Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design

Margaret S. Pepe, Ziding Feng, Holly Janes, Patrick M. Bossuyt, John D. Potter

Research methods for biomarker evaluation lag behind those for evaluating therapeutic treatments. Although a phased approach to development of biomarkers exists and guidelines are available for reporting study results, a coherent and comprehensive set of guidelines for study design has not been delineated. We describe a nested case–control study design that involves prospective collection of specimens before outcome ascertainment from a study cohort that is relevant to the clinical application. The biomarker is assayed in a blinded fashion on specimens from randomly selected case patients and control subjects in the study cohort. We separately describe aspects of the design that relate to the clinical context, biomarker performance criteria, the biomarker test, and study size. The design can be applied to studies of biomarkers intended for use in disease diagnosis, screening, or prognosis. Common biases that pervade the biomarker research literature would be eliminated if these rigorous standards were followed.

J Natl Cancer Inst 2008;100:1432–1438

With the emergence of new molecular biotechnologies, the intensity of biomarker research has increased greatly, yet the scientific rigor of biomarker research lags far behind that of therapeutic research (1–3). In therapeutic research, the randomized placebo-controlled, blinded clinical trial takes the central role in pivotal evaluation of a new therapy. Standards for the conduct of such studies have been agreed upon internationally (4) (www.ich.org), and books and journals have been devoted to nuances of their design and analysis. Analogously, in this commentary, we consider key aspects of design for pivotal evaluation of a biomarker's capacity to correctly classify a subject's outcome (ie, classification accuracy), in which a subject's outcome may be his or her current disease status or a future event for him or her. We propose a prospective-specimen-collection, retrospective-blinded-evaluation (PRoBE) design in which biologic specimens are collected prospectively from a cohort that represents the target population that is envisioned for clinical application of the biomarker. Specimens and clinical data are collected in the absence of knowledge about patient outcome. After outcome status is ascertained, case patients with the outcome and control subjects without it are selected randomly from the cohort and their specimens are assayed for the biomarker in a fashion that is blinded to case–control status. Although every biomarker study has its own special considerations, as does every randomized clinical trial, we propose that elucidation of the key design issues in this commentary will help move the field toward standards of practice.

Biomarkers are developed for many different purposes, including for classification and prediction, as surrogate outcomes in clinical trials, as measures of toxic or preventive exposures, or as a guide to individual treatment choice (5). In this commentary, we consider only the first category, which includes diagnostic, screening, and prognostic markers. A diagnostic marker is used in people with signs or symptoms to aid in assessing whether they have a condition. A screening marker is used in asymptomatic people to detect a disease or condition at an early stage. A prognostic marker is used in subjects diagnosed with a condition to predict subsequent outcomes, such as disease recurrence or progression. The PRoBE design is intended for all of these applications.

Development of a biomarker is a process that begins with biomarker discovery, is followed by rigorous evaluation of classification accuracy, and then terminates with the evaluation of the impact of the biomarker on clinical outcomes (6,7). Multiple studies may be involved in each stage. For example, in the discovery stage, a sequence of studies may be performed to identify biomarkers from among a large pool of candidates, to validate these individual markers in independent samples, and to optimally combine markers from a panel. We focus on the intermediate stage, after all discovery work is completed. We assume that a well-defined biomarker, or marker combination, is to be definitively evaluated for its classification accuracy in a specific clinical application. If it can be shown that the marker has acceptable classification accuracy, the marker should move to the final stage of evaluation, in which the net benefit to patients is assessed by incorporating effects of clinical interventions recommended for patients on the basis of their biomarker results. Clearly, before biomarker results are used in making medical decisions for individuals, it is crucial to know how accurately the marker classifies or predicts their outcomes.

Affiliations of authors: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (MSP, ZF, HJ, JDP); Department of Clinical Epidemiology, Academic Medical Center, Amsterdam, The Netherlands (PMB).

Correspondence to: Margaret S. Pepe, PhD, Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B500, Seattle, WA 98109-1024 (e-mail: mspepe@u.washington.edu).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djn326

Components of the PRoBE Study Design

The PRoBE study design includes four key components. These components relate to clinical context and outcomes, criteria for measuring biomarker performance, the biomarker test itself, and the size of the study. Items pertaining to each of the components are listed in Boxes 1–4.

Clinical Context

For what population and in what clinical setting is the biomarker intended? The context for clinical application should drive the study design (Box 1). Once the context has been defined, a random cohort of subjects from the target population is enrolled, pertinent clinical data are collected, and biologic specimens are collected and stored. A rigorous protocol for subject enrollment and specimen collection ensues. Generalizability of study results is an important consideration in pivotal evaluations, a concept that is well appreciated in therapeutic clinical trials. This consideration motivates the design of a protocol that is simple enough for general use, that includes several institutions in the study, and that has inclusion and exclusion criteria that provide an adequately heterogeneous population for the clinical application.

The outcome or condition that the biomarker is to classify must be defined and the procedures for evaluating it must be specified. These procedures may simply involve follow-up, or they may be invasive and/or costly. If procedures do not exist to measure the outcome of interest, study objectives may need to be modified to use an alternative clinically relevant but measurable outcome (8).

Box 1. Components of design relating to clinical context.

Clinical Application

- Define the target population and clinical setting intended for use of the biomarker.
- Define subject inclusion and/or exclusion criteria and process for enrollment.
- Define the setting for specimen collection.
- Ensure adequate generality in the population studied.

Outcome

- Define the outcome of interest.
- Specify procedures for ascertaining and measuring the outcome.
- Ensure prospective specimen collection before outcome ascertainment.

Case–Control Status

- Describe case patients (seek positive biomarker results in case patients) and subsets of case patients that are of interest.
- Describe control subjects (seek negative biomarker results in control subjects) and subsets of control subjects that are of interest.
- All subjects in the population must fit into a case or control category.

Selection

- Random selection of case patients and control subjects.
- Consider matching of control subjects to case patients on factors related to the biomarker *only if* scientific questions of interest can be addressed with matched data. Be aware of scientific limitations that result from matching.

The purpose of the biomarker is to distinguish those with a bad outcome, whom we call case patients, from those with a good outcome, whom we call control subjects. Case patients are those in whom we expect the biomarker to be positive, and control subjects are those in whom we expect the biomarker to be negative. Sometimes a positive biomarker result is defined as the absence or low level of that biomarker. Subgroups of case patients and control subjects may be of interest. For example, histology and disease stage can define subgroups of cancer case patients. Benign disease and normal healthy organ tissue define two subtypes of control subjects. All subjects in the target population must fit into a precisely defined case or control category.

The design requires that random selections be made from the population for each category. In our experience, random selections of the relevant case patients and control subjects can be achieved only with a prospective cohort of subjects from the target population, with collection and storage of specimens before determination of outcome. After outcome data become available (ie, when case or control designations are determined), random sets of case patients and control subjects are selected retrospectively and their specimens are retrieved from storage.

Classic confounding arises when case patients differ from control subjects on factors that are related to the biomarker and those factors are themselves predictive of disease. For example, case patients may be older than control subjects and, if the biomarker varies with age, some of the apparent difference in biomarker values between case patients and control subjects may simply be due to age discrepancies. Choosing control subjects to match case patients on such factors eliminates this sort of confounding. However, there are major disadvantages to matching that are not always appreciated (9). First, matched control subjects are no longer representative of the control population, making interpretation of false-positive rates problematic. Second, a simple analysis that compares matched control subjects with case patients can attenuate biomarker performance; that is, performance in the matched study as a whole appears worse than in subpopulations or strata in which matching factors are constant. Third, a matched study requires a covariate-adjusted analysis (10), which is conceptually an analysis that stratifies according to covariates. This is a disadvantage because a stratified analysis is more complicated to implement and interpret than an analysis that does not stratify by covariates. Interestingly, a matched design is most efficient for this stratified (ie, covariate-adjusted) analysis.

It should be noted that examining marker performance within covariate strata is not the same as including covariates in a statistical model for the outcome (9,10). The latter approach is concerned with performance of the combination of covariates and markers for classifying outcome. A serious problem with matching is that it actually precludes direct evaluation of the combination of markers and covariates. As a consequence, one cannot evaluate the increment in performance gained by combining a marker with the covariates compared with the covariates alone for classification. In summary, one must carefully consider whether a covariate-stratified

measure of performance is of primary interest in the clinical application. If it is, then matching is recommended because of its statistical efficiency. However, in many settings, a covariate-stratified measure of performance will not be of primary interest, and the major complexities in the analysis that are introduced by matching indicate that it be avoided.

Performance Criteria

What do we want the biomarker to achieve? Performance criteria must be set to provide a yardstick for measuring the success or failure of the biomarker. The PRoBE design revolves around determining whether these criteria are met (Box 2).

Performance measures and acceptable levels for these measures depend on clinical context. Consider first the diagnostic setting, in which the biomarker is to identify people with disease as being positive. The true-positive rate and the false-positive rate are the typical performance measures of interest. Also known as the sensitivity, the true-positive rate is the proportion of diseased people correctly detected as having disease by use of the marker. The false-positive rate (which equals 1 − specificity) is the proportion of nondiseased people incorrectly detected as having disease by use of the marker. Minimally acceptable values for both the true-positive rate and false-positive rate must be agreed upon at the design stage of the study. Current medical practice and effects of subsequent procedures or interventions resulting from positive and negative marker results impact on target values for true-positive rate and false-positive rate.

In a diagnostic study, the consequences of missing a case patient with invasive cancer may be fatal, which argues for a high true-positive rate. For example, a biomarker might be developed to guide women with suspicious lesions for breast cancer to undergo biopsy examination or not (for details of this study conducted by the Early Detection Research Network, see Supplementary Material, available online). Women with invasive cancer that is, under current protocols, detected with a biopsy examination should continue to be recommended for biopsy examination (ie, a very high true-positive rate is required). However, under this current practice, the false-positive rate is 100%, in the sense that all women with suspicious lesions who do not have invasive cancer are also subjected to biopsy examination. Study investigators consider that a reduction in the false-positive rate of even 25% would be beneficial because it would result in 25% fewer women undergoing unnecessary biopsy examination. In other words, a false-positive rate of 75% is considered minimally acceptable by study investigators.

For general population screening, in contrast, the false-positive rate must be very low to avoid huge numbers of people undergoing unnecessary costly medical procedures. It has been argued that for ovarian cancer screening, the false-positive rate should not exceed 2%. Because the goal of general population screening is to detect disease early, the proportion of case patients detected at some relevant time before the appearance of clinical disease is the appropriate true-positive rate performance measure; that is, the true-positive rate is a function of the time lag between marker measurement and subsequent diagnosis of disease that would occur in the absence of screening. For example, detecting even 20% of invasive ovarian cancers at least 1 year before clinical diagnosis would be enormously beneficial if such cancers could be successfully treated at that stage. Thus, the minimally acceptable false-positive and true-positive rates in ovarian cancer screening might be a false-positive rate of at most 2% and a true-positive rate (1 year before clinical diagnosis of invasive disease) of at least 20%.

Performance criteria may vary with subgroups of case patients and control subjects. For example, the false-positive rate for women with normal ovaries should be at most 2%, but a much larger false-positive rate may be acceptable in control women with benign ovarian disease. One might require a higher true-positive rate for a disease histology that is likely to be successfully treated than for a disease that is not.

Performance criteria for prognostic markers bear similarities to those for screening markers. For example, the time between biomarker measurement and outcome must be considered. A biomarker may be more sensitive to outcomes, such as disease recurrence, that occur soon after marker measurement than to those that occur later. Moreover, it may be more important to identify subjects with subsequent events who are likely to be successfully treated at the time the marker is measured. Patients with events occurring very soon after marker measurement may not benefit. The false-positive rate is the fraction of control subjects with false-positive results. How do we define control subjects for prognostic markers? A landmark time T after biomarker measurement may be chosen as 1 year or 5 years, with control subjects

Box 2. Components of design relating to performance criteria.*

True- and False-Positive Rates

- Define TPR as the proportion of case patients with positive results.
- Define FPR as the proportion of control subjects with positive results.
- Does time between obtaining the specimen and the occurrence of an outcome impact on criteria for defining case subjects and control subjects? If so, provide corresponding time-dependent TPR and FPR definitions.
- Are there subgroups of case patients and control subjects that are of interest? If so, will TPR and FPR be calculated separately for each subgroup? Which subgroups are of primary interest?

Minimally Acceptable Performance

- What are minimally acceptable values (or ranges) for the key TPR and FPR parameters in this clinical application?

Comparisons

- Does a classification method currently exist?
- What is the performance of that method?
- Will the biomarker alone be compared with the current classifier or will it be combined with the current method (ie, head-to-head comparison or evaluation of increment in performance)?
- What are target levels for comparative performance of methods?

* TPR = true-positive rate; FPR = false-positive rate.

defined as those who are event free at time T. Subjects who die of other causes before time T or who have other catastrophic events are included in the main control group in this approach. An alternative approach is to consider those subjects as a second control group and to calculate two false-positive rates, one for subjects event free at time T (the main control group) and one for subjects who have other events before time T (the secondary control group).

Sometimes biomarkers and/or predictors already exist for the clinical application (eg, CA-125 for ovarian cancer screening). Comparisons with existing biomarkers must be part of the pivotal classification accuracy study, and minimally acceptable improvements in performance must be specified. When there are difficulties or costs associated with existing markers, the goal may be to replace them with new markers. For example, pathologist assessment of nuclear grade of a ductal carcinoma in situ lesion is a marker for subsequent development of invasive breast cancer (11). One would like to replace this marker with a biomarker that is cheaper, more reliable, and more transportable than expert pathology review. Head-to-head comparisons are appropriate in this setting. However, when existing biomarkers and/or predictors are easily obtained, the objective may be to assess the increment in performance that is achieved by adding the new marker to them. For example, the TRANSBIG prognostic breast cancer study (12) found a relatively small increment in performance by adding the 70-gene signature to readily available data on clinical factors. Finally, if a marker is already part of standard clinical practice, it may be impossible to evaluate the inherent accuracy of a new marker. For example, because prostate-specific antigen testing is routine in the United States, one can now evaluate only certain types of improvements that can be achieved by combining new markers with prostate-specific antigen for prostate cancer screening (13) but not the performance of new markers used alone without prostate-specific antigen.

The Biomarker Test

What is the biomarker? It is defined in part by the biologic specimen, procedures and timing for specimen collection, processing, and storage, all of which must be detailed in the study protocol (Box 3). For example, in the diagnostic breast cancer study, blood is drawn preoperatively and centrifuged at 4°C within 5 hours of collection, serum is removed by pipetting, and aliquots are stored at −80°C. The PRoBE design ensures that these procedures are blinded to the patient's outcome status and to any information related to the outcome that is not available at the time of specimen collection. Retrieval of the specimen (eg, a serum aliquot) and the assay procedure itself must also be defined and blinded to outcome-related information. Blinding is a key component of the PRoBE design. Appropriate labeling of stored specimens ensures blinding. Only after the study is completed is the blinding broken so that outcome data can be linked with biomarker data.

Ideally, the assay used in the pivotal study should be the assay that is intended for general use. However, development of a commercially available assay is not always practical before the pivotal study. The initial study may therefore use a research assay, with the recognition that, if the study is positive and an alternative assay is developed for widespread use, the alternative assay must be evaluated further, preferably with the same specimens that were used in the pivotal study.

Box 3. Components of design relating to the biomarker.

Procedures

- Specify procedures for specimen collection, processing, storage, and retrieval.
- Specify assay procedures and how results are reported.

Blinding

- Are mechanisms in place to blind specimen handling, assay, and reporting of results to outcome status?

Combination

- Is the biomarker data to be combined with other information on the patient in the intended clinical application, including other clinical information, other markers, and previous measurements of the biomarker in the patient?
- The specific algorithm for calculating the combination must be defined (it cannot be developed during evaluation).

Other Biomarkers and Predictors

- If other biomarkers or predictors will be combined or compared with the study biomarker, describe in detail protocols and procedures for obtaining these data.
- Provide assurance that procurement of these items is blinded to patient outcomes.

It is now widely appreciated that the assessment of biomarker performance must be separated from biomarker discovery. In discovery research, if a biomarker is selected from a set of candidates because of its apparent good performance, its performance in those samples is biased in an overoptimistic direction. This is a statistical phenomenon that reflects elements of random variation in the particular samples chosen, specimen handling, and assay procedures. If the analysis were repeated with different specimens, the results would vary. The biomarkers that perform best in one dataset might not have the best performance in another. To estimate performance without bias, an independent dataset is ideal. Therefore, in the pivotal evaluation, the topic of this commentary, the marker is defined in advance and no selection of markers is involved.

A biomarker test may be defined as a combination of several biomarkers and possibly other predictors. The specific algorithm to combine the biomarker values into a score should be defined in advance of the pivotal evaluation; that is, we regard development of a combination of several biomarkers as part of discovery and not part of the PRoBE design. Statistical techniques such as cross-validation or bootstrapping (14) can sometimes be used to simultaneously discover and evaluate a marker combination, but these techniques require that all steps involved in developing the combination score be completely defined in advance, which is a tall order in practice. We and others (1,15) consider that instead a separate independent dataset is necessary to evaluate classification accuracy. This dataset may be obtained by splitting a large dataset into two components, one for discovery—the training dataset—and one for pivotal evaluation—the test dataset.

A profile of biomarker values over time may be more indicative of a subject's outcome status than a single measurement. Accordingly, the biomarker test may be defined by an algorithm that combines the

subject's historical and current biomarker values. For example, change from the average of two previous annual measurements could define the biomarker test result. The schedule for specimen collection must give rise to sufficient data for doing the calculation. In addition, because the manner in which the calculation is done amounts to defining how the current and historical biomarker data will be combined, the combination algorithm must not be derived during the pivotal evaluation but rather defined in advance of it.

Study Size

Conclusions must be drawn from the study. A positive conclusion is that the marker meets minimally acceptable performance criteria and a negative conclusion is that it does not (Box 4). In practice, we calculate a confidence interval (or region) for the performance measure (or measures), which is a set of plausible values for the measure given the data, and draw a positive conclusion if all values in the interval are at least minimally acceptable. For example, in the diagnostic breast cancer study, the biomarker performance measure is the false-positive rate that corresponds to a true-positive rate of 0.98—that is, the proportion of noncancers that are biomarker positive when we set the biomarker threshold to ensure that 98% of invasive cancers are positive. We would draw a positive conclusion if, for example, the 95% confidence interval for the false-positive rate was 0.50 to 0.60 because it indicates that while maintaining biopsy recommendations for at least 98% of invasive breast cancers only 50%–60% of noncancer lesions will continue to be recommended for biopsy examination. Proportions in this range are even better than the minimally acceptable value of 75% that was specified in the design.

At the design stage, when only pilot data or other information are available, we anticipate a desirable performance level for the biomarker and make sure that there is a high chance (or power) that a positive conclusion will be drawn from the study, if the true performance of the biomarker in the population is as good as is anticipated. These considerations give rise to procedures for sample size calculations. Note that a positive conclusion is expected only if the marker's performance is better than minimally acceptable (ie, at a desirable level). In statistical jargon, minimally acceptable performance constitutes the null hypothesis that we wish to rule out, whereas the anticipated desirable performance level constitutes the alternative hypothesis. We provide details of sample size calculations in the Supplementary Material (available online).

Box 4. Components of design relating to study size.

Null Hypothesis

- Recall minimally acceptable performance criteria (Box 2).

Alternative Hypothesis

- Define anticipated performance levels.
- Provide rationale preferably with evidence from pilot data.

Sample Size

- Calculate case and control sample sizes (see Supplementary Material, available online).
- Plan for prospective collection from a cohort until sufficient numbers of case patients and control subjects are enrolled.

Early Termination

- Plan for early termination of the study if appropriate.

For an inherently dichotomous biomarker that is either positive or negative, one specifies null (FPR_0, TPR_0) and alternative (FPR_1, TPR_1) values for the pair of performance measures, namely the false-positive rate and the true-positive rate. For a continuous biomarker, if some established threshold exists to define a biomarker result as positive, then again null and alternative values for the false-positive and true-positive rates will be specified for the dichotomized marker. More often, however, it will make sense either to set the false-positive rate at a minimally acceptable value, FPR_0, and to estimate the corresponding biomarker threshold and true-positive rate from the pivotal study, or to set the true-positive rate at a minimally acceptable value, TPR_0, and to estimate the corresponding biomarker threshold and false-positive rate from the pivotal study, as in the breast cancer study (Supplementary Material, available online). For the former, null (TPR_0) and alternative (TPR_1) true-positive rates are specified and then sample sizes are calculated. In addition, one must ensure that the actual false-positive rate associated with the estimated threshold is close enough to the target false-positive rate value, FPR_0, which places further constraints on the number of control subjects, as described in the Supplementary Material (available online). For the latter, sample sizes are based on specified null (FPR_0) and alternative (FPR_1) false-positive rates, and further requirements on the number of case subjects are made to ensure that the actual true-positive rate associated with the estimated threshold is close enough to the target value, TPR_0.

Sample size formulas that are detailed in the Supplementary Material (available online) specify the numbers of case patients and control subjects required. However, specimen collection is performed for a cohort. The formulas will therefore be used in conjunction with projected prevalence or incidence rates in the cohort to calculate total numbers of subjects to be enrolled. Adjustments may be necessary if the actual rates are found to be different from those projected.

It is sometimes reasonable to terminate a study early if the analysis of partially accumulated data indicates that the biomarker has poor performance. Data monitoring is commonplace in therapeutic research, in which ethical concerns motivate early termination (16). In biomarker research, evaluation often begins after the cohort is assembled and specimens are stored; therefore, ethical concerns do not motivate early termination. However, preservation of specimens and resources is important. Therefore, if initial results show clearly that a biomarker has poor performance, the study should terminate. Otherwise, the study should continue so that estimates of performance are refined. Standards of practice for study design in therapeutic research (16) stipulate that early termination rules be specified in advance. The same practice should be followed in biomarker research. One simple rule is to terminate at a preplanned data monitoring point if the 95% confidence interval for biomarker performance is below minimally desirable levels. One must be cognizant that allowing studies to terminate early causes bias in estimates of biomarker performance from studies that do not actually terminate early. Statistical methods to adjust for this bias are available (17).

Alternative Designs and Strategies

The most common bias in biomarker research involves systematic differences in subject selection and/or specimen collection between case patients and control subjects. For example, specimens collected from case patients at a treatment center will differ from those collected from healthy control subjects at a blood donation center because of such factors as differences in specimen processing protocols, stress levels, and medication use. The prospective uniform nature of specimen collection for all subjects from a single cohort in the PRoBE design eliminates these systematic biases by ensuring that specimens for case patients and control subjects are collected in exactly the same way.

Another common problem is that the population or clinical setting that is studied is not the setting for which the biomarker is intended. Performance of a biomarker in one setting may not reflect performance in the setting of interest. The PRoBE design avoids this extrapolation bias (18) by requiring that the clinical application be defined and that the study cohort be a random sample from the target population. Inclusion of several institutions in the study increases confidence that the results generalize across institutions.

Simple retrospective case–control studies are notorious for spectrum bias (18). A classic example is when selected case patients tend to have more severe or well-documented disease and selected control subjects are especially healthy, leading to overoptimistic estimates of biomarker performance. The PRoBE design avoids spectrum bias by identifying all subjects in the cohort as either case patients or control subjects and drawing randomly from the subgroups. Another problem with retrospective studies is that knowledge of the subject's outcome status may affect the interpretation of an assay result or the care with which the specimen is handled. This bias is avoided in the PRoBE design by storing specimens before outcome ascertainment and by blinding specimens for retrieval and assay procedures.

In strictly prospective studies, the biomarker value is ascertained for all subjects in a cohort and outcome status is determined subsequently. These studies are also subject to problems. First, they cost more because all samples are assayed instead of only a case–control subset. Second, ethical problems arise when the biomarker value is known but there is uncertainty about how it should affect patient care. Overtreatment is one such ethical concern that has been realized in the context of prostate cancer screening with prostate-specific antigen testing. The retrospective component of the PRoBE design avoids this ethical dilemma. Third, if outcome ascertainment is expensive or invasive, subjects with certain biomarker values may be more likely than those with other values to have the outcome ascertained. Incomplete ascertainment of outcome introduces verification bias, which typically inflates both true- and false-positive rates. Fourth, knowledge of the biomarker value may influence aspects of outcome determination, also leading to bias. The PRoBE design avoids all of these biases by ascertaining the outcome in a uniform manner for all subjects in the cohort and by timing biomarker measurement to occur after outcome assessment.

A major concern in biomarker research is overfitting bias. This bias occurs when a biomarker combination is evaluated with the same dataset that was used to develop it. By requiring completion of all discovery work before the pivotal evaluation, including development of marker combinations, the PRoBE design avoids overfitting bias. When the pivotal evaluation study is constituted as the test set derived from a larger study that is split into training and test components, the sample size calculations for the PRoBE design pertain to the size of the test component only. The threshold for marker positivity may or may not be defined in advance of the PRoBE study. If the threshold is derived from clinical settings that differ from the target setting or is derived from small studies, it is unwise to use that threshold. In contrast to previous approaches (19), the PRoBE design accommodates estimation of the threshold in its sample size recommendations. Moreover, our approach to sample size calculation guarantees a certain power to rule out tests with unacceptable performance. This feature differs from approaches that are based simply on estimating performance with specified precision (19).

Discussion

In this commentary, we have presented guidance on the design of a biomarker accuracy study. Many of the principles have been highlighted elsewhere (18,19) but a comprehensive design has not, to our knowledge, been detailed heretofore. The design is ideal but strict adherence to it may be difficult or impossible in some circumstances. For example, the target population for screening markers is the general population but subjects that enroll in research studies may not be representative of the general population. The design requires specification of minimally acceptable values for measures of biomarker performance such as true- and false-positive rates. However, it does not provide guidance on how to arrive at these values. Expertise in techniques of medical decision making (20) should play a major role in developing performance criteria that biomarkers should meet.

The design issues outlined in this commentary also have implications for biomarker discovery studies (1). To avoid bias and make best use of resources, discovery studies should use key elements of the PRoBE design, including randomized selection of case patients and control subjects from a well-defined prospective cohort that is relevant to the intended clinical application, rigorous protocols that precisely define data items and procedures to measure them, and mechanisms to ensure that biomarker and outcome assessments cannot influence each other. Nested case–control studies as described in this commentary would improve the quality of discovery research and increase the chances that truly valuable biomarkers will undergo definitive evaluation. One should ideally perform the pivotal PRoBE evaluation study for biomarkers that show promise in discovery studies that use the same clinical context and population. Simultaneous discovery and evaluation of the performance of a marker or marker combination can be undertaken by using a PRoBE design and randomly splitting the dataset into a training set for discovery and a test set for evaluation (19).

We propose the PRoBE design to evaluate diagnostic, prognostic, and screening biomarkers. In a previous study (6) pertaining only to screening markers, we outlined five phases for biomarker development. We observed that discovery studies often use convenient samples that do not satisfy the criteria described in this commentary. However, biomarker discovery has been plagued by false discoveries. Therefore, we now strongly encourage investigators to use population–science principles in the design of biomarker discovery studies too (21). The five-phase paradigm also proposed that the phase 2 study, which analyzes specimens collected from case patients at the time of their clinical diagnosis, could precede the pivotal

phase 3 study, which uses specimens collected from case patients before clinical diagnosis. The rationale was that phase 2 studies can provide preliminary evidence to convince owners of repositories to part with precious preclinical specimens for phase 3 studies that concern early detection of disease. However, the current research environment is much more conducive to biomarker research than it was in the past, and in some settings preclinical specimens are readily available. In these settings, one should skip phase 2 and use preclinical specimens for discovery work. This strategy is in keeping with the PRoBE design principle of using specimens that are directly relevant to the clinical application intended for the biomarker.

Repositories of specimens that can be used to evaluate biomarkers will greatly assist research in this field (22). In addition to having specimens available for researchers, repositories facilitate combining and comparing markers that are proposed by different groups and/or at different times, which allows research to proceed in a unified rather than a piecemeal fashion. Use of existing repositories from large observational or prevention studies is enabling research on markers for screening. Creation of additional biorepositories is crucial as the discovery and evaluation of biomarkers for use in clinical medicine becomes a national research priority. The value of repositories for biomarker evaluation relies on careful attention to the design of studies that will be conducted with the stored specimens. In creating repositories, we urge adherence to the elements of the PRoBE design as a way of maximizing their value.

References

1. Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol.* 2007;60(12):1205–1219.
2. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. *BMJ.* 2002;324(7335):477–480.
3. Hernandez-Aguado I. The winding road towards evidence based diagnoses. *J Epidemiol Community Health.* 2002;56(5):323–325.
4. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Stat Med.* 1999;18(15):1905–1942.
5. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol.* 2005;23(9):2020–2027.
6. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst.* 2001;93(14):1054–1061.
7. Baker SG, Kramer BS, McIntosh M, Patterson BH, Shyr Y, Skates S. Evaluating markers for the early detection of cancer: overview of study designs and methods. *Clin Trials.* 2006;3(1):43–56.
8. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med.* 1999;18(22):2987–3003.
9. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics.* 2008;64(1):1–9.
10. Janes H, Pepe M. Adjusting for covariates in studies of diagnostic, screening or prognostic markers: an old concept in a new setting. *Am J Epidemiol.* 2008;168(1):89–97.
11. Kerlikowske K, Molinaro A, Cha I, et al. Characteristics associated with recurrence among women with ductal carcinoma in situ treated by lumpectomy. *J Natl Cancer Inst.* 2003;95(22):1692–1702.
12. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst.* 2006;98(17):1183–1192.
13. Shaw P, Pepe M, Alonzo T, Etzioni R. Methods for assessing improvement in specificity when a biomarker is combined with a standard screening test. *Stat Biopharm Res.* In press.
14. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer-Verlag; 2007.
15. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol.* 2005;23(29):7332–7341.
16. Ellenberg S, Fleming T, DeMets D. *Data Monitoring Committees in Clinical Trials.* New York: John Wiley and Sons; 2002.
17. Pepe M, Feng Z, Longton G, Koopmeiners J. *Estimating Sensitivity and Specificity from a Phase 2 Biomarker Study That Allows for Early Termination.* UW Biostatistics Working Paper Series; 2007.
18. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press; 2003.
19. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: statistical guidelines for nested case-control studies. *BMC Med Res Methodol.* 2002;2:4.
20. Hunink M, Glasziou P, Siegel J, et al. *Decision Making in Health and Medicine. Integrating Evidence and Values.* Cambridge: Cambridge University Press; 2001.
21. Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. *Pharmacogenomics.* 2004;5(6):709–719.
22. Brenner DE, Normolle DP. Biomarkers for cancer risk, early detection, and prognosis: the validation conundrum. *Cancer Epidemiol Biomarkers Prev.* 2007;16(10):1918–1920.

Funding

National Institutes of Health (UO1 CA086368 to Z.F., M.S.P., and H.J.; RO1 GM054438 to M.S.P. and P.M.B.).

Notes

The authors shared full responsibility for conducting the research reported in this paper.

Manuscript received December 19, 2007; revised August 8, 2008; accepted August 11, 2008.