Journal of Cellular Biochemistry 17G:205–211 (1993)

Surrogate Endpoints in Chemoprevention of Breast Cancer: Guidelines for Evaluation of New Biomarkers

Susan G. Hilsenbeck, PhD and Gary M. Clark, PhD

University of Texas Health Science Center at San Antonio, Department of Medicine/Division of Medical Oncology, San Antonio, TX 78284-7884

Abstract Markers of early events in the development of breast cancer are potential candidates for surrogate endpoints in chemoprevention trials. There are many such markers and the challenge is to identify truly relevant markers. If successful, surrogate endpoints offer several potential benefits in the conduct of prevention trials, including: shorter latency and hence shorter trials; reductions in size and cost of trials; and the opportunity to study prevention measures where use of primary outcomes would be excessively invasive or unethical. Although there are currently no validated surrogate endpoints for breast cancer, criteria for the discovery and validation of surrogates have been proposed. Putative surrogate endpoints should be biologically plausible, represent an early event in the causal pathway, be measurable by a standardized and reliable assay, and exhibit a dose-response. Perhaps most importantly, surrogates should statistically capture the effect of the intervention on outcome. The identification and establishment of a biomarker as a valid surrogate for cancer is a stepwise process that involves both smaller "transitional" studies and larger second-generation chemoprevention trials in which both primary outcome and putative surrogates are measured. Transitional studies are used to move new markers from the laboratory into use in human populations, and are designed to address specific questions of assay validity, treatment/marker associations, marker/disease associations, and inter- and intra-individual variability. Promising markers should be added to current and planned, large, traditionally designed chemoprevention trials in order to definitively address the issues of optimal representation, and to test the adequacy with which the marker(s) captures the effect of treatment on outcome. Ancillary studies of markers attached to these second-generation prevention trials must be powerful enough to detect clinically important differences, to elucidate potentially complex multivariate markers, and to validate hypothesized relationships.

Key words: Biomarkers, breast cancer, prognostic factors, surrogate endpoints

It has been suggested that virtually any measurement of biological structure or process can be viewed as a biomarker [1]. There are numerous biomarkers in breast cancer, ranging from relatively crude measurements of tumor size or dietary fat consumption to the molecular detection of oncogene amplification or mutation of tumor suppressor genes. Biomarkers can be classified into types roughly corresponding to the disease state or study objective motivating their use. Potential uses include assessing exposure, quantifying susceptibility, monitoring compliance, detecting preclinical disease, monitoring intermediate endpoints, predicting prognosis, predicting therapeutic efficacy, and monitoring disease progression (Table I). Some biomarkers may fit several categories. For example, the appearance of a marker for preclinical disease in a previously normal patient might be used as a

Address correspondence to Gary Clark, PhD, University of Texas Health Science Center, Division of Medical Oncology, Department of Medicine, 7703 Floyd Curl Drive, San Antonio, TX 78284-7884.

TABLE I

Type of Marker	Examples of Possible Markers
Exposure	Smoking, alcohol, fat consumption, age
Susceptibility	Family history, breast cancer gene
Compliance	Tamoxifen metabolites
Preclinical disease	Hyperplasia, dysplasia, ductal carcinoma *in situ*
Surrogate endpoints	Response or relapse-free survival for overall survival
Predictors of prognosis	Tumor size, nodes, S-phase fraction
Therapeutic efficacy	Estrogen receptor, progesterone receptor
Disease progression	Carcinoembryonic antigen

surrogate marker for disease in a prevention trial. Although the use of prognostic factors to predict outcome is perhaps the best established use of biomarkers in breast cancer, much work must still be done to definitively establish the close relationship between even these biomarkers and the causal pathway they purport to measure. For example, estrogen receptor (ER) is a strong positive predictor of the benefit of hormone therapy, but not all ER-positive tumors respond, and not all ER-negative tumors fail to respond, to therapy.

BIOMARKERS AS SURROGATE ENDPOINTS

Markers of early events in the path to the ultimate development of breast cancer are potential candidates for surrogate endpoints in chemoprevention trials. There are many such markers, several of which are discussed elsewhere in this issue. New markers appear almost daily. The difficulty arises in deciding which markers are truly relevant. The remainder of this paper will focus on issues surrounding the selection and use of surrogate endpoints. We will first introduce the concept and rationale for surrogate endpoints, and then offer some suggestions for identifying and validating useful surrogate endpoints in the context of chemoprevention.

Examples

The use of surrogate endpoints is not unique to cancer prevention. In the arena of therapeutic clinical trials, where the primary endpoint of interest is overall survival, two commonly used surrogate endpoints are response rate and relapse-free survival. It seems biologically reasonable to expect that treatments that improve response rates or relapse-free survival will have a similar effect on overall survival. While differences in response rates to therapy are often reported as an indication of benefit, a recent review of 56 randomized Phase III trials comparing a control arm to a treatment arm and reporting both overall survival and response rates, found little or no relationship between the two endpoints [2]. Only 50% of trials reporting a significant improvement in survival also showed a difference in response rates. Similarly, differences in relapse-free survival do not always translate into differences in overall survival.

Rationale for Use of Surrogate Endpoints

Ideally, all studies would be designed to detect clinically interesting differences in primary outcome. In chemoprevention trials, the acid test is reduction in cancer occurrence. In reality, this is often not feasible. Surrogate endpoints offer potentially valuable benefits in several areas: shorter latency; reduction in size/cost/duration of trials; and avoidance of unethical practices. First, surrogate endpoints which theoretically lie on or are tightly linked to the causal path between intervention and disease should occur closer in time to the intervention than does the primary outcome (Fig. 1). The effects of treatment should be detectable earlier, with less intervening noise and fewer other factors to cloud the

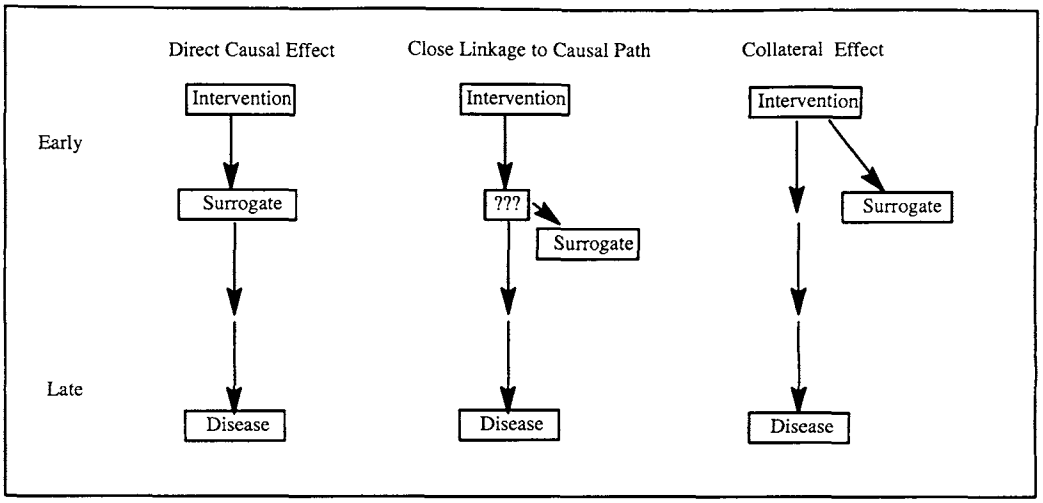

Fig. 1. Diagrammatic representation of the differences in surrogate endpoints directly involved in the causal pathway, tightly linked to an unobserved intermediate biomarker, or only collaterally associated with cancer causation.

relationship. In this sense, valid surrogate endpoints could provide a cleaner, more powerful assessment of treatment efficacy.

Similarly, all else being equal, trials using surrogate endpoints should produce an answer in less time. Even if the same number of participants is required, shorter follow-up could markedly reduce costs. In addition, a major deterrent to conducting cancer prevention trials of any type is the extremely low expected event rate, even in the control group. Huge numbers of participants with substantial follow-up are required in order to accumulate sufficient numbers of incident cancers. If a surrogate endpoint is truly linked to the outcome of interest, then the cumulative incidence of surrogate events will be equal to or greater than that of primary events, and the annual incidence will be greater. The required sample size should be substantially reduced. The use of valid surrogate endpoints will almost certainly result in shorter, smaller, and cheaper trials. Additional savings may accrue if measurement of the surrogate is cheaper or less invasive than the primary endpoint.

Finally, in some instances it may not be ethically feasible to conduct a trial based on the primary endpoint of interest. Measurement of the primary endpoint may be excessively uncomfortable, invasive, or detrimental. For example, a trial requiring repeated biopsies to monitor histologic changes or look for early disease would not be acceptable. Similarly, waiting to observe a change from early to frank malignancy instead of initiating appropriate therapy would not be acceptable.

Definition of a Valid Surrogate Endpoint

Exactly what do we mean by the term "valid surrogate endpoint"? Many authors have proposed definitions, all of which express more or less the same requirements. In essence, a biomarker is a valid surrogate endpoint for the development of breast cancer in relation to an intervention *if and only if* the biomarker captures the effect of the intervention on the development of cancer. In general, this implies that the biomarker is on the causal pathway leading to the development of breast cancer, or is very closely associated with another unobservable direct intermediate on the pathway [1]. The statistical interpretation of this definition is that testing hypotheses about treatment and outcome are equivalent to testing hypotheses about treatment and the marker [3]. This implies that treatment effects on outcome vanish when adjusted for the status of the surrogate [4].

CHARACTERISTICS OF CANDIDATE BIOMARKERS

Candidate markers should exhibit a number of characteristics in order to be considered as po-

Table II. Essential Features of Candidate Markers

Biological hypothesis
Linkage to causal pathway
Dose-response
Standardized, validated assay
Short latency
Statistical validation

tential surrogate endpoints (Table II). First, the selection of a candidate biomarker for use as a surrogate endpoint should be motivated by a biological hypothesis. That is, the marker should have "face validity." Its use should make sense. By extension, it should be, or closely represent, an intermediate step on the causal pathway. It should be clear that the marker is truly part of the pathogenetic pathway of disease and not merely a collateral adaptive response [5]. In the context of surrogate endpoints following carcinogenic exposure, the marker should exhibit a dose-response relationship. In the context of prevention, it seems reasonable to expect the marker to demonstrate a dose-response relationship with risk or susceptibility. In any event, the marker should parallel any dose-response relationship between the intervention and the outcome.

The biomarker should be measurable by a standardized, validated assay. For immunohistochemical markers, this implies the availability of commercially produced monoclonal antibodies, or standards kits. Ideally, the assay should be inexpensive and readily available. Assay validation implies laboratory characterization of sensitivity, specificity, reliability, and quality control. These important issues are addressed in detail elsewhere in these proceedings.

In order to be useful in reducing the size or duration of a prevention trial, a surrogate endpoint must appear substantially earlier than the primary endpoint. Otherwise, in the absence of other ethical or cost considerations, the trial would be completed just as quickly using the primary endpoint.

Finally, there should be convincing statistical evidence that trials based on the surrogate endpoint will draw the same conclusions regarding the efficacy of the intervention as trials based on breast cancer incidence. In this regard, suitably large trials involving the measurement of both primary and surrogate endpoints seem unavoidable and absolutely essential. Prentice [3] has formulated this requirement as a test of the hypothesis that, given the surrogate marker (S), the presence or absence of disease (D) is independent of treatment (T); that is, if the marker truly captures the effect of treatment on outcome, then knowing the treatment does not help predict outcome if you know the marker. This hypothesis can be written formally as conditional probabilities (Pr),

$$\Pr(D \mid S, T+) = \Pr(D \mid S, T-) = \Pr(D \mid S)$$

and could be used to design second-generation prevention trials. If first-generation trials center entirely on the primary outcome, and second-generation trials include both primary and surrogate endpoints, then having demonstrated the statistical equivalence of hypothesis tests based on primary and surrogate endpoints, third-generation trials could be undertaken based solely on surrogate endpoints.

STUDY DESIGN CONSIDERATIONS

Identification and validation of surrogate endpoints is a stepwise process (Table III). Hulka [5] has used the term "transitional study" to describe the collection of studies necessary to move a putative biomarker from the confines of the laboratory to human population studies. They are analogous to pilot studies in the development of prognostic factors [6]. These studies address preliminary questions such as assay reliability, specificity, and sensitivity. They demonstrate consistency with the hypothesized pathway of disease development and adequate assay performance. These studies represent a shift from the deterministic view of laboratory and animal models to the more stochastic, probabilistic view required in population studies.

Characterization of inter- and intra-person variability is essential, as are logistic considerations of measuring the marker, perhaps repeatedly, on large numbers of individuals under varying field conditions. Transitional studies will tend to be small, addressing specific questions. For example, retrospective case/control studies of archival or tissue bank material could help establish the relationship between marker and

TABLE III

Types of Studies	Objectives
Transitional	Establish biological consistency Assay validation Establish associations with intervention, outcome Demonstrate dose-response Assess intra- and inter-person variability
2nd Generation Prevention Trial	Measure and correlate panels of putative markers Optimize and verify representation of surrogate endpoints Validate equivalence of primary and surrogate endpoints
3rd Generation Prevention Trial	Show treatment effect on surrogate endpoints

disease. Other studies might involve very high-risk individuals, where the prevalence of both marker and disease should be high. Freedman and Schatzkin [7] formalized some of the questions that transitional trials might address and provided examples of sample size calculations.

Ultimately, it will be necessary to show the utility of a putative surrogate endpoint in a large prevention trial in which both the primary and surrogate endpoints are measured. Such a trial would be designed using the primary endpoint, but would include measurement and validation of the surrogate endpoint as an ancillary objective. At present there are no established surrogate endpoints in breast cancer, and few if any that have progressed beyond the earliest transitional stage of investigation. We still lack some of the most basic information in defining the causal path of disease development, let alone choosing appropriate markers of the process. Yet, considering the expense and long lead-time involved in planning and mounting prevention trials, every effort should be made now to include measurements of any potential surrogate endpoints in current traditionally designed trials.

There is precedent for such dual purpose studies in the cooperative group Phase III therapeutic trials, such as intergroup studies SWOG 8897 or ECOG 1180. For example, in the ECOG study, using relapse-free survival in low-risk, node-negative patients as the primary outcome, the main objective is to compare adjuvant CMF + prednisone to no adjuvant therapy. In an associated ancillary study, we are also measuring a panel of potentially useful prognostic factors and predictors of therapeutic efficacy. Pending validation, these factors will be used in future studies to define patient eligibility or refine risk groups. Similarly, a series of ongoing aerodigestive tract chemoprevention trials include the measurement of panels of biomarkers in the hope that one or more markers can be validated for subsequent use as surrogate endpoints [8].

The objectives of such ancillary studies are two-fold: to choose and verify the "best" marker(s); and then to validate the equivalence of the surrogate endpoint and primary outcome for testing treatment effect. Careful, independent verification is extremely important. As we [9] and others [10] have found in investigating prognostic factors, exploratory analysis to optimize biomarker performance can be very misleading. Simulations and theoretical work clearly show that Type I error rates can be 8- to 10-fold higher than nominal levels in optimized analyses, and some form of internal validation, p-value adjustment, or independent validation is essential. The same is true for investigation and optimal representation of biomarkers as surrogate endpoints.

If the putative surrogate endpoints are truly capable of reducing the size or duration of prevention trials, then trial designs based on the primary outcome may allow the creation of training and validation sub-studies of the surrogate endpoint within a single trial. Otherwise, separate studies will be required. In the therapeutic setting, demonstration of treatment equivalence at a clinically relevant level often requires as many or more patients than traditional trials [11]. Similarly, statistical demonstration of the

equivalence of the surrogate endpoint amounts to showing that the intervention effect on outcome disappears when adjusted for the surrogate endpoint. In an example based on the Polyp Prevention Trial, Freedman and Schatzkin [7] showed that the number of cases required to show equivalence with reasonable precision (n = 2140) was comparable to the number patients needed for the entire trial (n = 2000). Obviously there are no short-cuts.

Based on experience with prognostic factors where many appear promising initially but few stand the test of time, it is likely that many markers will have to be studied in order to select one or a few that truly capture the effect of intervention on the development of disease. Disease heterogeneity and multiplicity of developmental pathways may make the situation even more difficult. Rather than a single causal path leading linearly to frank malignancy, a more realistic view is likely to involve multiple, possibly overlapping networks (Fig. 2). Depending on the biological effect of the intervention, individual markers may fail to fully capture the effect. In addition, different types of chemoprevention agents will almost certainly exert influence at different points, leading to the need for agent-specific markers.

Finally, surrogate endpoints can be viewed as imperfect observations of the primary endpoint. In this sense, the epidemiological concept of misclassification is relevant. If misclassification, arising as a result of using the surrogate rather than the primary endpoint, is non-differential (*i.e.*, independent of the intervention) then it is well known that the effect of the treatment will be underestimated [12]. The power of a chemoprevention trial to detect differences in outcome based on the surrogate endpoint will be reduced. In fact, the increase in sample size required to compensate for classification errors could outweigh any benefit derived from using the surrogate endpoint. In a study modelling the use of changes in the distribution of stage at diagnosis of breast cancer as a surrogate for changes in mortality in the evaluation of cancer control programs, Austin *et al.* [13] have shown that sample size requirements increased by 7% for each 1% increase in misclassification of stage. A 5% misclassification rate would necessitate a 35% increase in sample size for any trial using the surrogate endpoint. If the misclassification is differential, as would be the case if the surrogate endpoint failed to capture some important aspect of true treatment effect, then the bias in the estimate of treatment efficacy can be in either direction. While adjustment methods are available, they generally require very precise estimates of classification probabilities, and hence large sample sizes. We see again the need to fully characterize the dependence of outcome on the surrogate endpoint.

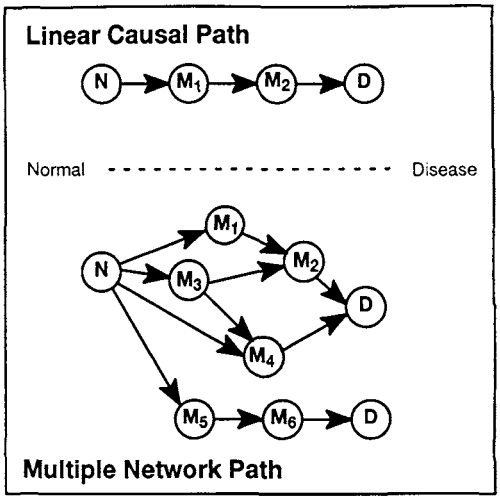

Fig. 2. Diagrammatic representation of simple linear causal pathways and network pathways.

FUTURE DIRECTIONS

Most work on the use of surrogate endpoints assumes the selection and substitution of a single surrogate endpoint for cancer incidence. In view of the heterogeneous nature of breast cancer and the likelihood that multiple, interconnected causal pathways will be required to characterize the development of breast cancer, it may be more productive to think in terms of constellations of intermediate biomarkers which, taken together, fully characterize an individual's location on the path toward carcinogenesis. The biomarker panel approach used in the aerodigestive tract prevention trials, with several markers each for proliferation, genomic damage, and differentiation, provides an interesting model [8]. Creation and validation of such multivariate endpoints will require re-design of both laboratory and transitional studies to carefully elucidate the diversity of possible causal paths, ancillary studies involv-

ing multiple endpoints as adjuncts to current and planned chemoprevention trials, and thoughtful multivariate statistical analyses.

ACKNOWLEDGEMENTS

This work supported in part by NIH grants NCI P01 CA30195, NCI P50 CA58183, and NCI P30 CA54174.

REFERENCES

1. Schatzkin A, Freedman L, Schiffman M. (1993) An epidemiologic perspective on biomarkers. J Intern Med 233:75–79.
2. Pappas PC, Lavin PT, Ross SD, Gold JE, Osband ME. (1992) Marked discrepancy between survival and tumor response endpoints in cancer clinical trials (meeting abstract). Proc Am Soc Clin Oncol 11:442–442.
3. Prentice RL. (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 8:431–440.
4. Schatzkin A, Freedman LS, Schiffman MH, Dawsey SM. (1990) Validation of intermediate endpoints in cancer research. J Natl Cancer Inst 82:1746–1752.
5. Hulka BS. (1991) Epidemiological studies using biological markers: Issues for epidemiologist. Cancer Epidem Biomark Prev 1:13–19.
6. McGuire WL. (1990) Breast cancer prognostic factors: Evaluation guidelines (editorial). J Natl Cancer Inst 83:154–155.
7. Freedman LS, Schatzkin A. (1992) Sample size for studying intermediate endpoints within intervention trials or observational studies. Am J Epidemiol 136: 1148–1159.
8. Lippman SM, Lee JS, Lotan R, Hittelman W, Wargovich MJ, Hong WK. (1990) Biomarkers as intermediate endpoints in chemoprevention trials. J Natl Cancer Inst 82:555–560.
9. Hilsenbeck SG, Clark GM, McGuire WL. (1992) Why do so many prognostic factors fail to pan out? Breast Cancer Res Treat 22:197–206.
10. Lausen B, Schumacher M. (1992) Maximally selected rank statistics. Biometrics 48:73–85.
11. Makuch R, Simon R. (1978) Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 62:1037–1040.
12. Diamond EL, Lilienfeld AM. (1962) Effects of errors in classification and diagnosis in various types of epidemiological studies. Am J Public Health 52: 1137–1144.
13. Austin DF, Fu CX, Roffers SD, Shields JM, Aubert RE. (1993) Breast cancer mortality model. Analytical Report #2. AACCR Cancer Surveillance & Control Program, Emeryville, CA 94608.
14. Freedman LS, Graubard BI, Schatzkin A. (1992) Statistical validation of intermediate endpoints for chronic diseases. Stat Med 11:167–178.
15. Schulte PA. (1987) Methodologic issues in the use of biological markers in epidemiologic research. Am J Epidemiol 126:1006–1016.