or (9), but biological or mechanistic considerations would typically also be necessary in order to argue that these conditions hold.

Even though condition (2) is framed in terms of a specific set of treatments x, empirical or theoretic support for a given surrogate may extend its use to a broader, related set of treatments. For example, an advanced breast cancer surrogate, as discussed in Section 4, may be suitable for evaluation of a range of mammographic screening proposals if once established to be appropriate for the comparison of a standard mammography screening programme versus control.

In spite of the hope for such extension I am somewhat pessimistic concerning the potential of the surrogate endpoint concept, as it is interpreted in this paper. One interpretation of criterion (2), or criterion (9), is that the surrogate endpoint must have precisely the same relationship to the true endpoint under each of the treatment strategies being compared. We need only look as far as the Multiple Risk Factor Intervention Trial[6] for an example in which important differences in prominent risk factors – namely blood pressure levels, smoking habits, and blood cholesterol – between intervention and control subjects evidently did not convey the anticipated difference in coronary heart disease mortality. Intuitively, a considerable period may often be necessary before a risk factor reduction conveys maximal true endpoint benefit, and the benefit may never fully compensate for the prior risk elevation. In these circumstances such risk factor histories will not serve as a surrogate for intervention versus control true endpoint comparisons.

The above presentation presumes that a surrogate response S is sought for a true response T that occurs later in time. One could also consider using a later response as surrogate for some, perhaps difficult to measure, earlier response. Criterion (2), or (9), would again provide the link under which the surrogate null hypothesis test also provides a valid test of the true endpoint null hypothesis.

ACKNOWLEDGEMENTS

This work was supported by grant GM-24472 from the National Institute of General Medical Sciences. The author wishes to thank Dr. Jay Herson for organizing a Biometrics Society session on this topic, and Dr. Ed Wagner for comments pertinent to the surrogate endpoint definition provided in Section 2.

REFERENCES

1. Ellenberg, S. S. and Hamilton, J. M. 'Surrogate endpoints in clinical trials: cancer', *Statistics in Medicine*, **8**, 405–413 (1989).
2. Wittes, J., Lakatos, E. and Probstfield, J. 'Surrogate endpoints in clinical trials: cardiovascular diseases', *Statistics in Medicine*, **8**, 415–425 (1989).
3. Hillis, A. and Seigel, D. 'Surrogate endpoints in clinical trials: opthalmologic disorders', *Statistics in Medicine*, **8**, 427–430 (1989).
4. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
5. Moolgavkar, S. H. and Knudson, A. G. 'Mutation and cancer: a model for human carcinogenesis', *Journal of the National Cancer Institute*, **66**, 1037–1052 (1981).
6. MRFIT Research Group. 'Multiple risk factor intervention trial; risk factor changes and mortality results', *Journal of the American Medical Association*, **248**, 1465–1477 (1982).