Biomarkers for Cancer Risk, Early Detection, and Prognosis: The Validation Conundrum

Dean E. Brenner\textsuperscript{1} and Daniel P. Normolle\textsuperscript{2}

\textsuperscript{1}Departments of Internal Medicine and Pharmacology, University of Michigan and VA Medical Center; \textsuperscript{2}Department of Radiology Oncology and Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan

Successful deployment of early detection technologies to population-based screening for colon, cervix, and, arguably, breast cancer has been responsible for recently observed reductions in disease-associated mortalities from these common malignancies, which are uniformly fatal in their advanced stages. With the advent of high-throughput technologies capable of interrogating genes, proteins, and their associated pathways, the promise of new, inexpensive molecular products for use in cancer early detection and cancer risk assessment has generated scientific enthusiasm. This enthusiasm has attracted substantial public and private resources aimed at discovering molecular products that can form the basis of simple, inexpensive, reproducible, highly sensitive, and specific tests capable of detecting malignancies before their metastatic dissemination. Two important reviews in this issue of Cancer Epidemiology, Biomarkers & Prevention (1, 2) address the historical origins, current progress, future promise, and challenges to this rapidly developing field.

Tensions between Mechanism-Based and High-Throughput Technology-Based Biomarker Discovery

Hundt et al. (2) thoroughly yet succinctly review the current state of the art of blood-based biomarkers for the early detection of colorectal adenocarcinoma. Despite a large and burgeoning literature documenting the discovery of such biomarkers derived from diverse discovery technologies that have identified protein, carbohydrate, lipids, mRNA, and DNA molecules, the authors conclude that the critical barrier to future successful deployment of early detection markers is the lack of "[l]arger prospective [validation] studies using study populations representing a screening population." Vineis and Perera (1) approach the issue of cancer risk assessment and early detection through their extensive background and experience in molecular epidemiology of environmental carcinogens. Using selected cases that emphasize the relationship between known carcinogen stress, primarily tobacco smoke and airway carcinogen-
esis, they suggest that "The first generation of biomarkers has ... related largely to genotoxic carcinogens. As a result, interventions and policy changes have been mounted to reduce risk from a number of important environmental carcinogens."

The advent of high-throughput, "omic" discovery technologies raises the challenge of validation of these products in a systematic fashion that permits rapid assessment of their efficacy as predictive instruments for cancer risk and early detection, their use as prognostic tools or therapeutic monitoring products, or a combination of these indications. As noted by Vineis and Perera (1), the high-throughput technologies "...allow massive investigations not based on a priori hypotheses." Such information can be a double-edged sword. On the one hand, the data provide powerful clues that accelerate causal or mechanistic investigations into the disease onset. On the other, the products being developed as biomarkers for disease detection, monitoring, or prognosis have minimal or no biological basis yet could enter clinical practice rapidly.

Definitions of Validation

Currently, there are no agreed definitions of "validation." The concept is so broad, as noted by Feinstein (4, 5) and elaborated by Ransohoff (6), as to sometimes be confusing. Yet within the broad concept are specific questions that can be translated to evaluate current research. The broad concept includes "efforts to confirm the accuracy, precision, and effectiveness of results" (4, 6). Vineis and Perera (1, 7) define validation as "...technical and field validation. Technical validation has to do with intrinsic measurement error and analytical sensitivity. Field (or epidemiological) validation is related to how a certain marker behaves in the population, depending on biological variability within..."
the population.' We use the term ‘analytical validation’ as equivalent to technical validation and ‘clinical validation’ to field validation.

**Fundamental Concerns in Clinical Validation**

We have considered three fundamental concerns related to validation. (a) Overfitting. This refers to the tendency of models trained on large numbers of variables measured on small numbers of samples to produce extraordinarily high sensitivity and specificity and then fail on independent validation sets. This problem has been exacerbated by new technologies, such as DNA microarrays and matrix-assisted laser desorption/ionization time of flight, which produce thousands of variables for each sample or subject. For example, in current high-visibility studies of RNA expression arrays to predict the prognosis of breast cancer, overfitting, without independent assessment of reproducibility, has been proposed as a major explanation for the high degree of discrimination reported (6, 8, 9). (b) Bias. Are results due to differences between the cancer and the control samples that do not exist in the cancer and control populations? This refers to misidentification of the cause of differences between samples. For instance, if a sample of patients is much older than a sample of controls, then differences due to age may be misattributed to disease, or bias is induced through variations in the handling and processing of specimens. Such bias may be an important explanation for current ‘discrimination’ (10, 11). (c) Robustness. Are results generalizable to appropriate clinical populations? This refers to the similarity of the distribution of markers or sets of markers between the samples studied and samples derived from a larger clinical or screening population.

**Addressing These Concerns**

Concerns a and b are so fundamental that they must be clearly and adequately addressed in every research study at every phase of biomarker development. After initial discovery, to discern ‘discriminatory’ characteristics (e.g., what distinguishes between subjects with early cancer versus no cancer) in a rigorous and reproducible fashion, well annotated, prospectively or retrospectively acquired biosamples, collected in compliance with a well-designed protocol that is powered to obtain statistically important outcomes is imperative. A disinterested investigator, preferably a biostatistician, should manage and analyze the data obtained from validation protocols. Many investigators tend to overlook the cost, effort, and skills required to assemble biosample banks (e.g., serum, urine, stool, DNA, and tissue) that can be reliably used to assess the ability of the markers to discriminate among groups of subjects in an unbiased manner. Such work is time consuming, tedious, and unglamorous but is an absolutely critical and rate-limiting step in the overall process of biomarker development and assessment. Biosample banks require standard operating procedures and tracking tools for sample harvest, management, and distribution.

The Early Detection Research Network has proposed to address concern c through the five-phased approach to biomarker validation described by Pepe et al. (12). Phase I is an early ‘exploratory’ study that addresses the discovery and preliminary assessment of the ability of a biomarker to detect a cancer. It can be based on results of an assay done in tumor tissue. Phase II focuses on biomarker detection, in people with clinically evident disease. Phase II also includes age, sex, race, within day and between day, and clinical and analytical quality control characteristics. Phase III assesses the performance of a biomarker in detection of asymptomatic preclinical disease (phase III). Later phases assess performance in prospective screening studies (phase IV) and in definitive large-scale population studies (phase V). Effort and expense substantially increase at each subsequent phase.

**Challenges to Biomarker Research in the Contemporary Research Environment**

**Cost and Infrastructure to Population-Based Longitudinal Validation Trials.** The final phases of validation (stages IV and V) are definitive, longitudinal, population-based studies. The highest-level scientific directorates of major industrial countries have experience in conducting such interventional studies. They cost hundreds of millions of dollars and are not sure successes. The inherent financial risk is magnified by a development paradigm that builds such experiments to test a single proposed treatment. The investment required to develop a vast infrastructure to recruit, monitor, and evaluate thousands of human subjects over years becomes a bet on a single treatment or biomarker. A better strategy might be to establish the testing infrastructure independent of the biomarkers or treatments to be tested, to test multiple markers in parallel, and to make periodic or even continual assessments, redistributing resources as some markers become more promising and others less so.

**Challenges to Industrial Aspects of Biomarker Validation.** There may be no feasible business plan for a given marker. The marker may be so rare that the cost of development may never be recovered. A treatment indicated by a positive marker may not be a product that can be sold, or a marker may have implications to the economics of a different industry that may attract the attention of the lobbyists and their lawyers. Most importantly, intellectual property issues become major barriers to joining biomarkers to panels that might require contributions from different competing industrial groups. Without intellectual property agreements, the validation process becomes fragmented and inefficient due to competing demands for limited clinical resources.

**Challenges to Academic Aspects of Biomarker Validation.** The basic hypothesis-testing paradigm of science may itself present difficulties to the elucidation of biological pathways and subsequent exploitation of those pathways. Science wants to test the effect of an isolated intervention on a strictly defined system, but biomarkers discovered with high-throughput technology are unlikely to have defined biological mechanisms or pathways.

To obtain resources, the proponents of a given biomarker will have to contend with different groups promoting different markers for different diseases. The current design and analysis paradigm focuses on receiver operating characteristic curves, which, although useful...
analytical tools, are primarily descriptive and tend to result in the conclusion that “more research is needed.” The current system has no consistent criteria, either for determining, early in the development of a particular biomarker its potential ultimate application, and hence its development path, or for accomplishing the graceful termination of that development. Resources allocated to a biomarker may continue even after its lack of an ultimate application has become apparent because the science is exciting; it is better than the current standard test (whose performance may be so weak that its primary application is to make other tests look good); the biomarker may be ultimately useful in a panel with other, yet unspecified, marker; or, finally, so much work has already been put into it, it is a shame to abandon it. Furthermore, the tests in phases II and higher are themselves criticized as “not innovative.” Thus, it is difficult to obtain competitive, peer-reviewed funding to conduct the tests necessary to advance biomarkers to useful products. If the tests are conducted, they may be designed in such a way that allow continued development irrespective of the results. The outcome of such a process is a system that rewards the discovery of new biomarkers but discourages their development into tools to enhance the health of the population. When the resources are allocated to development into tools to enhance the health of the population, they are diverted to support of more discovery of more biomarkers but not for their validation.

The “Good Enough” Approach to Phase-Based Biomarker Validation

Because each study of a marker is part of a developmental program, it should result in a decision about that program. We contend that simple estimation of sensitivity and specificity or construction of a receiver operating characteristic curve does not produce an adequate study design criterion. Rather, the design criterion is determined by answering the question “Are the sensitivity and specificity good enough, or potentially good enough, for the proposed application?” Consideration of the proposed application is important because it recognizes that the standards of efficacy will vary based on the mortality of the disease process; the alternative diagnostic or therapeutic tools available; and, in the case of early detection, the prevalence of the disease and the cost and morbidity of a subsequent diagnostic procedure. These considerations determine minimum acceptable sensitivity and specificity, which can be assessed in case-control studies designed to test parallel, one-sided null hypotheses that the sensitivity (measured on cases) and specificity (measured on controls) are less than some minimal criteria. Rejection of the null hypothesis leads to the next step in the developmental program. The sizes of the test sets may be determined using standard frequentist arguments or in a Bayesian context. The good enough criterion is then developed. Although receiver operating characteristic curves may still be estimated as a secondary end point, their use as the primary study objective does not motivates difficult decisions about application context (e.g., clinical tool versus population screen) or continued development.

Practical Considerations of Biomarker Validation

As called for by both reviews in this issue (1, 2), the resources required to complete the validation of biomarkers are beyond the capabilities of individual laboratories, investigators, or institutions. Collaborative research groups that are vertically integrated with basic, population, and clinical scientists, such as the National Cancer Institute–sponsored Early Detection Research Network, can offer the necessary resources to successfully validate promising biomarkers that may reduce cancer mortality.

It is too soon to determine if this particular initiative will be successful, but the Early Detection Research Network’s vertically integrated, team research model is attempting to rebalance the previously described incentive structure favoring repetitive discovery research over validation research. The dilemmas of intellectual property combined with the increased regulatory burden associated with the protection of privacy, and assurance of consent shouldered by investigators, make the kinds of studies required to validate potential biomarkers even more difficult and serve to hinder progress toward the goals of validation. The continued support and refinement of vertically integrated academic and industrial models, such as the Early Detection Research Network, will require patience, focus, careful management, and continued leadership support to sustain the promise of a balanced, appropriately calibrated system to develop and validate biomarkers for early cancer detection and risk assessment.

References

Biomarkers for Cancer Risk, Early Detection, and Prognosis: The Validation Conundrum

Dean E. Brenner and Daniel P. Normolle


Updated version
Access the most recent version of this article at:
http://cebp.aacrjournals.org/content/16/10/1918

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.