Disease Markers 20 (2004) 97–102
IOS Press

Development tracks for cancer prevention markers

Stuart G. Baker[a,*], Barnett S. Kramer[b] and Philip C. Prorok[a]
[a]*Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA*
[b]*Office of Disease Prevention, National Institutes of Health, Bethesda, MD, USA*

Abstract. We provide a general framework for describing various roles for biomarkers in cancer prevention research (early detection, surrogate endpoint, and cohort identification for primary prevention) and the phases in their evaluation.

1. Introduction

The evaluation of cancer prevention markers (biomarkers) depends both on the purpose to which they are used and the nature of the marker. The main purposes are early detection (with the goal of early intervention), surrogate endpoints (with the goal of having shorter studies) and cohort identification (of individuals with the greatest benefit to harm ratio associated with primary prevention). In cancer prevention, markers are also used to obtain better measures of nutrient intake or other exposures, but these markers are beyond the scope of this article.

2. Molecular markers for early detection

In the past, most markers for cancer screening involved changes in cell types (such as degrees of dysplasia) or single proteins in serum such as PSA or CA125. Recently there has been a large increase in the amount of data on molecular biomarkers to hopefully detect cancer early. One reason is simply the new types of biomarkers being developed, such as proteomic patterns. A second reason is the increasing number of biorepositories associated with randomized trials or cohort studies, making it possible to collect biomarker data retrospectively (e.g. Baker (1998)). This large increase in information, which sometimes goes under the name of bioinformatics, has created a challenge for cancer prevention. Without a clear purpose, bioinformatics may accomplish little. Pablo Picasso said, "Computers are useless. They only give us answers". Answers to the wrong question are of little value. It is important that that computations for evaluating biomarkers answer the appropriate question. With early detection, the appropriate question is what biomarkers are most likely to make effective triggers of early intervention. The following phases (which are applicable to both molecular markers and markers involving changes in cell type) provide a road map for answering this question.

2.1. Phase I. Molecular markers for symptomatic cancer

The easiest way to collect data for investigating molecular markers is to obtain specimens from subjects without cancer and subjects with symptomatic (clinical) cancer, or from tumor cells and non-tumor cells in the same subject. The goal is to identify markers that discriminate well between symptomatic cancer and no cancer. As discussed in Baker [4] in more detail, the recommended outcome measures are the false positive rate (proportion of non-cancers that are positive, or one minus specificity) and true positive rate (proportion of cancer specimens that are positive, or sensitivity). Each rule for classifying a marker as cancer or not is as-

*Corresponding author: Stuart G. Baker, Sc.D. National Cancer Institute, EPN 3131, 6130, Executive Blvd MSC 7354, Bethesda, MD 20892-7354, USA. Fax: +1 301 402 0816; E-mail: sb16i@nih.gov; URL: http://www.cancer.gov/prevention/bb/baker.html.

ISSN 0278-0240/04/$17.00

sociated with both a false and true positive rate. Some rules can be created by specifying different cutpoints for a particular marker. Other rules are generated by combinations of cutpoint levels among multiple markers. One can plot a point for each pair of false and true positive rates, where the horizontal axis is the false positive rate and the vertical axis is the true positive rate. The points highest and farthest to left form the optimal receiver operating characteristic (ROC) curve. Certain points on the ROC curve are of primary interest. Because the eventual goal is to identify a biomarker that is useful for cancer screening, it is important that the false positive rate be very small (e.g. 1% for breast cancer screening) to avoid unnecessary biopsies. The true positive rate should be large (e.g. 80% for breast cancer screening).

With many biomarkers (or combinations of biomarkers) under investigation, chance plays a large role in making some biomarkers or their combinations appear to perform well in discriminating between cancer and no cancer in the early phases of the study. To reduce the role of chance, we recommended that the data be randomly split into training and test samples. Then, if biomarkers (or combinations of biomarkers) with a small false positive rate and large true positive rate are identified in the training sample, we recommend definitive evaluation of their performance using an ROC curve based on the test sample. One example of the use of a test sample is a proteomic study that reported FPR=0.05 and TPR =1.00 in the test sample [25].

2.2. Phase II. Molecular markers for asymptomatic cancer

Biomarkers identified in Phase I that discriminate well between clinical cancer and non-cancer (as evidenced by the ROC curve) are investigated in Phase II to determine if they discriminate well between asymptomatic cancer and non-cancer. Although one could measure the biomarker in all asymptomatic subjects prospectively, a retrospective study using stored specimens is much more practical. Typically a biorepository is created to store specimens in a prospective study that may be unrelated to the cancer of interest. As part of a nested case-control study within the prospective study, biomarkers are measured in the stored sample from all subjects with the cancer of interest and a random sample of subjects without cancer.

Many of the approaches to evaluation are similar to those for Phase I. However, as discussed in Baker et al. [9], there are some special considerations and opportunities. One special consideration is related to the detection of cancer in the prospective trial. Ideally, for purposes of analysis, the subjects in the prospective study are those who were diagnosed with symptomatic cancer, while the sample used to measure the biomarker is obtained while the subject was still asymptomatic. If cancers in the prospective study were detected by screening of asymptomatic individuals and there was substantial overdiagnosis (i.e. detection of cancers that do not cause any medical problems in a person's lifetime), the performance results of the biomarker would not be relevant. One special opportunity with stored samples in Phase II that is not available in Phase I is the ability to investigate the performance of prediction rules based on changes in marker values over time, during the asymptomatic phase of the disease.

Another consideration at this phase is spectrum bias. This can occur if the spectrum of health conditions in the non-cancer control subjects differs from the spectrum in the target population for screening. For example, the target population for lung cancer screening may be people with a heavy cigarette smoking history. Such people may have a variety of health conditions which produce a false positive test including inflammatory lung conditions, upper respiratory infections, chronic obstructive pulmonary disease, cardiac disease, or even cancer of unrelated organ systems such as the bladder. If adequate attention is not paid to the possibility of spectrum bias, the false positive rate is likely to be much higher in practice than in Phase I or II studies.

2.3. Phase III. Molecular biomarkers as triggers for early intervention

Even if a molecular biomarker discriminates well between cancer and no cancer, when it triggers early intervention, it may not lead to a reduction in cancer mortality or an insufficient reduction in cancer mortality relative to harms such as unnecessary biopsies and work-ups. Therefore it is important to evaluate the biomarker as a trigger of early intervention. The most definitive approach is via a randomized trial with a cancer mortality endpoint [9]. Because such randomized trials are extremely large (usually with over 30,000 subjects) and costly, they should be used only for the most promising triggers for early intervention.

If the costs of the aforementioned randomized trail are prohibitive, and it is possible to delay making a definitive evaluation, we recommend observational studies as an intermediate step between a Phase II study and a randomized trial of biomarker as a trigger for early

intervention. One recently proposed approach, called periodic screening evaluation [9], rules out biomarker triggers for early intervention that would *not* likely lead to a reduction in cancer mortality in a randomized trial. Another approach is to combine before-and-after studies from various geographic regions (e.g. [13,16,26], with future work involving extensions of the methodology in Baker, Lindeman, and Kramer [10].

3. Imaging test for early detection

The results of imaging tests can be viewed as a marker for early detection. Because they involve a subjective component, the method of evaluation differs from that of molecular markers. Examples of imaging markers for the early detection of cancer are mammography and low-dose spiral CT. The evaluation of imaging markers follows the same general phases as the evaluation of molecular markers. However some details differ in Phases I and II.

3.1. Phase I. Imaging markers for symptomatic cancer

As with molecular markers, the first phase is to evaluate the ability of the markers to discriminate between subjects with clinical cancer and subjects without clinical cancer. However, unlike molecular markers, the classification of imaging markers involves a subjective element. For example, a typical classification of mammograms (normal, probably benign, indeterminate, suspicious, and highly suspicious) depends both on the image and the person reading the images. Fortunately the ROC curve can also accommodate subjectively determined categories. It is not as useful to report the false and true positive rate for a particular category because the determination of the category is fuzzy due to the subjective component. Instead, the preferred outcome measure is the area under the ROC curve for a range of small false positive rates [4] which can be thought of as an "average" true positive rate over the categories with the small false positive rates of interest.

3.2. Phase II. Imaging markers for asymptomatic cancer

If images are taken periodically in a prospective study and stored digitally for later evaluation, one can retrospectively read the images from subjects who later developed cancer and a random sample of subjects who did not develop cancer. The retrospective analysis follows the same recommendations as with the molecular markers. However, as mentioned above, due to the subjective nature of the readings, the preferred outcome measure is an area under part of the ROC curve.

Sometimes a prospective study is implemented for evaluating the performance of imaging in discriminating between asymptomatic cancer and no cancer. In the prospective study images are read, classified, and a recommendation for biopsy or no biopsy is made. Because cancer is rare in asymptomatic subjects, the prospective study needs to be large (e.g. 20,000 subjects.) However the prospective study has the advantage of a relatively short duration.

The difficulty for evaluating the prospective study is that a biopsy is performed only if the image is classified as positive, leading to a severe bias known as "confounding by indication." Therefore it is not possible to directly estimate false and true positive rates, which require knowing the number of subjects with and without cancer at the time of screening. However, because cancer is rare among screened subjects, one can obtain an estimate of the false positive rate by assuming that all subjects not biopsied are free of cancer. For estimating true positive rates, some investigators assume that all subjects negative on imaging who developed cancer within, say, one year, in fact, had undetected cancer at the time of screening. The validity of this assumption depends on the distribution of the length of time at which asymptomatic cancer is detectable on screening. See Day [14]. If classification data are collected, then, due to the subjective nature of the readings, the preferred outcome measure is an area under part of the ROC curve corresponding to small false positive rates, say 0.5% to 1.5%. This area can be used to compute an average TPR in the range of small false positive of interest. Ideally one would want an average TPR of 80%. Alternatively as this is a more definitive study than Phase I, one might base the criteria for going to Phase III on the performance of a decision to biopsy or not. In that case the numerical values of false positive rate (e.g. 1% for breast cancer screening) and high true positive rate (e.g. 80% for breast cancer screening) are analogous to the values for molecular markers.

3.3. Phase III. Imaging biomarkers as triggers for early intervention

The same considerations for evaluating molecular markers as triggers for early intervention apply to evaluating imaging markers as triggers for early intervention.

4. Surrogate endpoint markers

A surrogate endpoint is an endpoint that is obtained sooner, less invasively, or at less cost than a true endpoint for a health outcome. Surrogate endpoints have either been used to replace true endpoints or to predict true endpoints. Evaluation methodology is more controversial than with markers for early detection. Some endpoints that were thought to be good surrogate endpoints later turned out to give incorrect inference when compared to true endpoints [17]. Despite this strong caveat, we can outline various development stages and identify the gaps that need addressing. Most of the evaluation techniques can rule out poor surrogate endpoints. However, more methodological research is needed to develop criteria for deciding if a surrogate endpoint is useful in practice and thus avoid unintended harm based on incorrect assumptions. Given all the difficulties we discuss, the main use of surrogate endpoints may be in a preliminary study of an intervention to decide on whether or not to evaluate the intervention in a definitive randomized trial with a true endpoint.

4.1. Phase I. Surrogate and true endpoint in an observational study

A surrogate endpoint should be strongly associated with a true endpoint. To estimate association, data are needed on both a surrogate and true endpoints in individuals. In preliminay studies, the data will be observational. For example the strong association between human papilloma virus (HPV) infection and cervical intraepithelial neoplasia (CIN) [27] indicates the potential for using HPV as a surrogate endpoint for CIN. The challenge in this phase is deciding what level of association is sufficiently large to move to the next phase, and here the published methodology is lacking. Levels of association adequate or even high for an etiologic risk factor may not be sufficient to warrant further investigation as a surrogate endpoint since they may still explain a low proportion of variability in the true endpoint. More importantly, a high or even perfect correlation between surrogate and true endpoints does guarantee that the surrogate endpoint will provide reliable inference about the true endpoint in a randomized trial [4].

4.2. Phase II. Randomized trial of intervention known to affect true outcome

As a preliminary requirement for a surrogate endpoint to be effective, an intervention that is known to affect a true endpoint should also affect a surrogate endpoint. The best way to check is to randomize subjects to either an intervention with a known effect on a true endpoint or a control group, where the outcome is the surrogate endpoint. If the surrogate endpoint is promising for this phase, the intervention would affect the surogate endpoint in the same direction as the intervention is known to affect true endpoint. With a reasonable sample size, if there is no statistically significant difference in surrogate outcomes between randomization groups, one could likely rule out the biomarker as a good surrogate endpoint.

The dilemma is that it is often unlikely to have an intervention that is known to affect the true endpoint and be acceptable for a randomized trial. Nevertheless in some situations it is possible. For example, suppose that adenoma recurrence is a true endpoint. (In reality it is only a surrogate endpoint for colorectal cancer, but let us ignore that for now). Based on a randomized trial, calcium supplement was found to reduce adenoma recurrence [20]. Because calcium supplement did not affect rectal mucosal proliferation (surrogate endpoint) in another randomized trial [1], one can rule out rectal mucosal proliferation as a promising surrogate endpoint for adenoma recurrence.

If the outcome of this type of randomized trial is a panel of surrogate endpoint markers, it is important to adjust for multiple comparisons. The simplest, albeit conservative, approach is a Bonferroni correction: if there are k markers then statistical significance is not 5% (two-sided) but 5/k%. Although the magnitude of the effect is also important, it is not clear what size effect is needed for proceeding to the next phase.

4.3. Phase III. Randomized trials of multiple interventions with surrogate and true endpoints

The very nature and complexity of cancer pathogenesis makes the results of Phase I and II very tentative. Most cancers have multiple causal pathways and each pathway has multiple steps that can take years to evolve. If a surrogate marker occurs early in the pathway but the intervention blocks a “downstream” event, measurement of a surrogate endpoint can give a misleading answer even if the marker is in the causal chain. Likewise, if the marker is in an entirely different causal

pathway, its measurement can be misleading. This is why extrapolation of results from one particular set of surrogate and true outcomes can be hazardous. For this reason the most informative evaluation of surrogate endpoints comes from data on surrogate and true endpoints either in multiple randomized trials involving different interventions or a single randomized trial with multiple arms, each with a different intervention. Ideally the set of interventions should affect all major pathways so results are applicable to new interventions.

The methodological research in this area has focused on evaluating multiple trials with surrogate and true endpoints [19,23]. The data may be divided into a training set involving all but one trial and a test set involving the remaining trial. Based on data in the training set, a model is developed to predict the true endpoint from the surrogate endpoint. This prediction could be at a "meta-analytic" level [19,23] or based on an individual-level model (extending the methodology in Morrison 1989 and Day and Duffy, 1996 to multiple previous trials). Confidence intervals should incorporate sampling variability as well as variability among parameters in previous trials. For validation, the model is used to predict true endpoint in the test trial based on surrogate data in the test trial. The predicted true endpoint in the test trial is compared with the observed true endpoint in the test trial. However there is no guidance on how to decide when the overlap in confidence intervals warrants acceptance of the surrogate endpoint as a replacement for the true endpoint.

An alternative design, which we have not seen applied or discussed, is a single trial involving randomization to multiple interventions with surrogate and true outcomes. This design has several advantages over the multiple trial design. First it would likely take less time to implement. Second, it does not involve duplication of control groups. Third, it might be easier to set up a biorepository for implementing a nested case-control study in which the surrogate endpoint is measured in stored samples from a random sample of subjects [21].

Even if this more rigorous approach finds that a surrogate endpoint adequately predicts a true endpoint under a variety of interventions, there still could be a serious problem with using a surrogate endpoint. Most, if not all, cancer prevention trials involve both benefits and harms.A surrogate endpoint that is very accurate in predicting a true endpoint of benefit may not be very accurate in predicting harms. This could occur if the harms arise after the time of the surrogate endpoint. If trials are stopped early because of an apparently favorable change in a surrogate endpoint, a net harm associated with the intervention can be missed. For example if a chemoprevention agent were to delay or prevent polyps but cause strokes that occur between the detection of polyps and the occurrence of colon cancer, the harm from strokes would be missed if studies are stopped due to a "valid" surrogate endpoint based on polyps.

5. Cohort identification for primary prevention

Another use of biomarkers in cancer prevention is to target preventions to individuals with biomarkers that indicate a high ratio of benefits to harms, either by virtue of having greater benefits, fewer harms, or both. For simplicity of this discussion, we focus on benefits.

5.1. Phase I. Preliminary observational studies

One type of preliminary observational study is a small study in which all subjects receive the primary prevention, data are collected on biomarkers, and the outcome is a surrogate endpoint indicating response or non-response. The goal is to identify what biomarkers, if any, differ statistically significantly between responders and non-responders. Usually a very large number of markers are considered, so there is a need to adjust for multiple comparisons.

A second type of preliminary observational study is a case-control study to identify genes that substantially elevate the risk of cancer. The underlying premise is that subjects with genes indicative of a higher cancer risk have the greatest potential to benefit in absolute terms from a primary prevention. Again, if many biomarkers are under consideration an adjustment for multiple comparisons is necessary.

5.2. Phase II. Randomized trials

If a primary prevention was shown to be beneficial in a randomized trial, but the benefits did not outweigh the harms, it would be of great interest to investigate whether the benefits (i.e. reduction in cancer) would be greater for subjects with a certain gene. The problem is that testing for the gene in all subjects is not possible. Although only cases need to be tested [4,22] to estimate relative risk among subjects with the gene, relative risk is not the most clinically relevant outcome. A more relevant outcome is the age-specific absolute risk difference due to the intervention. As discussed in Baker and Kramer [26], this outcome can be estimated using

a nested case control design involving data from all subjects with cancer and a random sample of subjects without cancer in a randomized study.

If the randomized trial is in the planning stage and the focus is on a single gene, one could test for the gene in all subjects or subjects with a family history of cancer and randomize the subjects with the gene to intervention or not [5]. If many genes are involved, it would be preferable to randomize all subjects, collect specimens for a biorepository, and then perform the aforementioned nested-case control study.

6. Discussion

The purpose of this article is to outline development tracks for some markers that play a role in cancer prevention. Not all markers fit neatly into these development phases, as for example prospective studies of sputum cytology markers for early detection of lung cancer [8], but some of the same principles apply, such as using the appropriate portion of the ROC curves for evaluation. This article should help investigators understand how marker research fits into the general strategy of cancer prevention and identify the gaps in methodological work that need to be addressed in order to avoid potential unintended harms that can occur when biomarkers or surrogate endpoints are accepted prematurely.

References

[1] J.A. Baron, et al., Calcium supplements for the prevention of colorectal adenomas, *New England Journal of Medicine* **340** (1999), 101–107.

[2] S.G. Baker, Identifying combinations of cancer biomarkers for further study as triggers of early intervention, *Biometrics* **56** (2000), 1082–1087.

[3] S.G. Baker, The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer, *J. Natl Cancer Inst.* **95**(7) (Apr 2 2003), 511–515.

[4] S.G. Baker, D. Erwin, B.S Kramer and P.C. Prorok, Using observational data to estimate an upper bound on the reduction in cancer mortality due to periodic screening, *BMC Medical Research Methodology* **3:4** (2003), (06 Mar 2003).

[5] S.G. Baker and L.S. Freedman, Potential impact of genetic testing on cancer prevention trials, using breast cancer as an example, *Journal of the National Cancer Institute* **87** (1995), 1137–1144.

[6] S.G. Baker and B.S Kramer, Designing a genetic subset study for a randomized trial with a binary endpoint, *BMC Medical Research Methodology* **3**(16) (2003).

[7] S.G. Baker and B.S Kramer, A perfect correlate does not a surrogate make, *BMC Medical Research Methodology* **3**(16) (2003).

[8] S.G. Baker, B.S. Kramer and S. Srivastava, Markers for early detection of cancer: Statistical issues for nested case-control studies, *BMC Medical Research Methodology* **2** (2002), 4.

[9] S.G. Baker, B.S. Kramer and P.C. Prorok, Statistical issues in randomized trials of cancer screening, *BMC Medical Research Methodology* **2** (2002), 11.

[10] S.G. Baker, K.L. Lindeman and B.S. Kramer, The paired availability design for historical controls, *BMC Medical Research Methodology* **1** (2001), 9.

[11] S.G. Baker and M.S. Tockman, Evaluating serial observations of precancerous lesions for further study as a trigger for early intervention, *Statistics in Medicine* **21** (2002), 2383–2390.

[12] M. Buyse and G. Molenberghs, Criteria for the validation of surrogate endpoints in randomized experiments, *Biometrics* **54** (1998), 1014–1029.

[13] A.J. Coldman, N. Philips and T.A. Pickles, Trends in prostate cancer incidence and mortality: an analysis of mortality change by screening intensity, *Canadian Medical Association Journal* **168** (2003), 31–35.

[14] N.E. Day, Estimating the sensitivity of a screening test, *Journal of Epidemiology and Community Health* **39** (1985), 364–366.

[15] N.E. Day and S.W. Duffy, Trial design based on surrogate end points –application to comparison of different breast screening frequencies, *Journal of theRoyal Statistical Society A* **159**(1) (1996), 49–60.

[16] S. Duffy, et al., The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties, A collaborative evaluation, *Cancer* **95** (2002), 458–469.

[17] T.R. Fleming and D.L. DeMets, Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine* **125** (1996), 605–613.

[18] L.S. Freedman, B.I. Graubard and A. Schatzkin, Statistical validation of intermediate endpoints for chronic diseases, *Statistics in Medicine* **11** (1992), 167–178.

[19] M.H. Gail, R. Pfeiffer, H.C. Houwelingen and R.J. Carroll, On meta-analytic assessment of surrogate outcomes, *Biostatistics* **3** (2001), 231–246.

[20] P. RHolt, et al., Modulation of abnormal colonic epithelial cell proliferation and differentiation by low-fat dairy foods: a randomized controlled trial, *J. Am Med. Assoc.* **280** (1999), 1074–1079.

[21] M.D. Hughes, V. Degruttola and S.L. Welles, *Evaluating Surrogate Markers Acq Immun Def Synd* **10 S1-S8**(2) (1995).

[22] M. King, et al., Tamoxifen and breast cancer incidence among women with inherited mutations in BRCA1 and BRCA2 National Surgical Adjuvant Breast and Bowel Project (NSABP-P1) Breast Cancer Prevention Trial, *Journal of the American Medical Association* **286** (2001), 2251–2256.

[23] G. Molenberghs, et al., Statistical challenges in the evaluation of surrogate endpoints in randomized trial, *Controlled Clinical Trials* **23** (2002), 607–625.

[24] A.S. Morrison, Intermediate determinants of mortality in the evaluation of screening, *International Journal of Epidemiology* **20** (1991), 642–650.

[25] E.F. Petricoin, et al., Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet* **359** (2002), 572–577.

[26] L. Peron, L. Moore, I. Bairati, P. Bernard and F. Meyer, PSA screening and prostate cancer mortlaity, *Canadian Medical Association Journal* **166** (2002), 586–591.

[27] M.H. Schiffman et al., Human popillomavirus and cervical intraepithelial neoplasia response, *Journal of the National Cancer Institute* **85** (1993), 1868–1870.