COMMENTARY

Phases of Biomarker Development for Early Detection of Cancer

Margaret Sullivan Pepe, Ruth Etzioni, Ziding Feng, John D. Potter, Mary Lou Thompson, Mark Thornquist, Marcy Winget, Yutaka Yasui

1) Introduction

Recent developments in such areas of research as gene-expression microarrays, proteomics, and immunology offer new approaches to cancer screening *(1).* The surge in research to develop cancer-screening biomarkers prompted the establishment of the Early Detection Research Network (EDRN) by the National Cancer Institute *(2).* The purpose of the EDRN is to coordinate research among biomarker-development laboratories, biomarker-validation laboratories, clinical repositories, and population-screening programs. By coordination of research efforts, the hope is to facilitate collaboration and to promote efficiency and rigor in research.

With the goals of the EDRN in mind, the purpose of this commentary is to define a formal structure to guide the process of biomarker development. We categorize the development into five phases that a biomarker needs to pass through to produce a useful population-screening tool. The phases of research are generally ordered according to the strength of evidence that each provides in favor of the biomarker, from weakest to strongest. In addition, the results of earlier phases are generally necessary to design later phases.

Therapeutic drug development has had such a structure in place for some time *(3).* The clinical phases of testing a new cancer drug are as follows: phase 1, determinations of toxicity, pharmacokinetics, and optimal dose levels; phase 2, determinations of biologic efficacy; and phase 3, definitive controlled trials of effects on clinical endpoints. For each phase, guidelines exist for subject selection, outcome measures, relevant comparisons for evaluating study results, and so forth. Although deviations are common, the basic structure facilitates coherent, thorough, and efficient development of new therapies. A phased approach has also been proposed for prevention trials *(4,5).*

In a similar vein, we hope that our proposed guidelines or some related construct will facilitate the development of biomarker-based screening tools for early detection of cancer. Although deviations from these guidelines may be necessary in specific applications, our proposal will, at the minimum, provide a checklist of issues that should be addressed at each phase of development before proceeding to the next.

2) Objectives of Population Screening

The goal of a cancer-screening program is to detect tumors at a stage early enough that treatment is likely to be successful. Moreover, the screening tool must be sufficiently noninvasive and inexpensive to allow widespread applicability. A substance secreted by tumor tissue, not secreted by nontumor tissue, and easily and cheaply detectable in serum or urine is, therefore, an ideal biomarker because the cancer is detected specifically and noninvasively. Biomarkers, however, may be more complicated and/or indirect, involving, for example, measures of immune response to a developing tumor, hormonal changes induced by a tumor, or mass spectrometry profiles of serum protein. In this commentary, we use the term "biomarker" for cancer detection in a broad sense.

Cancer is a diverse disease, and it is unlikely that a single biomarker will detect all cancer of a particular organ with high specificity and sensitivity. Indeed, biomarkers, such as prostate-specific antigen (PSA), that purport to have high sensitivity tend to have low specificity because they do not detect cancer *per se* but rather a more general process. We note that maintaining high specificity (low false-positive rates) is a very high priority for population screening. Even a small false-positive rate translates into a large number of people subjected to unnecessary costly diagnostic procedures and psychologic stress. Thus, biomarkers need to be highly specific for cancer, and the use of several such biomarkers of cancer will likely be necessary for an overall screening program that is both sensitive and specific.

3) Five Phases of Screening Biomarker Development

We propose that biomarker development be conceptualized as occurring in five consecutive phases as depicted in Fig. 1. In this section, we outline the key objectives of each phase and discuss aspects of study design for achieving the primary aim.

3.1) Phase 1—Preclinical Exploratory Studies

The first step in the search for biomarkers often begins with preclinical studies, comparing tumor tissue with nontumor tissue. These are exploratory studies to identify characteristics unique to tumor tissue that might lead to ideas for clinical tests for detecting cancer. Immunohistochemistry and western blots have been extensively used for this purpose. More recent technology includes gene-expression profiles based on microarrays that yield information regarding expression for thousands of genes *(6),* protein expression profiles based on mass spectroscopy *(7),* and levels of circulating antibodies against thousands

Affiliations of authors: M. S. Pepe, Department of Biostatistics, University of Washington, and Fred Hutchinson Cancer Research Center, Seattle; M. L. Thompson, Department of Biostatistics, University of Washington; R. Etzioni, Z. Feng, J. D. Potter, M. Thornquist, M. Winget, Y. Yasui, Fred Hutchinson Cancer Research Center.

Correspondence to: Margaret Sullivan Pepe, Ph.D., Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195–7232 (e-mail: mspepe@u.washington.edu).

See "Notes" following "References."

of cancer antigens. The objective of a phase 1 gene-expression or proteomics study is to identify genes or clusters of genes (or proteins) that appear to be overexpressed or underexpressed in tumor tissue relative to control tissue. Organ tissue, however, cannot usually be used for clinical screening purposes because its procurement is too invasive. Thus, the development of a clinical assay based on serum levels of proteins expressed by the identified genes, say, or on serum antibody level to those proteins would be the task of the next phase. The aims of a phase 1 study are as follows:

Primary Aims

1) To identify leads for potentially useful biomarkers.
2) To prioritize identified leads.

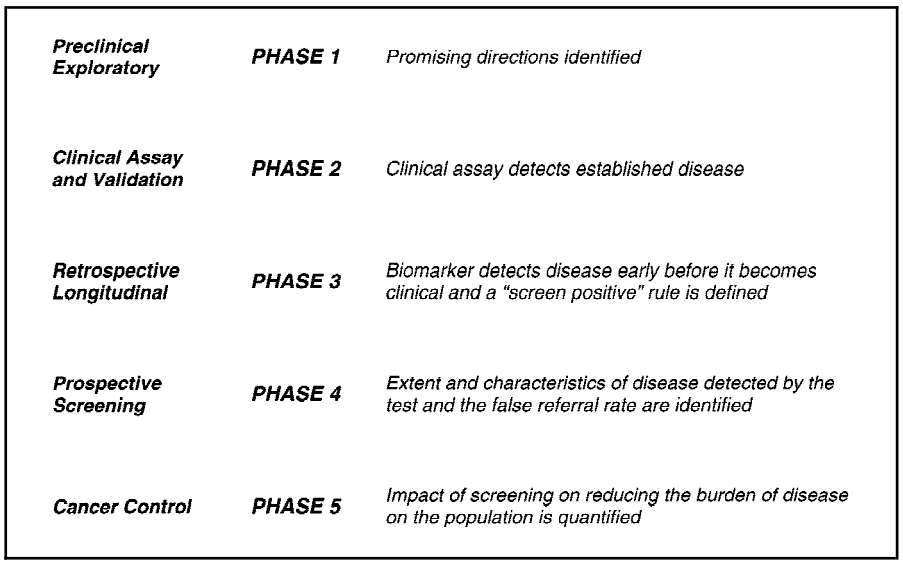

Fig. 1. Phases of biomarker development.

Specimen Selection

Tumor tissue from case subjects should be obtained at diagnosis and before treatment because treatment may interfere with the behavior of the biomarker. It seems appropriate that a wide spectrum of tumors be evaluated in this exploratory phase *(8),* with attention being paid to variability in patient demographics, histology, prognosis, stage, and mode of detection. There may be particular interest in tumors that become clinically evident only at a late stage, since earlier detection of such tumors is clearly warranted.

Nontumor specimens are derived from various sources: Noncancerous organ tissue from the cancer patient, normal organ tissue from noncancer patients, and abnormal but noncancerous tissue from noncancer patients (such as inflamed tissue or benign growth tissue) are all useful controls in phase 1 studies. Ideally, a biomarker will be evident in tumor tissue but not in nontumor tissue of the case subjects or in organ tissue from noncancer patients. Matching of control subjects with case subjects seems desirable in phase 1 studies because factors other than cancer might affect the biomarker and confound associations if such factors differ between case and control subjects. In a small study, random selection is likely to result in imbalance on some factors. Normal tissue from a subject with cancer is, by definition, perfectly matched for all factors that vary between individuals. Noncancer control subjects should be selected so that factors potentially influencing the biomarker, other than the cancer itself, are tightly matched to those of the cancer case subjects. These factors might include age, sex, race, and possibly lifestyle-related characteristics, such as smoking habits.

Primary Outcome Measures

The outcome measures, or primary data items, for analysis in phase 1 are the values of the biomarkers. In a gene-expression study, these might entail several hundred (or even thousands) of overexpressed or underexpressed species of messenger RNA (mRNA). The assays should be reliable and reproducible. Substantial variability in assay results could obscure biomarkers that are promising.

Evaluation of Study Results

For each biomarker under consideration, one needs to ascertain how well it distinguishes between case and control subjects. If a biomarker is measured on a binary scale (positive versus negative), the true-positive rate (TPR), i.e., the proportion of case subjects who are biomarker positive, and the false-positive rate (FPR), i.e., the proportion of control subjects who are biomarker positive, summarize its ability to discriminate between disease and nondisease.

Sensitivity and specificity are commonly used terms for TPR and 1 – FPR. If a biomarker result can take many values, with larger values, say, being more strongly indicative of disease, a receiver operating characteristic (ROC) curve is used *(9–11).* Fig. 2 shows data on a pancreatic cancer marker *(12).* Advantages of the ROC curve over simple frequencies and summary statistics for raw biomarker data are *(a)* that it does not depend on the scale of raw-data measurements, which greatly facilitates comparison of the discriminatory capacities of different biomarkers; and *(b)* that it displays true- and false-positive rates, quantities that are more relevant for screening purposes than the raw biomarker values themselves. Since we have argued that low false-positive rates are of interest for disease screening, that portion of the ROC curve relating to low FPRs should be the focus of data analysis.

The development of statistical algorithms for selecting promising biomarkers from a large pool of biomarkers is an active area of research *(13).* One simple approach is to rank the biomarkers on the basis of a summary statistic, such as the area under the ROC curve [or under that part pertaining to low FPRs *(14)* or other restricted region *(10)*] and to select those that rank highest.

Exploratory data analysis is an integral part of phase 1. However, spurious results due to random variation occur in exploratory data analysis. If many biomarkers are under evaluation, one or more may appear by chance alone to have good discrimination ability. Sampling variability should be assessed, and statistical cross-validation methods should be applied when possible. In addition, it is prudent to perform a well-controlled confirmatory study in phase 1 with the use of a new set of tissue specimens. New outcome measures might be chosen at the con-

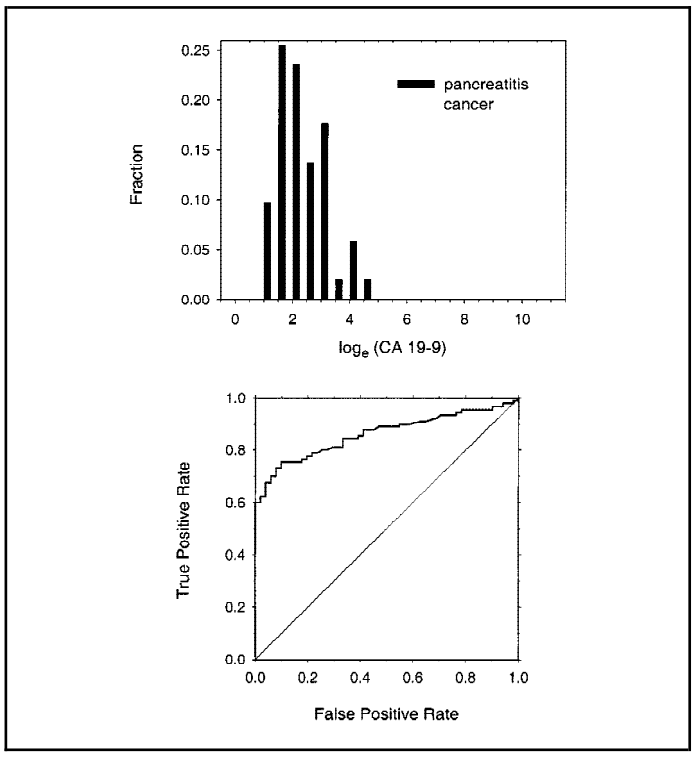

Fig. 2. Histograms and receiver operating characteristic (ROC) curve for a cancer antigen, CA 19–9, as a biomarker for pancreatic cancer *(14).* **Each point** on the ROC curve is the fraction of cancer case subjects with CA 19-9 exceeding a threshold (true-positive rate) versus the corresponding fraction of control subjects (false-positive rate). Higher thresholds yield ROC points at the lower ranges of the axes.

firmatory stage—for instance, protein expression in a study that follows initial investigation at the mRNA level.

Sample Sizes

How many specimens should be tested in phase 1? The number depends on the objective of the study and the extent of the variability of the biomarker in the study. When the objective is to select a subset of biomarkers from a pool, the following factors contribute to variability: the number and relative prevalence of the cancer subtypes among the study samples, the capacities of the biomarkers to discriminate among the different cancer subtypes, the number of biomarkers under study, the number of case and control subjects, and the statistical algorithm used to select promising biomarkers. Simple methods for recommending sample sizes, therefore, are not feasible. In particular, traditional sample size calculations that are based on statistical tests of hypotheses are not relevant. We suggest that computer simulations guide the choice of sample sizes; that is, simulation studies of hypothetical study data should be performed with guidance from investigators on biologically plausible models to generate data. By varying the numbers of case and control subjects, one can assess at what sample sizes a reasonable proportion of promising biomarkers are likely to be selected for further study.

3.2) Phase 2—Clinical Assay Development for Clinical Disease

A clinical assay based on a specimen that can be obtained noninvasively is developed in phase 2. An immune response to a protein uniquely expressed by tumor and measured with serum antibodies would be an example of such a biomarker *(15).* The clinical assay must be shown to distinguish subjects with cancer from those without cancer, in order to be considered promising for screening. Note, however, since the case subjects in a phase 2 study have established disease, with clinical biomarker assay results that are concurrent with their clinical disease, this phase does not determine if disease can be detected early with a given biomarker. The aims of phase 2 are as follows:

Primary Aim

To estimate the TPR and FPR or ROC curve for the clinical biomarker assay and to assess its ability to distinguish subjects with cancer from subjects without cancer.

Secondary Aims

1) To optimize procedures for performing the assay and to assess the reproducibility of the assay within and between laboratories. The assay should be reasonably simple, and results should be reproducible when the assay is repeated on the same tissue at the same or different laboratory, if it is to be used for widespread screening. In preparation for phase 3, the assay should also work well on stored clinical specimens.
2) To determine the relationship between biomarker measurements made on tumor tissue (phase 1) and the biomarker measurements made on the noninvasive clinical specimen (phase 2). For example, one should confirm that patients with high expression of mRNA in tissue are the same patients for whom an associated biomarker protein is measured in serum.
3) To assess factors, such as sex, age, smoking behavior, etc., that are associated with biomarker status or level in control subjects. If such factors affect the biomarker, thresholds for screen positivity may need to be defined separately for screening subpopulations to keep the FPR at a low level for each.
4) To assess factors associated with biomarker status or level in cancer case subjects—in particular, disease characteristics such as stage, histology, grade, and prognosis. Understanding the nature and characteristics of cancer that is detected with the biomarker is a key issue. A biomarker that detects cancer in early stage is more valuable than one that detects only late-stage cancer. A biomarker that misses an aggressive subset of cancers will be less valuable than one that detects such cancers.

Specimen Selection

The principles described for phase 1 selection of case and control subjects also apply to phase 2. The population from which case and normal control subjects are selected needs careful consideration. Ideally, case and control subjects would be representative of those from a target screening population. However, case subjects and control subjects with benign growths are often identified from a surgical clinic, having been referred for biopsy or surgery on the basis of suspicious clinical findings. Control subjects from such clinics may not be representative of control subjects recruited from the population because they, too, have been referred for some reason to the clinic. Control samples from a blood bank might be used; however, again, these samples may differ systematically from random samples from

the target screening population. We suggest that, although selection based on convenience may be necessary early in phase 2, final conclusions should be based on population-based studies, if possible.

Primary Outcome Measure

The result of the clinical biomarker assay is the primary data unit for analysis.

Evaluation of Study Results

Estimates of the TPR and FPR or ROC curve should be calculated. Ideally, some target values for minimally acceptable TPR and FPR, which we denote by TPR_0 and FPR_0, respectively, should be decided on before evaluation of the data. One can then evaluate statistically the joint null hypothesis H_0: $TPR \leqslant TPR_0$ or $FPR \geqslant FPR_0$, using simple tests of proportions if the biomarker assay yields a binary result. The null hypothesis is that the TPR of the biomarker is too low or that the FPR is too high. For biomarkers measured on a continuous scale, the minimally acceptable false-positive rate FPR_0 implies a decision threshold for the biomarker. The null hypothesis for the corresponding true-positive rate, H_0: $TPR \leqslant TPR_0$, can equivalently be written as H_0: $ROC(FPR_0) = TPR_0$, where $ROC(f)$ denotes the true-positive rate (ROC value) at a false-positive rate f. Adjustments for multiple comparisons will be necessary if case subjects are to be compared with multiple control groups. Moreover, if frequency matching of case and control subjects distorts the distribution of covariates from those in the target population, a statistical reweighting adjustment in estimating the true- and false-positive rates in the target population may be warranted.

Sample Sizes

The sample sizes will be determined by how precisely one needs to estimate TPR and FPR or ROC (FPR_0). One should choose desirable values for the true- and false-positive rates, which we denote by TPR_1 and FPR_1, respectively, and specify an alternative hypothesis as H_1: $TPR = TPR_1$ and $FPR = FPR_1$. The sample sizes should be large enough so that, despite random variation, a biomarker with this level of discrimination, (TPR_1, FPR_1) will have a high probability, $1 - \beta$, of rejecting the null hypothesis H_0 that operating characteristics are below target values.

3.3) Phase 3—Retrospective Longitudinal Repository Studies

Clinical specimens collected from cancer case subjects before their clinical diagnosis and compared with those from control subjects (i.e., subjects who have not developed cancer) provide evidence regarding the capacity of the biomarker to detect preclinical disease. If the levels of the biomarker in case subjects only deviate from those in control subjects close to the time of clinical diagnosis, then the biomarker shows little promise for screening. On the other hand, if the levels in case subjects reach levels distinct from those in control subjects months or years before clinical symptoms appear, then the biomarker's potential for early detection is increased. In that event, criteria for defining a positive screening result that will be used for prospective screening are defined in phase 3. The specific aims of phase 3 are as follows:

Primary Aims

1) To evaluate, as a function of time before clinical diagnosis, the capacity of the biomarker to detect preclinical disease.
2) To define criteria for a positive screening test in preparation for phase 4.

Secondary Aims

1) To explore the impact of covariates on the discriminatory abilities of the biomarker before clinical diagnosis, including demographics, disease-related characteristics, and other clinical information about the subject. If the biomarker appears to discriminate well only in certain subpopulations, this information might be used to select appropriate populations for prospective screening with the new biomarker.
2) To compare markers with a view to selecting those that are most promising.
3) To develop algorithms for screen positivity based on combinations of markers. Although earlier phases might suggest that particular combinations of markers work well together, formal algorithms for combining biomarker results can be developed only in phase 3, where the relative behaviors of biomarkers over the preclinical interval are established.
4) To determine a screening interval for phase 4 if repeated screening is of interest.

Specimen Selection

Repositories of clinical specimens, collected and stored from a cohort of apparently healthy subjects monitored for development of cancer, are used in phase 3 of the biomarker evaluation. Subjects who develop cancer are identified, as is a set of appropriate control subjects from the cohort. The composition of the study cohort should reflect the target population for screening in relation to cancer and biomarker processes. Moreover, it is important that a well-defined and appropriate protocol be used for collection, storage, and processing of specimens. The specimen collection should provide samples representative of those that would be collected from a target screening population. Note that interventions or screening practices may interfere with inference about the behavior of the biomarker. Interventions may alter the cancer or biomarker processes or both, and screening that is more intensive than is usual for the target population will affect the nature of cancer that is detected clinically and the estimated time by which diagnosis is advanced by the new biomarker.

Case subjects may be identified within the study or by linkage to cancer registries. The same criteria as in earlier phases might be used for selecting case subjects. Although multiple sequential samples are not necessary for addressing the primary aims, subjects with more specimens and a longer history of prediagnosis specimens may be preferable because they provide more information about the prediagnostic trajectory of the biomarker within individuals. However, because such sampling can bias the case group toward slow-growing cancers, random selection of case subjects should also be considered. Control subjects are defined as individuals who have not developed cancer during a given follow-up time, and we note that consideration of the length of this follow-up time can be a difficult issue. Matching on enrollment date and on compliance with the specimen-collection protocol might be useful, in addition to matching on subject-related characteristics.

Primary Outcome Measure

The result of the biomarker assay again constitutes the primary outcome.

Evaluation of Study Results: an Example

We recently analyzed PSA data from a phase 3 prostate cancer case–control study *(16)* nested in the Beta Carotene and Retinol Efficacy Trial *(17)*. The ROC curves of Fig. 3 address the key question pertaining to primary aim, item 1. They describe the capacity of a biomarker to distinguish subjects destined to develop cancer *T* years after their biomarker is measured from control subjects for various values of *T*. For both serum PSA measures—total PSA and ratio of free total PSA—discrimination decreases between case and control subjects as the time interval between specimen collection and cancer diagnosis increases for case subjects. At an FPR of 5%, the TPR for total PSA is roughly 80% at diagnosis, 60% at 2 years before diagnosis, and 40% at 4 years before diagnosis. Thus, for example, 60% of cancers could have been detected 2 years before their clinical diagnosis by using total PSA as a biomarker and allowing for a 5% false-positive rate.

Consideration of the FPR is the natural starting point for choosing a threshold that will be applied for defining screen positivity in phase 4. With FPR_0 denoting the largest acceptable FPR, the corresponding biomarker threshold can be determined with the use of data from control subjects. The time-dependent ROC curves determine corresponding TPRs at time lags before clinical diagnosis. The threshold chosen will be one that achieves an acceptable trade-off between the time-dependent TPRs and the FPR.

Sample Sizes

Three sample sizes need to be considered when designing a phase 3 study to evaluate a biomarker: the number of case subjects, the number of control subjects, and the number of clinical specimens per subject. The sample sizes should ensure that, for each preclinical time lag of interest (e.g., 1 year, 2 years, 4 years), there are enough specimens from control subjects and from case subjects taken close to those intervals so that biomarker accuracy can be estimated with sufficient precision, as described earlier, for phase 2.

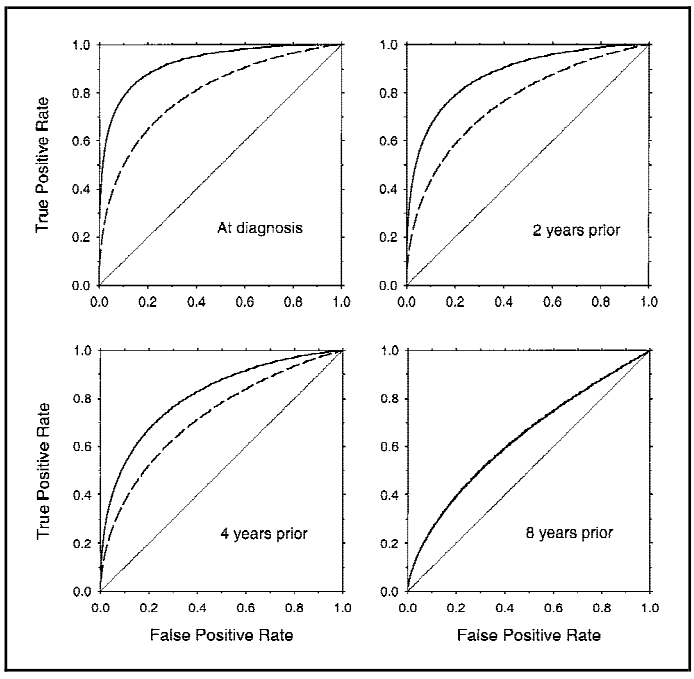

Fig. 3. Receiver operating characteristic (ROC) curves for total prostate-specific antigen (PSA) and ratio of free to total PSA at various times before diagnosis, calculated with the use of a retrospective longitudinal nested case–control study. **Solid line** = total PSA. **Dashed line** = PSA ratio.

Although the ROC analyses of Fig. 3 do not require longitudinal data, we suggest that a series of biomarker values over time from a relatively small number of subjects are preferable to more subjects contributing fewer measurements each. The longitudinal data will allow assessment of within-subject variability and more powerful comparisons of time-specific ROC curves, thereby providing better statistical evaluation of time trends in the ability to discriminate between control subjects and case subjects.

3.4) Phase 4—Prospective Screening Studies

The retrospective phase 3 study determines whether tumors can be detected early before clinical diagnosis with the biomarker, but it does not establish the stage or nature of the cancer at the time that it can be detected. In a prospective screening study, the screen is applied to individuals, and definitive diagnostic procedures are applied at that time to those screening positive. Thus, the number and nature of cases detected with the screening tool are determined in phase 4, as are the numbers of subjects falsely screening positive and referred for work-up.

It should be noted that, in contrast to studies in phases 1, 2, and 3, which are conducted on retrospective analysis of stored specimens, studies at this stage involve screening people and lead to diagnosis and treatment. Ethical considerations, therefore, play a greater role. Moreover, because disease prevalence in cohort studies is low (since subjects are not selected on the basis of their disease status), large sample sizes are required for such studies. Adequate planning and piloting of studies are, therefore, very important in phase 4. Finally, we note that population screening can be implemented as a one-time event (prevalence screen) or repeatedly at intervals over time. Although repeated screening over time may yield the best benefit, for simplicity, we focus on prevalence screening here. The specific aims of phase 4 are as follows:

Primary Aim

To determine the operating characteristics of the biomarker-based screening test in a relevant population by determining the detection rate and the false referral rate.

Secondary Aims

1) To describe the characteristics of tumors detected by the screening test—in particular, with regard to the potential benefit incurred by early detection. Patients with tumors that cannot be successfully treated or those with tumors that regress spontaneously or tumors that are very slow growing derive little benefit from screen detection.
2) To assess the practical feasibility of implementing the screening program and compliance of screen-positive subjects with work-up and with treatment recommendations. Understanding factors that can result in poor compliance, for example, can lead to improvements in the screening program.

3) To make preliminary assessments of the effects of screening on costs and mortality associated with cancer.
4) To monitor tumors that occur clinically but that are not detected by the screening protocol.

Subject Selection

The screening cohort should be from a population that might be targeted for a screening program. One might impose additional inclusion/exclusion criteria based on disease risk, compliance potential, and characteristics identified as being related to improved test performance in phase 3. Such constraints serve to provide a setting where a useful test has the best chance of showing itself to be so, even if this benefit comes at the expense of some generalizability. This approach seems prudent, since a negative conclusion in this phase is unlikely to be expanded on with further studies, whereas a positive conclusion will stimulate further research. An unscreened control arm should be considered in phase 4 (secondary aim, item 3). This provides preliminary data for the design of a phase 5 randomized trial in which cost and mortality are the primary outcomes.

Primary Outcome Measure

Ideally, along with the screening test result, a definitive test for the presence of cancer would be available for all study subjects in phase 4 so that true- and false-positive rates of the screening tests could be calculated. However, procedures for definitive testing, such as surgical biopsy, are invasive and ethically can be undertaken only for subjects who screen positive. Thus, negative screens are not identified as true or false negatives, which means that neither the disease prevalence nor the true- and false-positive rates are identifiable from phase 4 studies. The primary outcome measure for phase 4, which we call the detection outcome, is one of three categories: screening test positive and disease confirmed, screening test positive and disease not confirmed, and screening test negative. These measures yield detection rates and false-referral rates that are defined formally below.

Evaluation of Study Results

The detection rate is the proportion of screened subjects who test positive and have the disease, and the false-referral rate is the proportion who test positive but do not have the disease. Statistical techniques for binomial proportions yield confidence intervals for these parameters. Binary regression methods are used to evaluate factors that affect the rates. Methods for comparing screening tests while adjusting for covariates and for pooling data across multiple study sites have been described previously *(18).*

Sample Size

The size of the study cohort will be driven by how precisely the detection rate and the false-referral rate are to be estimated. In a study comparing two screening protocols with possibly different biomarkers, a hypothesis regarding relative performance can drive the sample size. This has been described for both equivalence and superiority studies in an unpublished manuscript (by T. A. Alonzo, M. S. Pepe, and C. S. Moskowitz), where equivalence studies evaluate if detection and false-referral rates for the different screens are sufficiently close that the screens are considered to be equivalent and superiority studies determine if one screen has better performance than the other.

3.5) Phase 5—Cancer Control Studies

The final phase addresses whether screening reduces the burden of cancer on the population. Even if the biomarker detects disease early, there are several reasons why it might not have an overall benefit for the screened population. These include *(a)* ineffective treatments for screen-detected tumors, *(b)* poor compliance with the screening program or difficulties with implementing the program in community practice, *(c)* prohibitive economic or morbidity-associated costs of screening itself and of the diagnostic work-up of subjects who falsely screen positive for disease, and *(d)* the overdiagnosis of cancers that, in the absence of a screening program, would not be detected and would in some cases regress *(19).* If population screening is to be justified, there should be little doubt about its net benefit. Unfortunately, for some of the screening tests currently in place, we do not yet have firm evidence of such a benefit. Prostate cancer screening is a case in point *(20,21).* The aims of phase 5 are as follows:

Primary Aim

To estimate the reductions in cancer mortality afforded by the screening test.

Secondary Aims

1) To obtain information about the costs of screening and treatment and the cost per life saved.
2) To evaluate compliance with screening and work-up in a diverse range of settings.
3) To compare different screening protocols and/or to compare different approaches to treating screen-detected subjects in regard to effects on mortality and costs.

Subject Selection

Subjects should be randomly selected from populations in which the screening program is likely to be implemented if it is found to be successful. A standard parallel-arm randomized clinical trial is ideally undertaken in phase 5, with one arm consisting of subjects undergoing the screening protocol and the other arm consisting of unscreened subjects.

Primary Outcome Measure

Time from entry into the study until death is the key outcome measured. Some studies consider mortality only from the specific screened cancer to be of primary relevance, whereas other studies consider death from any cause to be relevant.

Evaluation of Study Results

Survival analysis methods for censored data are used to compare the study arms with regard to overall mortality and cause-specific mortality. Recently, methods for comparing costs and quality of life for randomized trials have been developed *(22,23).*

Sample Size

To detect a 20% reduction in cause-specific mortality with 80% power at the .05 two-sided significance level, standard calculations *(24)* indicate that 650 deaths would need to be observed. Studies in this phase are, therefore, huge undertakings, since the population mortality from a specific cancer is extremely low over reasonable study periods. Few such studies have actually been undertaken *(25,26).*

Computer-modeling methods can be used as a preliminary step for assessing the necessity for and composition of phase 5 studies *(27,28).* Such models synthesize what is known about the natural history of the cancer and biomarker, treatment effects on tumor and survival, cost information, and population behavior. Screening practices are then superimposed on the model to assess the potential benefits and costs associated with screening. These models necessarily use information gained from phases 2, 3, and 4 in their construction and might be considered as an intermediate phase 4.5 between phase 4 and phase 5. The approach of Etzioni et al. *(29),* which models PSA levels in healthy men and in men with prostate cancer, uses estimates of sensitivity and specificity from phase 3 studies and cancer detection rates from phase 4 studies to compare annual and biannual PSA testing. Although such phase 4.5 modeling strategies can provide insight, only a randomized phase 5 study can provide conclusive evidence on the actual impact of screening.

4) Discussion

Not all biomarkers will need to progress consecutively through the five phases of evaluation outlined here. For example, mass spectrometry to identify a given protein in serum as a biomarker might begin in phase 2. A clinical assay that potentially detects only early-stage cancer might skip phase 2 in favor of phase 3. Moreover, insights provided by studies in later phases might prompt development of an alternative biomarker that would, again, need evaluation at early phases. Although the process is not necessarily linear, we believe that the conceptual structure provided by the five phases is nevertheless useful for planning and coordinating biomarker research.

Our proposal is intended to provide guidance but not a rigorous structure paralleling the existing structure for therapeutics. Indeed, the same considerations do not apply exactly, particularly in early phases where therapeutic studies involve patient care, while biomarker studies do not. The involvement of government regulatory agencies, therefore, may not be required until phase 4.

Nevertheless, research groups and funding agencies do need formal guidelines for biomarker development, and we hope that our proposal will be helpful in this regard. One additional important step is to formulate criteria for when a biomarker can reasonably progress from one phase of development to the next. With limited funding and specimen resources (phase 3 repositories for rare cancers are particularly precious), the research community should agree on such criteria, so that resources can be allocated in a sensible and fair fashion. Such criteria could include specification of minimally acceptable true- and false-positive rates for population-screening tests. Acceptable rates will vary with the cancer and the context in which the biomarker is to be applied and will undoubtedly require multidisciplinary panels of experts for their definition.

Such recommendations will greatly facilitate study design. This commentary has identified additional issues to be addressed for the design of biomarker studies. Choices of case and control subjects, for example, are complex and need careful consideration. Statistical methodology needs development. Indeed, the statistical issues are quite different from more classic fields for biostatistics, namely, therapeutic and epidemiologic studies. For example, methodology for sample-size calculations is lacking, particularly for phase 1, 2, and 3 studies, as are algorithms to combine the results of multiple biomarkers for detecting disease.

Finally, we note that the five-phase structure that we have formulated was derived in part from our exposure to the many research proposals of the EDRN. We hope that our proposal will serve as a foundation for dialogue that will lead to improvement in the efficiency and rigor of biomarker research in EDRN and in the broader scientific community.

References

(1) Henson DE, Srivastava S, Kramer BS. Molecular and genetic targets in early detection. Curr Opin Oncol 1999;11:419–25.

(2) Srivastava S, Kramer BS. Early detection cancer research network [editorial]. Lab Invest 2000;80:1147–8.

(3) International Conference on Harmonisation E9 Expert Working Group. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Stat Med 1999;18:1905–42.

(4) Greenwald P. New directions in cancer control. Johns Hopkins Med J 1982;151:209–13.

(5) Greenwald P. Epidemiology: a step forward in the scientific approach to preventing cancer through chemoprevention. Public Health Rep 1984;99:259–64.

(6) Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, et al. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 1999;238:375–85.

(7) Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. Electrophoresis 2000;21:1164–77.

(8) Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134:587–94.

(9) Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York (NY): Academic Press; 1982.

(10) Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. Biometrics 2000;56:1082–7.

(11) Pepe MS. Receiver operating characteristic methodology. J Am Stat Assoc 2000;95:308–11.

(12) Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76:585–92.

(13) Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. Tech Rep #576. Berkeley (CA): Department of Statistics, University of California Berkeley; 2000.

(14) McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–5.

(15) Sahin U, Tureci O, Schmitt H, Cochlovius B, Johannes T, Schmits R, et al. Human neoplasms elicit multiple specific immune responses in the autologous host. Proc Natl Acad Sci U S A 1995;92:11810–3.

(16) Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. Med Decis Making 1999;19:242–51.

(17) Thornquist MD, Omenn GS, Goodman GE, Grizzle JE, Rosenstock L, Barnhart S, et al. Statistical design and monitoring of the Carotene and Retinol Efficacy Trial (CARET). Control Clin Trials 1993;14:308–24.

(18) Pepe MS, Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. Biostatistics. In press 2001.

(19) Woods WG, Tuchman M, Robison LL, Bernstein M, Leclerc JM, Brisson LC, et al. A population-based study of the usefulness of screening for neuroblastoma. Lancet 1996;348:1682–7.
(20) Kramer BS, Gohagan JK, Prorok PC. Is screening for prostate cancer the current gold standard?—"no." Eur J Cancer 1997;33:348–53.
(21) The International Prostate Screening Trial Evaluation Group. Rationale for randomised trials of prostate cancer screening. Eur J Cancer 1999;35: 262–71.
(22) Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. Biometrics 1997;53:419–34.
(23) Zhao H, Tsiatis A. A consistent estimator for the distribution of quality adjusted survival time. Biometrika 1997;84:339–48.
(24) Schoenfeld D. The asymptotic properties of nonparametric-tests for comparing survival distributions. Biometrika 1981;68:316–9.
(25) Kramer BS, Gohagan J, Prorok PC, Smart C. A National Cancer Institute sponsored screening trial for prostatic, lung, colorectal and ovarian cancers. Cancer 1993;71(2 Suppl):589–93.
(26) Gohagan JK, Prorok PC, Kramer BS, Cornett JE. Prostate cancer screening in the prostate, lung, colorectal and ovarian cancer screening trial of the National Cancer Institute. J Urol 1994;152(5 Pt 2):1905–9.
(27) Ross KS, Carter HB, Pearson JD, Guess HA. Comparative efficiency of prostate-specific antigen screening strategies for the prostate cancer detection. JAMA 2000;284:1399–405.
(28) Urban N, Drescher C, Etzioni R, Colby C. Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. Control Clin Trials 1997;18:251–70.
(29) Etzioni R, Cha R, Cowen ME. Serial prostate specific antigen screening for prostate cancer: a computer model evaluates competing strategies. J Urol 1999;162(3 Pt 1):741–8.

NOTES

Supported by Public Health Service grants GM5443805 (National Institute of General Medical Sciences) and CA8636801 (National Cancer Institute), National Institutes of Health, Department of Health and Human Services.

We thank Nora Disis, Garnet Anderson, and Nicole Urban (University of Washington, and Fred Hutchinson Cancer Research Center, Seattle) for their helpful discussions, Early Detection Research Network investigators whose research proposals stimulated and guided our work, and Sandy Walbrek, Lian Schmidt, and Gary Longton (Fred Hutchinson Cancer Research Center, Seattle, and University of Washington) for their help with preparing the manuscript.

Manuscript received December 11, 2000; revised May 15, 2001; accepted May 17, 2001.