Minireview

Joint National Cancer Institute-Food and Drug Administration Workshop on Research Strategies, Study Designs, and Statistical Approaches to Biomarker Validation for Cancer Diagnosis and Detection

Padma Maruvada and Sudhir Srivastava
Cancer Biomarkers Research Group, Division of Cancer Prevention, National Cancer Institute, NIH, Bethesda, Maryland

Abstract

Cancer remains the second leading cause of death in the United States, in spite of tremendous advances made in therapeutic and diagnostic strategies. Successful cancer treatment depends on improved methods to detect cancers at early stages when they can be treated more effectively. Biomarkers for early detection of cancer enable screening of asymptomatic populations and thus play a critical role in cancer diagnosis. However, the approaches for validating biomarkers have yet to be addressed clearly. In an effort to delineate the ambiguities related to biomarker validation and related statistical considerations, the National Cancer Institute, in collaboration with the Food and Drug Administration, conducted a workshop in July 2004 entitled “Research Strategies, Study Designs, and Statistical Approaches to Biomarker Validation for Cancer Diagnosis and Detection.” The main objective of this workshop was to review basic considerations underpinning the study designs, statistical methodologies, and novel approaches necessary to rapidly advance the clinical application of cancer biomarkers. The current commentary describes various aspects of statistical considerations and study designs for cancer biomarker validation discussed in this workshop. (Cancer Epidemiol Biomarkers Prev 2006;15(6):1078–82)

In spite of remarkable advances in cancer research, it remains the second leading cause of death in the United States. Successful cancer treatment depends not only on improved therapies but also on improved methods to assess an individual’s risk of developing cancer and to detect cancers at early stages for early intervention. Biomarkers for early detection of cancer enable the screening of asymptomatic populations. The Biomarkers Definitions Working Group convened by the NIH and the U.S. Food and Drug Administration (FDA) in 1999 has suggested various definitions for biomarkers, clinical end points, and surrogate end points. Several workshops have since been convened to address the use of biomarkers in cancer detection, diagnosis, and treatment (1). Participants in these workshops have emphasized the need for proper designs and conduct of clinical trials and the need for databases for such trials.

Cancer research has reached a strategic inflection point, enabling researchers to generate a wealth of critical data. The current challenge is to understand how these data and related technologies can be applied to clinical use. Cancer biomarkers must therefore be critically evaluated for their clinical applications. Several recent papers have described various aspects of the biomarker development process (2-4). For the purpose of this article, “validation” refers to the confirmation of accuracy, reproducibility, and precision or effectiveness of biomarkers in detecting the intended end points [preneoplastic lesions, incidence, etc. (5)]. However, the approaches for validating biomarkers have yet to be addressed clearly. In an effort to delineate the ambiguities related to biomarker validation and related statistical considerations, the National Cancer Institute (NCI), in collaboration with the FDA, conducted a workshop in July 2004 entitled “Research Strategies, Study Designs, and Statistical Approaches to Biomarker Validation for Cancer Diagnosis and Detection.”

Experts from the statistical, epidemiologic, and clinical communities deliberated on current statistical designs and discussed approaches to biomarker validation for cancer diagnosis and detection in a 2-day workshop. This article summarizes the discussions, critically evaluates the existing approaches, and provides the recommendations of the participants.

Performance Metrics of Biomarkers

A perfect biomarker for cancer prediction among asymptomatic populations would be able to yield a test (classification rule) that is either positive or negative with 100% sensitivity and specificity. Sensitivity indicates the proportion of individuals who test positive for a given biomarker and reflects the true-positive rate. Specificity is the proportion of normal individuals who yield a negative test and reflects false-positive rate. Both true-positive and false-positive rates reflect benefits and adverse effects, respectively, related to screening by a biomarker. For a given biomarker for cancer detection, true-positive and false-positive rates should be estimated from subjects with and without cancer and summarized using receiver operating characteristic curves, which play a central role in evaluating tests for the early detection of cancer (6-9). When validating early detection biomarkers as triggers of early intervention or for the evaluation of cancer screening,
it is necessary to estimate benefits and adverse events (e.g., unnecessary biopsies) in a particular study or a screening program (10). It is important to use mortality related to the disease as the end point rather than overall mortality in assessing the efficacy of the marker. The most commonly used epidemiologic approaches to measure performance characteristics are observational studies and case-control studies. In observational studies, the results suffer from self-selection bias; that is, subjects who receive screening have differing risks of cancer from those not tested. Case-control studies and mathematical models that combine variables from different sources increase the potential for self-selection bias, whereas periodic screening evaluation and a paired self-control ability to decrease this potential (10-14). These approaches are further biased by high-throughput, high-dimensional genomic and proteomic data—a potential source of candidate biomarkers. This has shifted the original concept of "one marker-one disease" to "multiple markers-one disease." A new approach therefore is warranted to avoid chance (something that happens unpredictably without discernible human intention or observable cause), bias, and overfitting, that is, to incorporate multivariable correlations in a statistical model.

When investigating the performance of a classification rule based on multiple markers, the problem of overfitting can be avoided by selecting a classification rule based on a training sample and estimating marker performance in a separate test sample. Because of noise associated with high-dimensional data, the top few features that differ between cases and controls should be identified up front and classification rules for combinations of these features should be investigated. For example, one could consider reestimating the top 20 features (genomic and/or proteomic) in another application versus the top 20 features that work together as joint behaviors.

Dr. Martin McIntosh (Fred Hutchinson Cancer Research Center, Seattle, WA) suggested that many approaches may be available to estimate classifiers that satisfy optimal criteria. But the choice of method should depend on practical needs, such as modest sample sizes and the need to control for observational biases. For this reason, he favors using logistic regression or other semiparametric binary regression approaches among the theoretically justifiable methods. In practice, discovery and validation platforms are rarely applicable in the clinic for monitoring the joint behavior of markers; thus, one needs to consider the need for optimizing a marker panel early on during discovery or validation. For example, the proteins that work best together when measured by matrix-assisted laser desorption ionization are not the same that will work best together when measured by an ELISA. Because of this, suggestions were made that researchers consider the entire development process to recognize the risks inherent in rejecting a marker in the discovery phase for one with inferior performance when considered alone solely because it may work well with other markers selected. Furthermore, criteria to help identify biomarkers, such as investigating their function and performance in subgroups based on stage, histology, and survival, must be developed.

Whenever decisions must be made about which markers to advance to subsequent study phases in the selection and validation processes, biology should drive the choice because, statistically, there is no justification to optimize the complex methods when the function of these markers and methods may change downstream. Small changes may have large downstream effects; if a selection decision is made too early, key markers may be overlooked. For instance, five genes that change subtly may make more difference than one gene that undergoes dramatic alteration. Moreover, the scale of a particular change is less important than the biological significance of the change. However, a large sample size is necessary to correctly identify multiple markers that work together rather than find chance combinations that apparently do well. In preliminary performance studies, the longitudinal behavior of markers in controls should be investigated to determine marker stability over time, as this characteristic is a good indicator for the retrospective performance phase. Final performance evaluation should be based on an external independent sample.

**Strengths and Weaknesses of Longitudinal and Cohort-Based Designs: A “Piggybacking” Approach through Treatment and/or Prevention Trials**

Common errors in study designs for screening include lack of a prospective statement of hypotheses and the potential application of the biomarker, lack of a clear protocol for patient selection and specimen collection, inadequate accounting of heterogeneity among patients, inattention to multiple testing, and differential specimen handling between cases and controls. Such problems can be addressed if a prospective study is designed with a verifiable endpoint–based screening protocol. Well-defined standardized specimen protocols (emerging technology needs, standards for frequency of collection, and processing and storage) with clear human subject processes (appropriate flexibility in research scope and coordination of multiple institutions) should be developed with appropriately targeted outcome data collections. Participants discussed strategies for choosing a series of time points for early detection in the setting of an observational cohort without screening for disease in its initial design. Because the period between the time a disease detected by a new test and the time a disease diagnosed clinically may vary—for early detection varies by disease—frequent blood collection beyond the original study as a main objective is difficult to justify. Additionally, specimens from existing cohorts should be used only when the lead time of the test fits the time of specimen collection. Biomarkers should be assessed by a masked observer to avoid bias. Some of the problems associated with patient heterogeneity, such as chance and bias, can be addressed by enrolling adequate numbers of patients in the study and randomizing intervention assignments. In addition, the experimental design should avoid bias from time-, laboratory-, or experiment-based factors (such as masking, randomizing order, and blocking of samples for laboratories).

Recommendations were made for piggybacking biomarkers already proven for their accuracy and performance characteristics in preceding phases of randomized trials. The control groups from the large trials can serve as longitudinal cohorts to get a prospective assessment of the relationship with biomarkers of subsequent disease. It is possible to use permutation tests to compare intervention groups with respect to multiple biomarker changes observed between groups. Biomarker validation backed up by randomized trials will yield more convincing results, remove bias, and balance predictive factors (known and unknown). Factorial designs have also been proposed to piggyback biomarkers on trials with one or more endpoints in one single study. For example, in the Physicians’ Health Study in which 22,000 physicians were evaluated for the effects of aspirin on mortality due to cardiovascular disease and β-carotene on cancer incidence. Using $2 \times 2$ factorial designs, it was possible to assign physicians to one of four groups taking aspirin placebo, aspirin plus β-carotene placebo, aspirin placebo plus β-carotene, or aspirin plus β-carotene. The aspirin component of the study was terminated early because of significant reduction in myocardial infarctions, whereas the β-carotene study was continued for unaffected results by aspirin. This approach allows addressing two separate questions relating to entirely different diseases in a single study and yields unaffected study
results for both of the study objectives (15). Some additional specific examples illustrating the piggyback approach are provided below.

In the first example, it was noted the Prostate Cancer Prevention Trial, in which 18,882 men were randomized to finasteride or placebo for 7 years and required to undergo biopsy at the end of the study, offered unique research opportunities. These included independent confirmation of disease progression, markers that predict recurrence or progression early, existent funded clinical trial infrastructure, collected covariates, and relatively little competition for samples, thereby providing a unique cohort for validating new biomarkers. By using prostate-specific antigen profiles in the Prostate Cancer Prevention Trial and recurrence times from 1,011 patients treated >7 years, prostate-specific antigen was determined to be an early predictor of recurrence of prostate cancer via a study design that had no verification bias (16, 17). Yet, there is clearly a need for analytic methods to optimize inference that is subject to design constraints. Moreover, in a piggyback approach the relationships should be made as early as possible to provide a strong justification for a particular study, embed correlative studies into the design phase, and assemble analytic techniques.

A second example is the Women’s Health Initiative, a randomized trial designed to study the effects of dietary modification, hormone therapy, calcium, and vitamin D on disease outcomes in >160,000 women initiated in 1992. Initially, the study was designed to use the specimens to explain intervention effects in the randomized clinical trial, examine disease mechanisms, identify/confirm biological risk factors, develop risk strata, and describe the natural history of disease biomarkers. It was noted that the study offered unique cohorts for a variety of biomarker validation with well-defined specimens (type, collection times, and volumes), outcomes (adequate time to accumulate a sufficient number of events and quality of outcome data), study population (relevance of biomarkers to a specific population), clinical practice (availability of screening and complementarity of a biomarker to existing modalities), and consent and Health Insurance Portability and Accountability Act authorization offer. In addition, 26 studies have been approved that use Women’s Health Initiative blood specimens, and 17 of these feature principal investigators who are not Women’s Health Initiative investigators.

### Trials for Biomarker Validation

For validating cancer early detection biomarkers, it was suggested that there should be two trials: one for identification and another for validation of biomarkers in a prospective manner. All potential markers must be identified in the first trial and confirmed prospectively in the second trial; otherwise, the test set becomes the training set. The odds ratio is a way of comparing whether the probability of a certain event is the same for two groups and is often used to evaluate the success of a cancer biomarker in a trial. Current study designs involve planning a single study to observe a large effect (odds ratio, 30) followed by a phase II study with a smaller effect (odds ratio, 3). It was suggested that it may be better to carry out more smaller multisite trials with a larger study population to observe a smaller effect (odds ratio, 10) followed by a phase II study with yet smaller reduction in the effect (odds ratio, 5). But, this paradigm makes the initial biomarker discovery more expensive. It may be justifiable if the initial investigation is comprehensive (e.g., genome-wide single-nucleotide polymorphisms, cDNA microarrays, or protein profiles). One example of such design was presented by Dr. Ross Prentice (Fred Hutchinson Cancer Research Center) for an ancillary genome-wide single-nucleotide polymorphism study to search for, and validate, cancer and cardiovascular disease single-nucleotide polymorphisms. The initial investigation involves 1,000 colon cancer cases and 2,000 matched controls followed by subsequent validation studies with smaller numbers of cases and controls (18).

The importance of modeling and how cancer biomarkers may be used as auxiliary variables in a seamless phase II/III trial design was also discussed. Dr. Don Berry (M. D. Anderson Cancer Center, Houston, TX) stated that conventional drug development paradigms contain a lag time of 9 to 12 months between phases II and III. However, with biomarkers as auxiliary variables, a drug-versus-placebo phase II study carried out at select centers, each enrolling, say, 10 to 20 patients monthly, could significantly reduce this lag time. Such a predictive prototype of biomarkers are encouraging in phase II, the trial can be expanded to phase III, carried out at many centers that enroll higher numbers, say, >40 patients monthly. In a single trial, survival data from both phases can be combined in the final analysis. Frequent updating of data evidence allows judgments about accrual and continuation of the trial. Such an adaptive design allows fewer patients to be enrolled, enables a smooth transition between phases II and III, and uses data from all patients to assess phase II end point and the relationship between the biomarker and survival (19, 20).

Dr. Don Berry noted further that, when using longitudinal markers (e.g., CA-125 in ovarian cancer), data available from the trial are used to model the relationship over time between the biomarker and survival, depending on therapy. By calculating predictive distributions for each patient and using covariates, the seamless phase II/III model can be applied. Such an approach enables key decisions in trial design, including adding or discontinuing study arms or changing doses. However, it has been suggested that, to use the auxiliary variables nonparametrically, a strong relationship between variables is necessary to yield useful results from intermediate information and final outcomes.

Using flexible genomic drug trial design scenarios for the purpose of population stratification, Dr. Sue Jane Wang (FDA, Rockville, MD) discussed how genomic biomarkers can be developed for drug response. She proposed five phases of genomic biomarker development for drug response. Briefly, the genomic biomarkers are explored for disease severity detection followed by clinical assay and validation of established disease severities in early phases of genomic biomarker development. In mid-phases, genomic biomarkers that detect drug toxicity or drug response should be identified through retrospective (longitudinal) evaluation. Clinical confirmation of its predictability is to be assessed preferably through prospective validation. The final phase is the effect of the pharmacogenomic diagnostic screening test that reduces the burden of disease on the population via therapeutic/diagnostic treatment intervention.

Dr. Richard Simon (NCI, Bethesda, MD) described the elements and value of proper cross-validation in the evaluation of biomarker indices and indicated that cross-validation is valid only if the test set is not used in any way to develop the model. With proper cross-validation, the model is developed from scratch for each “leave-one-out” training set (21-23). He noted that, for smaller studies, cross-validation is preferable to split-sample validation; internal validation is limited by the precision in the estimated error rate and the data used for the developmental study. Whenever working with high-dimensional data, samples should be split between validation and development of markers.

The group noted that each phase of biomarker development requires validation, but true confirmation of patient benefit is established in phase V, a large randomized trial to determine the effect of a test on mortality, as described in the phases of biomarker development by the Early Detection Research Network of NCI (4). Because of the expense, however, phase IV, a prospective study in which a biomarker test that triggers
a workup test should be carried out, and phase V can be completed only for a select few biomarkers. When validating via a simulation, it was noted that a model can never replace the observation. True validation means that results may be reproduced by other laboratories in other settings and in independent populations. Thus, banking specimens for sharing is critical.

Clinical and Biological Challenges: Biological Specimens from Large Institutional Trials

For successful biomarker validations, there is a great need for good specimen collections according to NCI workshop participants. As the technology is fast changing, sample storage and processing play a critical role in determining the suitability of the specimens for various technologies. Whenever possible, pristine samples should be prioritized for discovery efforts. Open interaction among steering committees of large trials, large cohort studies, and cancer biomarker consortia should be encouraged for the free exchange of ideas and specimens for biomarker validation. This will ensure proper use of specimens across research community. By establishing networks of cooperative human tissue banks or resources, existing resources could be made available for discovery rather than through single-tissue banks. Additional resources should be established to catalogue tissue specimen collections acting as “tissue collector” with clearly developed standard operating procedures and common data elements. Systematic monitoring of specimen quality over time should be established with measures to randomly test the quality of samples or other checks for maintaining the integrity of the specimen bank. Academic laboratories should be encouraged to adapt industrial standards for compliance of good laboratory practices and necessary quality control measures. Authorship and ownership issues should be clearly defined by these repositories to avoid future conflicts related to the use of specimens. The NCI is currently developing the National Biospecimen Network to address these issues (24, 25).

Considerations for Biomarker Validation Regulatory Requirements for Commercialization

Dr. Theresa Mullin (FDA) described the Critical Path Initiative and the NCI-FDA interagency task force established in an effort to expedite the delivery of safe products to patients. Dr. Maria Chan (FDA) discussed the Safe Medical Device Amendments Act in 1976 (http://www.fda.gov/cdrh/oivd/index. html) and other regulations associated with in vitro diagnostic devices established to set general controls for product registration, listing, good manufacturing practices, and postmarket surveillance. She discussed three core FDA review issues: analytic performance, clinical performance, and labeling.

To evaluate the performance of a biomarker, the FDA prefers a “yardstick of truth,” such as analysis with receiver operating characteristic curves and the ability to understand biomarker behavior in intended populations. It has been suggested that industry should have a clear idea of the intended use of the biomarker, shown by a good study design, analytic and data collection methods, and the demonstration of good science. The agency in turn should develop clear guidelines to approve cancer diagnostics (e.g., analyte-specific reagent definitions and criteria for classification and showing clinical use of biomarkers that affect therapeutic decision making). It was recommended that industry should engage in early dialogue with the FDA for developing clear roadmaps with necessary regulatory emphasis and priorities on various test developments (e.g., multiplex testing, genomics, quality standards and benchmarks, and combination product guidelines for device/drug or device/biological applications).

Summary

Participants in this NCI-FDA workshop discussed a variety of issues related to the validation of biomarkers for the early detection of cancer. To validate an early detection marker, it is necessary to evaluate the marker as a trigger for early intervention using randomized trials or observational studies for phase V. To evaluate cancer biomarkers for clinical applications, investigators are encouraged to take advantage of resources created for existing randomized trials and large cohort studies. However, biology must drive the decisions about the selection and validation processes among the markers to advance into subsequent study phases. It was also noted that, when investigating the performance of a classification rule based on multiple markers, overfitting may be avoided by selecting a classification rule based on the training sample and then estimating marker performance in a separate test sample. It was also observed that industry must present a clear idea of the intended use of a biomarker to enable a smooth regulatory process and for subsequent commercialization. The development and implementation of standards for specimen handling, annotation, and analysis will enhance the reproducibility and use of biomarker validation studies.

Acknowledgments

We thank Drs. Gregory Campbell, Maria Chan, Steven Hirschfeld, Ralph Kodell, Robert O’Neil, Lakshmi Vishnuvajjala, and Sue Jane Wang (FDA), Gregory Downing (NCI Office of the Director), Ziding Feng, Martin McIntosh, and Ross Prentice (Fred Hutchinson Cancer Research Center), Lance Liotta and Emmanuelle Petricoin (George Mason University, Fairfax, VA), Jose Costa (Yale University, New Haven, CT), Sue Ellenberg (University of Pennsylvania, Philadelphia, PA), Sylvan Green (University of Arizona, Tucson, AZ), William Grizzle (University of Alabama, Birmingham, AL), Richard Schilsky (University of Chicago, Chicago, IL), Yu Shyr (Vanderbilt University, Nashville, TN), Steven Skates (Massachusetts General Hospital, Boston, MA), and Dean Brenner (University of Michigan, Ann Arbor, MI) for their valuable input and helpful discussions in preparation of this article and Charles A. Goldthwaite, Jr. for his assistance in writing this report.

References


15. Freedman LS, Green SB. Statistical designs for investigating several