Improving the Biomarker Pipeline to Develop and Evaluate Cancer Screening Tests

Stuart G. Baker

The biomarker pipeline to develop and evaluate cancer screening tests has three stages: identification of promising biomarkers for the early detection of cancer, initial evaluation of biomarkers for cancer screening, and definitive evaluation of biomarkers for cancer screening. Statistical and biological issues to improve this pipeline are discussed. Although various recommendations, such as identifying cases based on clinical symptoms, keeping biomarker tests simple, and adjusting for postscreening noise, have been made previously, they are not widely known. New recommendations include more frequent specimen collection to help identify promising biomarkers and the use of the paired availability design with interval cases (symptomatic cancers detected in the interval after screening) for initial evaluation of biomarkers for cancer screening.

Cancer screening involves the testing of asymptomatic persons for the presence of a precancerous lesion, an early-stage cancer, or a biomarker, such as a change in protein level, and is followed by early intervention if the test is positive. In recent years, extensive searches for new biomarkers that could be used to trigger an early intervention have given rise to a biomarker pipeline to develop and evaluate cancer screening tests, which is sometimes referred to as the phases of biomarker development (1,2). However, some important design and analysis considerations related to this biomarker pipeline have been underappreciated, insufficiently disseminated, or not previously discussed. By taking these considerations into account, researchers can improve this biomarker pipeline. The three main stages in this pipeline—identification of promising biomarkers for the early detection of cancer, initial evaluation of biomarkers for cancer screening, and definitive evaluation of biomarkers for cancer screening—are discussed in turn.

Identification of Promising Biomarkers for the Early Detection of Cancer

A biomarker test refers to either a single biomarker or a combination of biomarkers that is classified as positive or negative for the early detection of cancer. An important principle in designing a study to identify promising biomarker tests for the early detection of cancer is to consider how the study relates to the ultimate goal of using the biomarker test in cancer screening as a trigger for early intervention (3). This principle, which is reminiscent of the famous saying of architect Eliel Saarinen, “Always design a thing by considering it in its next larger context—a chair in a room, a room in a house, a house in an environment, an environment in a city plan” (4), motivates many of the following recommendations to improve studies to identify promising biomarker tests.

Evaluate Stored Specimens Early in the Pipeline

In the standard phases to evaluate biomarkers for cancer screening, an early phase involves the search for promising biomarkers in biological specimens that are collected when a person has symptomatic cancer (1,2). However, biomarkers that are identified at this phase may reflect changes that occur a long time after the initiation of cancer and hence have little value for the early detection of cancer. Therefore, whenever possible, the initial investigation of the performance of a biomarker test should be based on stored specimens that were collected from individuals with no symptoms of the target cancer at the time of specimen collection who later developed the target cancer (case patients) and from individuals with no symptoms of the target cancer at the time of specimen collection who did not develop the target cancer (control subjects). These stored specimens often come from repositories that are created in conjunction with long-term prospective clinical trials that may have a primary outcome unrelated to cancer.

A good strategy for identifying a promising biomarker test for the early detection of cancer is to randomly split the combined set of control and case specimens into training and test samples, with the selection of a biomarker test made in the training sample and the performance of the biomarker test in discriminating between case and control specimens ascertained in the test sample (5). Importantly, the biomarkers identified as promising in the training sample are directly relevant to discriminating between case and control specimens in the test sample because both the training and the test samples contain specimens from asymptomatic persons. Ideally, if a biomarker test performs well in the test sample, its performance should be evaluated in a validation sample comprising stored specimens from a different population (3,6), which argues for establishing many repositories.

Affiliation of author: Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD.

Correspondence to: Stuart G. Baker, ScD, National Cancer Institute, EPN 3131, 6130 Executive Blvd MSC 7354, Bethesda, MD 20892-7354 (e-mail: sb16i@nih.gov).

See “Funding” and “Note” following “References.”

DOI: 10.1093/jnci/djp186

Published by Oxford University Press 2009.
Consider More Frequent Specimen Collection

Little attention has been given to the frequency at which specimens should be collected and stored. For practical reasons, specimens are often collected infrequently (i.e., once or twice from a subject or patient), but depending on certain biological considerations, more frequent collections may be advisable. The optimal frequency of collecting specimens depends on the theory of carcinogenesis to which one subscribes. Under the somatic mutation theory, cancer begins with an initial mutation and proceeds through subsequent mutations. If the somatic mutation theory is correct, an investigation that includes infrequently collected specimens could identify long-term irreversible genetic changes that would lead to a good biomarker test for the early detection of cancer. Under the tissue organization field theory (7) and a related theory involving morphogenetic fields (8,9), the proximal cause of cancer is a disruption of cell communication that precedes irreversible changes associated with cancer. Biomarkers that are related to the disruption of cell communication may not detect cancer early because such disruptions may be reversible or of insufficient duration to initiate cancer. If the tissue organization field theory is correct, frequent specimen collection is recommended to identify biomarkers that arise in the small window of time bracketing the first occurrence of irreversible changes associated with early-stage cancers that are potentially amenable to early treatment.

Focus on the Optimal Part of the Receiver Operating Characteristic Curve in a Test Sample

Classification accuracy of a biomarker test is often summarized by the false- and true-positive rates in the test sample. The false-positive rate is the fraction of control subjects who are classified by the biomarker test as positive for cancer, and the true-positive rate is the fraction of case patients who are classified by the biomarker test as positive for cancer. A single biomarker or a combination of biomarkers yields a pair of false- and true-positive rates for each possible cut point designating a positive classification. A plot of true- vs false-positive rates from the test sample yields a receiver operating characteristic (ROC) curve that is useful for biomarker evaluation. The optimal slope of the ROC curve, which corresponds to the point on the ROC curve that maximizes expected benefits minus expected costs, equals the product of two ratios: the ratio of one minus the prevalence of cancer to the prevalence of cancer, and the ratio of the cost of a false-positive finding for early detection cancer to the benefit of a true-positive finding for early detection cancer (10,11). Because the prevalence of cancer is small, the optimal slope is typically steep and occurs at the leftmost part of the ROC curve, which corresponds to a low false-positive rate (Figure 1). For a promising biomarker for the early detection of cancer, there should be a moderate or high true-positive rate associated with a low false-positive rate.

Reduce the Effect of Chance

The effect of chance can be reduced by using the appropriate sample size. A training sample with 100 control specimens and 100 case specimens should be adequate even when searching through thousands of candidate markers in high-throughput platforms (12). A test sample with 110 control specimens and 70 case specimens yields reasonable precision for estimating classification accuracy with a target false-positive rate of 0.01 and a target true-positive rate of 0.80 (5), although a larger sample size would be needed for hypothesis testing related to classification accuracy (13). Evaluating the performance of the biomarker test in a validation sample involving a different population would require an additional sample with 110 control specimens and 70 case specimens.

Reduce Bias

In this context, bias refers to the incorrect estimation of classification performance regardless of sample size. Bias can be reduced by using the same blinded techniques to collect and handle specimens from the case patients and the control subjects and by adjusting for risk factors that are associated with the level of the biomarker and the likelihood of cancer (3,13–15).

An important but underappreciated bias relates to overdiagnosis, which is the detection of an early-stage cancer that, in the absence of detection, would not cause any medical problems during the person’s lifetime. When cancers are diagnosed on the basis of a positive finding in an established screening test, such as a prostate cancer diagnosis after a positive prostate-specific antigen test, some may be overdiagnosed. A biomarker that has a high true-positive rate when a large fraction of cancers are overdiagnosed may have a low true-positive rate for the relevant cases with cancers that are not overdiagnosed. To avoid this bias from overdiagnosis, case patients should be restricted to persons whose cancer was detected due to clinical symptoms (5).

Keep the Biomarker Test Simple

When the search for biomarkers involves high-throughput methods, numerous biomarkers are often included in the combination of biomarkers that comprise the biomarker test. However, after the few most discriminating biomarkers are used to create a biomarker test, the inclusion of additional markers often yields little gain in classification performance (16,17). A simple biomarker test may be preferable to a more complicated biomarker test if both
have similar classification accuracy and the former is easier to implement as a trigger for early intervention.

**Initial Evaluation of Biomarkers for Cancer Screening**

A biomarker test that is promising for the early detection of cancer will not necessarily reduce the mortality rate from cancer when used as a trigger for early intervention in cancer screening. Therefore, the biomarker test must be evaluated as part of cancer screening. Because a definitive evaluation of a biomarker for cancer screening is an expensive and long-term process that involves a randomized trial with a cancer mortality endpoint, an initial evaluation is needed to determine which biomarker test should be definitively evaluated for cancer screening.

This initial evaluation can involve the random allocation of subjects to either an established cancer screening test or a new biomarker test, and computing the difference in the fraction who are interval cases (i.e., symptomatic cancers detected in the interval after screening) between the two randomized groups (18). A new biomarker test that substantially reduces the fraction who are interval cases would be a good candidate for a definitive cancer screening evaluation (2). An appealing aspect of this design is that the results are not affected by overdiagnosis. However, although this design involves a smaller sample size and a shorter follow-up period than a cancer screening trial with a cancer mortality endpoint, it could still be difficult to implement due to the need for randomization.

However, because randomization might not be critical at this intermediate stage in the biomarker pipeline, it might be sensible to instead implement a modification of the paired availability design, which is a rigorous nonrandomized trial involving multiple before-and-after studies (19,20). Under this modification of the paired availability design, a new biomarker test is introduced at various screening centers where an established screening test has been given. Data are collected on the number of interval cases associated with screening in time periods before and after the introduction of the new biomarker test. For each screening center, the estimated effect of the biomarker test on interval cases is the difference in the fraction of interval cases between the time periods divided by the difference in the fraction of persons receiving the new biomarker test between the time periods. These estimated effects are averaged over screening centers. The main requirements for unbiased estimation are that each screening center serves a stable population with little in- or out-migration and that the criteria for symptomatic cancer detection change little from one time period to the next, in essence making each time period analogous to a randomized group. These requirements are reasonable when the duration of the study is short and with the appropriate choice of screening centers, such as those in geographically isolated locations. Hence, biases due to the lack of randomization should be small.

**Definitive Evaluation of Cancer Screening Tests**

The final stage in the pipeline is the definitive evaluation of the biomarker as a trigger for early intervention in a randomized cancer screening trial with a cancer mortality endpoint. The endpoint of overall mortality, although theoretically desirable, is not practical because the variability between the randomized groups in the numbers of deaths unrelated to cancer screening necessitates an extremely large sample size (21). For purposes of evaluating cancer screening, cancer mortality is defined as deaths from the cancer that is the target of screening, deaths that are directly caused by screening such as perforation of the colon on colonoscopy, and deaths related to treatment. To reduce bias, deaths from uncertain causes should not be preferentially attributed to cancer in persons diagnosed with cancer (22). With some types of cancer screening, such as screening for cervical cancer, an endpoint of invasive cancer might also be considered.

A typical sample size for a randomized cancer screening trial with a cancer mortality endpoint ranges from 50 to 100,000. Sample sizes are usually based on statistical power and the type I error, but with multiple trials under consideration, one could compute the sample sizes that maximize expected benefits subject to a budget constraint (23,24). After randomization, subjects are typically screened for a few years and followed up for 5–10 years. Most analyses of cancer screening trials compare the cumulative cancer mortality rates in the two randomization groups. Although the relative risk is widely reported, the absolute risk difference facilitates a direct comparison of benefits and harms between randomization groups. The following recommendations can improve the definitive evaluation of cancer screening tests.

**Adjust for Noncompliance or Contamination**

In many randomized cancer screening trials, some subjects who are assigned to screening refuse screening (i.e., noncompliance) and some subjects who are assigned to no screening receive screening outside of the trial (i.e., contamination). The estimated difference in cancer mortality rates between the two randomized groups is diluted by noncompliance and contamination. Statistical methods have been developed to estimate the effect of screening on cancer mortality in the absence of noncompliance and contamination that occur soon after randomization, which is called all-or-none compliance (19,25,26). In particular, the risk difference adjusted for all-or-none compliance equals the difference in cancer mortality rates between the two randomized groups divided by the difference in the fraction receiving screening soon after randomization between the two randomized groups. For unbiased estimation, an assumption of rationale preferences is needed, namely, that no person would receive screening if randomly assigned to no screening and no person would receive no screening if randomly assigned to screening. Also needed under all-or-none compliance is the plausible assumption that the probability of cancer mortality does not depend on the randomization group among persons who would receive the same intervention, either screening or no screening, regardless of their group assignment.

**Adjust for Postscreening Noise**

Perhaps the most important and least appreciated adjustment in the analysis of data from cancer screening trials concerns postscreening noise. Postscreening noise refers to cancer deaths that occur during the follow-up period that could not have been prevented by screening because the cancers arose during the follow-up period or were not detectable by screening before the follow-up period. The number of cancer deaths associated with postscreening noise is expected to be similar in the two randomization groups. With increasing follow-up, the number of cancer deaths associated with postscreening noise...
increases, thereby decreasing the precision of the risk difference and the relative risk and reducing the relative risk. If the data are analyzed very soon after randomization, the full effect of screening on the cancer mortality rate may not have materialized. If the data are analyzed very late after randomization, postscreening noise may have masked any effect of screening on the cancer mortality rate. According to one approach (21), the optimal time of analysis occurs when the ratio of signal to noise, that is, the estimated difference in the cancer mortality rate in the two randomization groups divided by its SE, is largest.

Investigate Possible Early Reporting of Results
Most cancer screening trials have a planned maximum follow-up time that is based on the anticipated number of cancer deaths each year. However, during the follow-up period, there may be an opportunity to report the results of a cancer screening trial sooner than planned if waiting longer would add little information because of postscreening noise. This early reporting of results allows health benefits, either the reduction in the cancer mortality rate or the avoidance of unnecessary biopsies, to reach the general population sooner. Early reporting is indicated when the optimal time of analysis, and hence the estimated effect of screening on the cancer death rate and its confidence interval, would likely change little with further follow-up. Early reporting was validated using data from randomized trials of breast cancer and lung cancer screening (27).

Estimate Harms
Because screening evaluation involves weighing benefits and harms, the amount of harm should also be estimated. One type of harm is a false-positive screen, that is, when screening is positive for cancer but no cancer is detected on further testing or biopsy. Of particular interest is the cumulative rate of false-positive screens in a periodic screening program, which can be estimated from data in cancer screening trials (28–31). Another type of harm is unnecessary therapy resulting from overdiagnosis. Overdiagnosis is likely if there are more cumulative cancer cases in the group randomly assigned to screening than there are in the control group, but the overall mortality rates are the same in the two groups (26). The fraction of persons overdiagnosed can be estimated using a statistical model for the natural history of cancer (33).

References

Funding
This work was supported by the Division of Cancer Prevention in the National Cancer Institute and the National Institutes of Health.

Note
Manuscript received November 25, 2008; revised May 26, 2009; accepted May 29, 2009.