Use of Archived Specimens in Evaluation of Prognostic and Predictive Biomarkers

Richard M. Simon, Soonmyung Paik, Daniel F. Hayes

The development of tumor biomarkers ready for clinical use is complex. We propose a refined system for biomarker study design, conduct, analysis, and evaluation that incorporates a hierarchal level of evidence scale for tumor marker studies, including those using archived specimens. Although fully prospective randomized clinical trials to evaluate the medical utility of a prognostic or predictive biomarker are the gold standard, such trials are costly, so we discuss more efficient indirect "prospective–retrospective" designs using archived specimens. In particular, we propose new guidelines that stipulate that 1) adequate amounts of archived tissue must be available from enough patients from a prospective trial (which for predictive factors should generally be a randomized design) for analyses to have adequate statistical power and for the patients included in the evaluation to be clearly representative of the patients in the trial; 2) the test should be analytically and preanalytically validated for use with archived tissue; 3) the plan for biomarker evaluation should be completely specified in writing before the performance of biomarker assays on archived tissue and should be focused on evaluation of a single completely defined classifier; and 4) the results from archived specimens should be validated using specimens from one or more similar, but separate, studies.

J Natl Cancer Inst 2009;101:1446–1452

Many cancer patients do not benefit from the systemic treatments they receive. For example, adjuvant chemotherapy that is considered highly effective may often improve the disease-free or overall survival rate by only 5–10 percentage points. Also, chemotherapy for metastatic disease often provides sustained benefit for a small portion of the patients treated. Therefore, the practice of oncology has been very inefficient, with exposure of far more patients than will benefit to the cost and toxicity of these agents . Although this overtreatment is understandable in dealing with life-threatening diseases, the ability to better "personalize" treatment decisions could have important benefits for patients as well as medical costs. In spite of developments in biotechnology and genomics, the pace of acceptance of new markers to inform treatment decisions for patients with cancer has been slow. The limited introduction of effective biomarkers is partly because of the substantially lower reimbursement for tumor marker tests, as compared with therapeutics by health insurers, but is also because of a shortage of prospective studies of marker utility and the lack of reproducibility and reliability among the many published retrospective studies of prognostic and predictive markers (1,2).

Several committees and authors have proposed specific guidelines that might be used to evaluate and report a given marker. For example, in 1996, the members of the American Society of Clinical Oncology Tumor Markers Guidelines Committee recommended five Levels of Evidence (LOEs) that might be used to determine the clinical utility of a tumor marker (3). This LOE scale has been widely cited and used as a template for deciding whether to recommend the use of a tumor marker in clinical practice and for design and conduct of tumor marker studies (4,5). The criteria for reporting the results of marker studies (designated the REMARK criteria) have been published in several journals, and at least a few journals have incorporated REMARK into the required submission format (6,7).

In this article, we will address the nature of the methodological difficulties involved in studying tumor markers, both prognostic (ie, predictive of prognosis, independent of treatment) and predictive (ie, in terms of best choice of therapy). We will also propose that there are conditions in which archived specimens can be used to provide reliable evaluations of the clinical validity or medical utility of prognostic and predictive biomarkers.

Prospective Randomized Trials to Address Tumor Marker Utility

The gold standard for establishing clinical utility of a new medical intervention is the prospective randomized clinical trial. Several authors have proposed prospective randomized clinical trial designs for evaluation of prospective or predictive diagnostic markers (8–13). In the latter circumstance, the medical utility of the candidate predictive biomarker can be established by evaluating the benefit of the new drug according to marker status (positive or negative) in adequately sized patient subgroups using a prospectively specified analysis plan within a randomized clinical trial that compares a regimen containing the new drug to a control.

One might consider a prospective clinical trial in which the test itself is the investigational intervention to be the ultimate validation

Affiliations of authors: Biometric Research Branch, National Cancer Institute, Bethesda, MD (RMS); Division of Pathology, National Surgical Adjuvant Breast and Bowel Project, University of Pittsburgh, Pittsburgh, PA (SP); Breast Oncology Program, University of Michigan Comprehensive Cancer Center, Ann Arbor, MI (DFH).

Correspondence to: Richard M. Simon, DSc, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434 (e-mail: rsimon@nih.gov).

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

DOI: 10.1093/jnci/djp335

Published by Oxford University Press 2009.

Advance Access publication on October 8, 2009.

of a prognostic or predictive tumor marker. That is, a trial may be designed so that a patient's care would be determined based on random assignment to use the test or not, as referred to as the marker strategy design by Simon and Wang (14). In such a trial, treatment decisions are made for patients who are randomly assigned to the control group using standard prognostic factors and practice guidelines. For patients who are randomly assigned to the investigational group, the test, or marker, is used in treatment determination, perhaps in conjunction with standard prognostic factors. The test would be performed only for patients who are randomly assigned to the test group, and the trial would be evaluated by comparing outcomes overall for the two randomization groups. The outcomes must be compared overall because the new test is not used for the "control" group. In many cases, this restriction seriously limits the information that can be gleaned from the design. Results can be particularly confounded and diluted in cases where the standard of care is variable among physicians.

The marker strategy design is also generally very inefficient in terms of the number of patients required for randomization. Sample size requirements for randomized clinical trials are often proportional to the reciprocal of the square of the size of the treatment effect to be detected with a specified statistical power. For the marker strategy design, only the overall treatment effect between the two randomized groups can be evaluated, and the size of that effect is generally quite small because many patients will receive the same treatment regardless of the group to which they are randomized. If the analysis is to demonstrate that withholding a standard therapy for test-negative patients is not inferior, then sample size problems are compounded, and even with a huge sample size, the results are unlikely to be convincing.

An alternative approach requires that all patients be tested for marker status "upfront." In this case, the evaluation can be focused on subsets of patients for whom the treatment assignment that is based on the test differs from treatment assignment that is based on standard of care. For example, suppose the standard of care is to use chemotherapy for stage II patients but not for stage I patients and the test purports to identify patients who are likely to benefit from chemotherapy regardless of stage; test-positive patients will receive chemotherapy and test-negative patients will not. In this case, the only patients randomly assigned are stage I patients with a positive test and stage II patients with a negative test. The design enables the effectiveness of chemotherapy to be evaluated separately for these subsets of patients. This design presumes, however, that the standard of care, as a function of standard prognostic variables, is determined.

This strategy of testing all patients up-front is used by two current clinical trials, the Microarray in Node-Negative Disease may Avoid Chemotherapy (MINDACT) study in Europe (15) and the Trial Assigning Individualized Options for Treatment (Rx) (TAILORx) study in North America (16). Although the designs of both trials are complex and somewhat different, they both address the medical utility of withholding standard of care chemotherapy from women with node-negative estrogen receptor–positive breast cancer who have a predicted low risk of recurrence, based on a predefined gene expression–based risk score. The MINDACT study evaluates a 70-gene classifier, and the TAILORx study evaluates a 21-gene classifier. Even though these designs are more efficient than the randomized marker strategy trial design, both of these studies will require many thousands of patients, and nearly a decade each from the time, accrual was begun until the first results are anticipated. The TAILORx and MINDACT studies will cost millions of dollars or Euros to conduct, and with the current speed of the evolution of technology, the test being evaluated may have become obsolete by the time such studies are completed.

It is common for a new marker to be identified after the definitive trials have demonstrated benefit for a specific agent or class of agents or even type of modality (such as chemotherapy in general). We maintain that, in many cases, it may be possible to use archived specimens collected in the past from appropriate previously conducted therapeutic trials and to preserve the focus, control of type I error, and statistical power of properly designed fully prospective studies. Indeed, when there is substantial preliminary evidence that a new marker predicts benefit from a specific drug, it may sometimes be possible to assay the marker in archived specimens from randomized clinical trials that were conducted to evaluate the drug, as was done for *KRAS* in colorectal cancer (17,18).

When suitable archived tissue is available and can be used reliably, it can facilitate and expedite delivery of valuable cancer diagnostics that may be of considerable benefit to patients. Nonetheless, there are certainly also risks to patients from the unreliable use of archived tissues. We have tried here to clarify the key features involved in using these resources in a reliable manner, and we propose a refinement to the previously published LOE scale that permits a more critical analysis of the quality of tumor marker studies using archived specimens.

Prospective vs Retrospective Studies: A Matter of Semantics

Although biomedical scientists and biostatisticians are taught that "prospective" studies are preferable to "retrospective" studies, the distinction between prospective and retrospective is often confused with the distinction between "experimental" and "observational." We propose that for studies of prognostic and predictive biomarkers in oncology, the term retrospective is in some cases misleading.

In cancer epidemiology, both retrospective case–control studies and prospective cohort studies are observational, rather than experimental, studies. Neither type of study involves random assignment of exposure, and hence, observed associations between exposures and disease do not provide as strong a basis for claims of causality as in experimental studies. The most serious limitation of epidemiological studies is their nonexperimental nature, not whether they are retrospective or prospective.

In therapeutics, many retrospective analyses are also nonexperimental, with treatment selection based on patient factors and referral pattern rather than on randomization. Such studies are also often conducted without a written protocol and are unfocused, with numerous patient subsets and endpoints compared without control for the overall chance of a false-positive conclusion. In contrast, prospective randomized clinical trials contain internal control of treatment assignment, careful and proscribed data collection (including outcomes and endpoints), and a focused analysis plan that is developed before the data are examined.

Many biomarker studies are conducted with convenience samples of specimens, which just happen to be available and are assayed for the marker, with no prospectively determined subject eligibility, power calculations, marker cut-point specification, or analytical plans. Such studies are very likely to result in highly biased conclusions and truly deserve to be pejoratively labeled as "retrospective." However, if a "retrospective" study is designed to use archived specimens from a previously conducted prospective trial, and if certain conditions are prospectively delineated in a written protocol before the marker study is performed, we argue that it might be considered a "prospective–retrospective" study. Such a study should carry considerably more weight toward determination of clinical utility of the marker than a simple study of convenience, in which specimens and an assay happen to be available. Having multiple studies of different candidate biomarkers based on archived tissues from the same prospective trial would, however, present a greater opportunity for false-positive conclusions than a single fully prospective trial focused on a specific biomarker. Consequently, independent confirmation of findings for specific biomarkers in multiple prospective–retrospective studies is important (see below).

Using Archived Tissue to Establish the Medical Utility of a Marker

In assessing the use of archived specimens in the evaluation of prognostic and predictive biomarkers, it is useful to consider the three requirements for clinical acceptance of a tumor marker that were first proposed by Henry and Hayes (2): 1) the specific setting and utility of the marker must be clear, 2) the magnitude in either outcomes or treatment effects between those patients who are "positive" for a marker must be sufficiently different from those who are "negative" for that marker that the clinician and/or patient would accept different treatment strategies for the two patients, and 3) the estimates of that magnitude must be reliable.

These criteria relate to establishing the clinical utility of the marker. It is useful to clarify the use of the term "validation" as applied to diagnostic tests. Hunter et al. (19) distinguished three types of validity in terms of genetic tests: "First, there is the question of a test's analytic validity, its ability to accurately and reliability measure the genotype of interest . . .. Second, one must consider clinical validity, or the ability of the test to detect or predict the associated disorder . . .. Finally, there is the issue of the test's clinical utility, or the balance of its associated risks and benefits if it were to be introduced into clinical practice." Clinical utility requires that the test is "actionable," that the clinical context and medical indication for use of the test is clear, and that the magnitude of outcomes or treatment effects associated with different results of the test are sufficiently great as to influence treatment decisions. A serious defect of most retrospective studies of prognostic markers is that the patients are not selected for addressing a defined medical indication for use of the marker. Such studies may establish a correlation with clinical outcome but not the medical utility of the marker.

The consideration of reliably establishing the magnitude of marker effect may be further divided into the following three conditions: 1) the technical and analytical properties of the marker assay must be accurate and/or robust and reproducible; 2) the clinical study design and analysis must be appropriate and adequate to address the utility of a precise intended clinical use; and 3) the results should be verified, or validated, in more than one study set, with similar estimates of the magnitude in separate populations of patients that resemble each other. Each of these conditions is potentially subject to considerable bias in most retrospective studies using archived specimens, especially those of convenience. Even if the investigation is a prospective–retrospective study, careful attention to each of these concerns will reduce the bias and inconsistent results obtained with studies of convenience, and we believe that it will further hasten the introduction of useful tumor markers into clinical practice.

Analytical Concerns

"Analytical validation" generally refers to reproducibility and robustness of the test or assay value. This generally includes minimizing variation with regard to both preanalytical factors, such as tissue collection, processing, storage, and preparation, as well as analytical factors, such as reagent choice, incubation time and conditions, and method of readout (including cut-point determination) (20,21).

For a clinical biomarker evaluation using archived tissues to be interpretable, it is necessary that the assay results from the archived sample reflect what would happen in a true clinical setting. The following are examples of how archived tissue might differ from true clinical specimens.

1) Preanalytical issues. It is possible that samples collected in the past, and specifically for the bank in hand, might be handled differently than they are in current practice. Examples of differences might include whether a precollection diagnostic biopsy was performed (which might affect various gene expression and tissue processes), the time after the sample was removed from the patient and processed (fixed, frozen, etc), procedures for fixation or freezing, how the sample was stored (temperature, exposed to room air, as a tissue block or a section on a slide, etc), and how many cycles it was frozen and thawed.

2) Analytical issues. For a tumor marker study to be sufficient to change clinical practice, the test itself should be ready for clinical practice. For studies to change clinical practice, the investigator should carefully and prospectively plan to use reagents, conditions, and cut points that have been previously determined to be accurate and reproducible. These considerations include fixed reagent supply sources, concentrations, and incubation times among many other possible variables. In addition, the investigator should have demonstrated with statistical confidence the analytical concordance of results between archived specimens and clinical samples for that specific assay. Examples of these concerns include whether the sample was prepared for analysis in a tissue microarray or as a whole section, and whether and how it was subjected to antigen retrieval.

As a precaution against bias that may result from incomplete analytical and preanalytical validation, marker studies using archived specimens should have the assays performed blinded to all clinical data, including treatment and patient outcome.

Clinical Study Design

As noted in the first required condition, the investigator should have a clear idea of the specific intended use for the assay. In general, this will be as a prognostic factor to decide if any further treatment is necessary or as a predictive factor to determine whether a particular type of therapy is likely to be effective. To establish medical utility of a prognostic marker, a randomized trial is sometimes not necessary. For example, a prospective single-arm trial in which chemotherapy is withheld from patients at a low risk of recurrence is used in the portion of the TAILORx clinical trial designed to validate the very favorable prognostic outcomes in the low recurrence score population. Assuming that preanalytical factors are well controlled and match current practice activities and that the clinical data are collected in a fashion typical of a clinical trial, archived tissue from a sufficiently large population of untreated patients may be adequate to permit accurate estimates of recurrence based on tumor marker subgroups for determination of clinical utility of the marker.

Tumor response data from a single-arm phase II clinical trial of a specified treatment can be used to establish the clinical validity of a biomarker for predicting response to that treatment, but a larger randomized trial with a survival or progression-free survival endpoint is generally required to establish the medical utility of the predictive marker.

Suggested Revision of LOEs

In the original American Society of Clinical Oncology LOE scale, “retrospective studies” were determined to be LOE II or worse (3). We now propose an updated revision of the LOE scale, in which more precise definitions are provided for the types of studies that might be used to analyze the clinical utility of a biomarker and in which retrospective studies using archived specimens might reach level I evidence. The LOE for the medical utility of a biomarker relates to key factors involving patients, specimens, assays, and statistical analysis plans (Tables 1 and 2).

Scientifically, the clinical utility of a biomarker in a particular situation is best addressed by a prospective randomized clinical trial (Table 1, category A). Patients are entered, treated, and followed

Table 1. Elements of tumor marker studies that constitute Levels of Evidence determination*

Category Element	A Prospective	B Prospective using archived samples	C Prospective/ observational	D Retrospective/ observational
Clinical trial	PCT designed to address tumor marker	Prospective trial not designed to address tumor marker, but design accommodates tumor marker utility Accommodation of predictive marker requires PRCT	Prospective observational registry, treatment and follow-up not dictated	No prospective aspect to study
Patients and patient data	Prospectively enrolled, treated, and followed in PCT	Prospectively enrolled, treated, and followed in clinical trial and, especially if a predictive utility is considered, a PRCT addressing the treatment of interest	Prospectively enrolled in registry, but treatment and follow-up standard of care	No prospective stipulation of treatment or follow-up; patient data collected by retrospective chart review
Specimen collection, processing, and archival	Specimens collected, processed, and assayed for specific marker in real time	Specimens collected, processed, and archived prospectively using generic SOPs. Assayed after trial completion	Specimens collected, processed, and archived prospectively using generic SOPs. Assayed after trial completion	Specimens collected, processed and archived with no prospective SOPs
Statistical design and analysis	Study powered to address tumor marker question	Study powered to address therapeutic question and underpowered to address tumor marker question	Study not prospectively powered at all. Retrospective study design confounded by selection of specimens for study	Study not prospectively powered at all. Retrospective study design confounded by selection of specimens for study
		Focused analysis plan for marker question developed before doing assays	Focused analysis plan for marker question developed before doing assays	No focused analysis plan for marker question developed before doing assays
Validation	Result unlikely to be play of chance	Result more likely to be play of chance that A but less likely than C	Result very likely to be play of chance	Result very likely to be play of chance
	Although preferred, validation not required	Requires one or more validation studies	Requires subsequent validation studies	Requires subsequent validation

* PCT = prospective controlled trial; PRCT = prospective randomized controlled trial; SOPs = standard operating practices.

Table 2. Revised determination of Levels of Evidence using elements of tumor marker studies*

Level of evidence	Category from Table 1	Validation studies available
I	A	None required
I	B	One or more with consistent results
II	B	None or inconsistent results
II	C	2 or more with consistent results
III	C	None or 1 with consistent results or inconsistent results
IV–V	D	NA†

* Levels of Evidence (LOEs) revised from those originally proposed by Hayes et al. (3).

† NA = not applicable because LOE IV and V studies will never be satisfactory for determination of medical utility.

prospectively according to a prewritten protocol; the study is prospectively powered specifically to address the tumor marker question; and specimens are collected, processed, and assayed for the marker in real time. The randomized trial will generally not use a "marker strategy design" as described above, however, because of the serious limitations of that design. Although further confirmation in a separate trial of the results gained from a category A prospective trial is always welcome, compelling results from such a trial would be considered definitive and no other validating trial would be required. This strategy was included in the original LOE scale proposed by American Society of Clinical Oncology as LOE I and continues to be the "gold standard."

In the revised LOE scale, a second strategy to obtain level I data would be to perform a tumor marker study using archived specimens from a prospective trial that addresses a therapeutic question (or another marker question) and accommodates the current marker question (Table 1, category B). To evaluate prognostic markers that are intended to identify patients for whom prognosis is so good that further therapy would be withheld, the clinical trial in some cases may not need to be randomized. For example, in the TAILORx study, the low recurrence score group receives only endocrine therapy and is followed to determine if risk of recurrence is as low as predicted by the 21-gene recurrence score. To evaluate a predictive marker, the prospective trial would generally need to be a randomized trial that compares the treatment with an appropriate control treatment. As in study design A, patients are prospectively enrolled, treated, and followed, and specimens are prospectively collected, processed, and archived using generic standard operating procedures. The tumor marker question might be identified during the conduct of the trial or after its completion, but the specification of the tumor marker hypothesis should be based on results completely external to the trial. In fact, tissues archived from the trial should not be assayed until a new protocol has been written that focuses on the evaluation of the specified new marker with a completely specified statistical analysis plan. Before undertaking the study, the assay should be analytically and preanalytically validated for use with archived tissue, and the assay should be performed blinded to the clinical data. Because the trial was designed to address the therapeutic question, it will often be underpowered to establish the statistical significance of treatment by marker interaction (22). It may, however, be adequately sized to reliably identify a large treatment effect in "test-positive" patients, as might be expected for a predictive biomarker. Nevertheless, even with these caveats, results from such a study will be more likely to arise from chance than those from a fully prospective approach.

It is clearly desirable that the available specimens from the archived bank should be representative of the patients who were accrued to the study as a whole, although there is no guarantee that the study patients are themselves representative of the general population of patients. Although there are no minimal requirements that can be universally applicable, we suggest that the correlative study should include at least two-thirds of the total accrued patients or that the patients be selected in a way that strives to avoid selection bias. For example, if the investigator wishes to minimize resource utilization, or wishes to use intrastudy specimen sets for test and validation, one might use a mathematical randomization scheme to select a sample of specimens for study that mirror the known important prognostic and predictive factors of the population as a whole (5).

For a category B study to be sufficient to change practice, we maintain that the results must be confirmed using specimens from a second category B study based on archived tissue from a different trial that has been designed, conducted, and analyzed in a similar, if not identical, manner. The results of these two studies must be equally compelling to change clinical practice. Furthermore, these validation studies need to be performed using the same assay or similar assays that clearly identify the same marker. For example, different investigators have used several different assays for p53 status, including direct sequencing for genetic abnormalities, immunohistochemistry to determine protein expression, or even functional assays. These assays provide very different indications of p53, and therefore, the available data are very difficult to interpret (5). Validation studies must also address the same endpoint and that endpoint should reflect medical utility.

Using nearly 1500 archived specimens collected within a prospective randomized clinical trial, Hayes et al. (23) reported that node-positive, estrogen receptor–positive, and human epidermal growth factor receptor 2–negative patients did not appear to benefit from addition of adjuvant paclitaxel chemotherapy after four cycles of doxorubicin and cyclophosphamide. Although these observations were provocative, results from a completely separate, but similarly designed, prospective randomized clinical trial did not confirm these findings (24), and the question regarding selection of patients for adjuvant paclitaxel remains open (25). Thus, this issue is still considered to be LOE II in Table 2. By contrast, the recently observed association of presence of *KRAS* mutations with lack of benefit from monoclonal antibodies directed against the epidermal growth factor receptor , such as cetuximab and panitumumab (17,18), provides an example of successful use of category B archived samples to establish medical utility. Several prospective randomized trials have demonstrated a small but statistically significant benefit from these antibodies, either alone or in combination with chemotherapy, for treatment of patients with advanced colorectal cancer (26). Preliminary, LOE II or III studies suggested that cetuximab and panitumumab are only active in patients whose cancers carry a wild-type *KRAS* (27). These data have now been validated in a retrospectively performed study using archived samples from large prospectively randomized clinical trials and therefore would achieve LOE I in our modified scale (Tables 1 and 2) (28).

Category C (Table 1) biomarker studies use prospective patient registries in which subjects are treated and followed according to standards of care. Specimens are collected, processed, and archived prospectively, using generic standard operating procedures, but are assayed after the study has completed patient accrual. Tumor marker studies conducted using these specimens are often not prospectively powered at all. Because of the lack of control of treatment assignment, specimen collection, and data collection, such settings are generally more susceptible to selection biases for patients, specimens, and clinical data that include outcomes. This concern may not be the case in some tightly controlled population-based registries. Category C studies are more likely confounded by unrecognized biases, and their results are more likely to result from chance than those of categories A and B. Category C studies may be validated to LOE II if two or more subsequent studies provide similar results (Table 2). However, it is unlikely that category C studies would ever be sufficient to change practice, except under particularly compelling circumstances.

Category D studies (Table 1) are the most common type of reported tumor marker analyses: studies of convenience in which specimens were collected for unknown reasons, processed and stored in a variety of ways, and happen to be available for assay. The results from these types of studies are highly unstable and likely to be because of chance alone.

Summary

Ideally, any new medical intervention will be adopted into clinical practice only in the setting of level I evidence, and ideally, such evidence is generated in a prospective randomized clinical trial. However, such trials are not always practical. In the case of tumor markers, practice guidelines and the availability of other diagnostic procedures can sometimes make it very difficult to perform new clinical trials because such trials may involve withholding of therapy that is considered standard of care. Even when they are considered ethical, such trials usually require many years to conduct and are quite expensive. For new drug development, in many cases, an analytically validated companion diagnostic test will not be available or the appropriate biological measurement may not be clear at the time that the pivotal trials of the drug are initiated, as for the use of *KRAS* mutation as a predictive biomarker for EGFR inhibitors in colorectal cancer (17,18,28).

Archived tissue specimens from high-quality datasets can therefore be of great importance for establishing the medical utility of a prognostic or predictive biomarker. We argue that it is appropriate to use archived tissue specimens from large prospective clinical trials to do so. For such an evaluation to be more useful than just for generating hypotheses, however, several conditions must be satisfied:

1) Archived tissue, adequate for a successful assay, must be available on a sufficiently large number of patients from the pivotal trials to permit appropriately powered analyses and to ensure that the patients included in the biomarker evaluation are clearly representative of the patients in the pivotal trials. Although no minimal requirement can be stated as universally applicable, we would suggest that samples from at least two-thirds of the patients be available for analysis.

2) Substantial data on analytical validity of the test must exist that ensure that results obtained from the archived specimens will closely resemble those that would have been obtained from analysis of specimens collected in real time. Assays should be conducted blinded to the clinical data.

3) The analysis plan for the biomarker evaluation must be completely developed before the performance of the biomarker assays. Both the analysis plan for the biomarker study and the design of the trial(s) whose samples were selected for analysis should be appropriate for the evaluation of a companion diagnostic had it been undertaken at the outset. The analysis should be focused on a single, completely defined, diagnostic classifier. For multigene classifiers, the mathematical form of combining the individual components, weights, and cut points should be specified beforehand. In general, the analysis should not be exploratory, and practices that might lead to a false-positive conclusion should be avoided.

4) The results must be validated in at least one or more similarly designed studies using the same assay techniques.

Physicians need improved tools for selecting treatments for individual patients. Cancers of the same primary site are in many cases heterogeneous in molecular pathogenesis, clinical course, and treatment responsiveness. Current approaches for treatment development, evaluation, and use result in treatment of many patients with ineffective drugs. Advances in cancer genomics and biotechnology are providing increased opportunities for development of more effective therapeutics and prognostic and predictive biomarkers to inform their use. These opportunities have enormous potential benefits for patients and for containing health-care costs. However, the complexity of cancer biology and the increased complexity of development of biomarkers with drugs offer formidable challenges to the transition to a more predictive oncology. In some cases, it is either ethically or practically impossible to evaluate the medical utility of prognostic and predictive biomarkers in a fully prospective manner.

It is essential to ensure that cancer patients are offered the benefits of valuable prognostic and predictive tests as soon as they are rigorously and reliably evaluated. In this article, we have tried to clarify some of the uncertainty in the field about the validation of prognostic and predictive biomarkers and to propose an update of a LOE schema that has been widely used for evaluating the medical utility of biomarkers in oncology. We believe that this update is important for improving the conduct of validation studies and, in some cases, for expediting the adoption of important diagnostic tools.

References

1. Ransohoff DE. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer.* 2004;4(4):309–314.
2. Henry NL, Hayes DF. Uses and abuses of tumor markers in the diagnosis, monitoring and treatment of primary and metastatic breast cancer. *Oncologist.* 2006;11(6):541–552.

3. Hayes DF, Bast RC, Desch CE, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst.* 1996;88(20):1456–1466.
4. Locker GY, Hamilton S, Harris J, et al. ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *J Clin Oncol.* 2006;24(33):5313–5327.
5. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol.* 2007;25(33):5287–5310.
6. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst.* 2005;97(16):1180–1184.
7. Hayes D, Ethier S, Lippman M. New guidelines for reporting tumor marker studies in breast cancer research and treatment: REMARK. *Breast Cancer Research and Treatment.* 2006;100(1):237–238.
8. Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol.* 2005;23(29):7332–7341.
9. 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.
10. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res.* 2005;11(21):7872–7878.
11. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst.* 2007;99(13):1036–1043.
12. Simon R. Using genomics in clinical trial design. *Clin Cancer Res.* 2008;14(19):5984–5993.
13. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. *Expert Rev Mol Diagn.* 2008; 2(6):721–729.
14. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. *Pharmacogenomics J.* 2006;6(3):166–173.
15. Bogaerts J, Cardoso F, Buyse M, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat Clin Pract Oncol.* 2006;3(10):540–551.
16. Sparano JA. TAILORx: Trial Assigning Individualized Options for Treatment(Rx). *Clin Breast Cancer.* 2008;7(4):347–350.
17. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol.* 2008;26(10):1626–1634.
18. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med.* 2008;359(17):1757–1765.
19. Hunter DJ, Khoury MJ, Drazen JM. Letting the genome out of the bottle—will we get our wish. *N Engl J Med.* 2008;358(2):105–107.
20. Goldstein NS, Hewitt SM, Taylor CR, Yaziji H, Hicks DG. Recommendations for improved standardization of immunohistochemistry. *Appl Immunohistochem Mol Morphol.* 2007;15(2):124–133.
21. Goldstein NS, Hunter S, Forbes S, Odish E, Tehrani M. Estrogen receptor antibody incubation time and extent of immunoreactivity in invasive carcinoma of the breast: the importance of optimizing antibody avidity. *Appl Immunohistochem Mol Morphol.* 2007;15(2):203–207.
22. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 1 x k factorial design when time-to-failure is the outcome. *Control Clin Trials.* 1993;14(6):511–522.
23. Hayes DF, Thor AD, Dressler LG, et al. HER2 and response to paclitaxel in node-positive breast cancer. *N Engl J Med.* 2007;357(14):1496–1506.
24. Martin M, Rodriguez-Lescure A, Ruiz A, et al. Randomized phase 3 trial of fluorouracil, epirubicin, and cyclophosphamide alone or followed by paclitaxel for early breast cancer. *J Natl Cancer Inst.* 2008;100(11): 805–814.
25. Hudis C, Dang C. The taxane limbo: how low can we go? *J Natl Cancer Inst.* 2008;100(11):761–763.
26. Labianca R, La Verde N, Garassino MC. Development and clinical indications of cetuximab. *Int J Biol Markers.* 2007;22(1 suppl 4):S40–S46.
27. Messersmith WA, Ahnen DJ. Targeting EGFR in colorectal cancer. *N Engl J Med.* 2008;359(17):1834–1836.
28. Allegra C, Jessup JM, Somerfield MR, et al. American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy [published online ahead of print]. *J Clin Oncol.* 2009;27(12):2091–2096.

Funding

Supported in part by the Fashion Footwear Charitable Foundation of New York/QVC Presents Shoes on Sale (D.F.H.).

Notes

The authors take full responsibility for the analysis and interpretation of the data, the writing of the manuscript, and the decision to submit the manuscript for publication.

The authors would like to acknowledge very helpful discussions with their colleagues and members of the American Society of Clinical Oncology Tumor Markers Guidelines Committee that contributed to much of this work. D. F. Hayes reports research and other support from Astra Zeneca, GlaxoSmithKline, Pfizer, Novartis, Ayerst-Wyeth, Veridex (Johnson & Johnson), and Sanofi-Aventis, Predictive Biosciences, Curalogie SAB, Incyte Corporation, DNA Repair, and Compendia Bioscience.

Manuscript received December 5, 2008; revised August 4, 2009; accepted August 27, 2009.