Perspective

Cancer Research

Why Your New Cancer Biomarker May Never Work: Recurrent Patterns and Remarkable Diversity in Biomarker Failures

Scott E. Kern

Abstract

Less than 1% of published cancer biomarkers actually enter clinical practice. Although best practices for biomarker development are published, optimistic investigators may not appreciate the statistical near-certainty and diverse modes by which the other 99% (likely including your favorite new marker) do indeed fail. Here, patterns of failure were abstracted for classification from publications and an online database detailing marker failures. Failure patterns formed a hierarchical logical structure, or outline, of an emerging, deeply complex, and arguably fascinating science of biomarker failure. A new cancer biomarker under development is likely to have already encountered one or more of the following fatal features encountered by prior markers: lack of clinical significance, hidden structure in the source data, a technically inadequate assay, inappropriate statistical methods, unmanageable domination of the data by normal variation, implausibility, deficiencies in the studied population or in the investigator system, and its disproof or abandonment for cause by others. A greater recognition of the science of biomarker failure and its near-complete ubiquity is constructive and celebrates a seemingly perpetual richness of biologic, technical, and philosophical complexity, the full appreciation of which could improve the management of scarce research resources. *Cancer Res; 72(23); 6097–101. ©2012 AACR.*

Introduction: What Is Biomarker Development?

Fundamentally, biomarker development is observational, or empiric, research. As such, it shares the shortcomings characteristic of observational methods. It lacks the strengths of a designed experiment. For example, interventional clinical trials permit the use of randomization and other powerful tools and controls not available to observational clinical studies. The latter may disprove, but cannot alone prove, causal relationships.

At the inception of most "new" biomarkers, the research is generally ambiguous as to the key decisions needed to conduct the study (1). Investigators may not yet have fixed the scope or the variety of data types to be collected, the quality control rules to use for data inclusion/exclusion and data processing, when to interrupt data collection to interrogate the data, and whether to resume data collection after some data have been analyzed. Because this interrogation can be part of the quality-control mechanisms, it is not feasible to prohibit all forms of premature data analysis. Initially, also unsettled will be the precision and the number of questions to be directed toward the data. The lockdown rules under which a final data set is defined may be delayed; such decisions may not be possible before starting data analysis. These lockdown rules should be published in all emerging biomarker reports, but generally are not. Experimental research, in contrast, need not violate these ideals as often.

"Outcomes data" are an essential component of both experimental and observational research. In biomarker development, the markers are intended to find diseases yet unknown or predict events yet to unfold. These outcomes data, however, are often unreliable. Underreporting and erroneous reporting are common. Sources of outcomes data, such as chart review and adverse-event reporting, have frustrating limitations. For example, it is possible that the outcome of interest to the biomarker study was never systematically gathered. A simple outcome inquiry, such as "Were you ever diagnosed to have a cancer?" may not have been posed as a question to the patient. A measure of significance, such as "Was the progression of the cancer the cause of death?" may not have been recorded in the medical record examined. Different data types may have differing levels of flaws, with serious biases and confounded variables selectively affecting the more unreliable data types.

Novel Markers Seldom Stand the Test of Time

Pervasive problems are not limited to biomarker development. Referring to biomedical research as a whole, Ioannidis offered a strong argument that most published scientific results are wrong (2). Referring to research published in psychology, Simmons explained that "It is unacceptably easy

Author's Affiliation: The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: Scott E. Kern, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Department of Oncology, 1650 Orleans Avenue, Baltimore, MD 21287. Phone: 410- 614-3314; Fax: 443-287-4653; E-mail: sk@jhmi.edu

doi: 10.1158/0008-5472.CAN-12-3232

to publish 'statistically significant' evidence consistent with any hypothesis" (1).

The considerable resources devoted to biomarker research belie a similar, sobering truth, that few biomarkers enter clinical practice. If one were to define as "success" the creation of an insurance-reimbursable biomarker test used to alter clinical decision making, less than 1% of new cancer biomarkers have been successful. A sizable group of highly successful markers is identified by the reimbursements paid to pathologists to interpret new immunohistochemical markers for tumor classification. Another group, comprising unsuccessful markers, is defined by the paucity of new blood-based screening or prognostic markers among the clinical tests offered by medical clinics. Diamandis observed that "very few, if any, new circulating cancer biomarkers have entered the clinic in the last 30 years" (3). The problem of identifying novel cancer biomarkers must have been persistently underestimated (4, 5).

Underlying the difficulty is a general failure to validate clinically that a marker has utility. An ability to change a clinical decision arises from the marker's predictive value or its ability to discriminate among disease classifications. Such clinical validation differs from a technical validation. The latter is stringently provided when the results of one assay are in agreement with those of another, orthogonal, assay. For example, the results from hybridization-based mRNA profiling may indicate that the more aggressive tumors express gene X at a high level. A technical validation is provided when a PCR-based or a protein-based assay confirms the elevated levels of gene X in the same samples. Functional studies may further confirm technically that the activity of gene X is also elevated in the tumors. The investigators may claim, with some justification, that the markers had been "rigorously validated." Nonetheless, a subsequent attempt at a clinical validation, using a different panel of samples, may still find that gene X is not generally associated with tumor aggressiveness. This is possible because technical validations are inherently a circular process of discovery. For example, a finding in sample set Q will be technically validated as true when retested with varying methods applied in the same sample set, Q. When the newly discovered clinical association is tested in a different sample set, sample set R, the analytic strategy is no longer circular; clinical weaknesses may now be revealed.

It is typical to use high-dimensional data to identify new biomarkers for cancer. In this situation, the number of features determined for each sample (p, the quantity of parameters or dimensions in the data) greatly exceeds the sample size (n, the number of samples analyzed), producing the "large p, small n" or "$p >> n$" paradigm. Such data arise from exceedingly parallel assays such as microarrays, mass spectroscopy, or DNA sequencing. Most of these marker publications have not been validated clinically by outside investigators or, when they were, failed the external validation attempt (6). For example, among 35 detailed studies reporting a new molecular tumor-classifier (selected for studies having both a statistical cross-validation procedure used on an initial or "discovery" population and a validation attempt conducted on a separate "external" or "validation" population), which in turn were cited by a total of 758 subsequent publications, only one of the citations constituted an additional independent validation of the reported classifier (7). As another example, the NIH's Early Detection Research Network (EDRN) conducted a large validation study of 35 reported markers for early diagnosis of ovarian cancer to attempt to improve on the available but inadequate marker, CA125. They found none that surpassed the inadequate marker (8). Even among the top 5 markers in this study, unless the specimens were obtained within a few months before the time of the initial clinical cancer diagnosis, the sensitivity ranged from 0.73 to 0.40 at marker cutoff levels, producing a specificity of 0.95. Even this level of specificity was generous. That is, it presented the new markers in a favorable light. The 5% rate ($1 - 0.95 = 0.05$) of false positives would likely be prohibitively high if used to screen asymptomatic populations.

Why Is Marker Failure Important?

Failures in marker development equate to lost resources from consuming money, calendar years, labor, talent, and credibility for the field. The volume of misleading publications raises false hopes, poses ethical dilemmas, and triggers improper policy changes and purposeless debate.

An improved recognition of the patterns of failure should improve future performance of the field. As Heisenberg had recalled Einstein saying, "It is the theory which decides what can be observed." Optimal scientific procedures thus require that we consider a great diversity of theories to interpret any observed pattern, not excluding those theories that notify us of our failures, so that we can move toward a situation wherein "theory guides, but experiment decides." This diversity represents in a sense our marker-development vocabulary. Repeatedly, however, the literature in a biased manner presents the arguments in favor of marker successes; this bias will limit our vocabulary for communicating failures.

Biomarker development is typically a team process. Seldom, however, does a single team member command the topic completely. Some aspects of biomarker evaluation are obscure to most biomedical researchers. Omissions or conceptual blind spots among the team members can thus go undetected. It can be difficult to adjudicate among team members when differences of opinion arise as to the strategy, the final quality desired for the studies, and the scope and limitations of the conclusions justifiably drawn from the studies. A substantial attempt could be made to formulate a more complete common vocabulary. And because failures are more common and diverse than the successes in biomarker development, this effort could be voluminous. Such an effort to collate, to organize, and to teach this vocabulary is presented in the Supplementary document, titled "Flaws to Anticipate and Recognize in Biomarker Development," and builds upon focused prior efforts.

Systematic Attempts to Improve Marker Development and Adoption

A number of general road maps have been offered to guide successful marker development and offer a beginning vocabulary of marker failure. Pepe and colleagues described a

sequential series of 5 stages of biomarker discovery, validation, and implementation (9). Comprehensive reviews of the best practices were offered for proteomic biomarker development (10) and for the validation steps for molecular biomarker signatures (11). Professional associations facilitate the incorporation of new, evidence-based methods into standards of care; examples include the National Comprehensive Cancer Network Cancer Care Guidelines and the American Cancer Society Cancer Screening Guidelines.

Funding agencies such as the NIH launched supportive platforms for biomarker development. They created disease-focused opportunities, assembled qualified samples sets, and provided resources for reference laboratories. Efforts were manifest in the creation of the Early Detection Research Network, the fostering of team science by special funding mechanisms (SPORE and PO1 grants by the NIH), and efforts to lower regulatory and intellectual property hurdles.

Guidelines are available as prepublication checklists to be used in manuscript creation and peer review. These reflect broad to narrow interests: Guidelines for Reporting of Tumor Marker Studies (REMARK; 12), Standards for Reporting of Diagnostic Accuracy (STARD; 13, 14), Guidelines for Minimum Information for Publication of Quantitative Real-time PCR Experiments (MIQUE; 15), a checklist for statistical analysis and reporting of DNA microarray studies for cancer outcome (16), recommendations for improved standardization of immunohistochemistry and *in situ* hybridization (17, 18), guidelines for tissue microarray construction representing multicenter prospective clinical trial tissues (19), and recommendations to record and lock down the statistical methods used in biomarker discovery and validation (11).

Instructive anecdotes relating specifically to cancer biomarkers have been collected and lessons derived. Productive authors in this area have included Ransohoff, Diamandis, McShane, Pepe, Berry, Baggerly, and Coombes. The published commentary at the BioMed Critical Commentary website addresses flaws in hundreds of cancer biomarker publications.

The accompanying review (see the Supplementary document) incorporates an outline of lessons derived from the above sources.

In brief, investigators involved in developing a cancer biomarker are likely to encounter one or more of the following flaws. (i) The marker may provide a valid classification of disease or risk, and yet still lack clinical utility. Alternately, a marker may not provide a valid classification despite promising indications. For example, the marker may have appeared promising (ii) because of hidden structure in the data (a form of bias producing an invalid classification), (iii) because of the assay being technically inadequate, (iv) because of use of inappropriate statistical methods, (v) because of normal variation in the results dominating more than any possible useful information from the marker, (vi) because of a low prior probability of the marker being useful (i.e., an implausible marker), (vii) because of deficiencies inherent to the studied population, or (viii) because of deficiencies in the investigator system (including the research team, the research institution, the funding agencies, and the journal and its review process), or (ix) owing to the marker, perhaps quietly, having already been disproved or abandoned by other knowledgeable investigators.

Many of these categories reflect the influence of a bias. According to Ransohoff, "Bias will increasingly be recognized as the most important 'threat to validity' that must be addressed. . .a study should be presumed 'guilty'–or biased–until proven innocent. . . .Because there are many potential biases and some do not have consensus names or definitions, the process of identifying and addressing bias is not amenable to a simple checklist" (20). Many others reflect the error from statistical "overfitting" of high-dimensional data (11, 21) or from, as Berry warned, the ubiquitous and often insidious threat of multiplicities when the same patients have been analyzed repeatedly (22, 23).

Reasons for Failure: The List, with Expansions

Exploring the above categories of failure, one notes a remarkable complexity of failure modes.

1. ***A marker providing a truly valid disease classification may fail to be implemented due to lack of clinical utility.*** The marker by its nature may fail to justify changing the clinical decisions about the disease, or standards of care may be so ingrained as to unfortunately prohibit improvements. The distinct condition identified by the marker may be rare or may pose an analogous problem of excessive disease heterogeneity. The information conveyed by the marker may already be contained in the known clinical variables, and indeed the marker may reflect the comorbidities of the disease rather than the disease itself. The marker may differ among differing populations. The condition identified by the marker may be a poorly or undefined entity. The setting for biomarker discovery might have been poorly chosen such that it cannot translate to a realistic clinical setting. Use of the marker may lead to harm through overdiagnosis (24). The marker's utility may be limited to a temporary window of opportunity, one that is closing. The marker may not by nature be binary, impairing the certainty of its information. And it is possible that a marker, despite its theoretical strengths, may nonetheless not be implemented because of cost, inconvenience, or other disadvantages.

2. ***A marker may seem promising, and yet be invalid because it merely reflects a hidden structure in the original data.*** Such biases can arise in the studied populations, within the procedures used for the analysis (20), or from a covariation among variables that may not be handled well in the analysis. Some of these biases are investigator specific and persist because of omission of the necessary "blinding." Some are especially common to initial reports (the "beginner's luck bias" or "declining effect phenomenon").

3. ***The assay may be technically inadequate.*** The performance characteristics (sensitivity for the analyte, specificity, precision, and accuracy) may be unknown or suboptimal. The assay may be nonrobust, not locked down or reduced to a single standard procedure, or not portable from one laboratory to another. There may be no available orthogonal confirming assay or no appropriate control samples. The claimed performance of the assay may be impossible, its procedure may be inaccurately described, or its performance may not be

replicated. It may be a stopgap technology or impractically expensive. It may not be capable of distinguishing among opposite phenomena: the inhibition of the assay versus direct changes in the analyte under study. A single assay result may be insufficient to characterize a heterogeneous sample or for other scenarios in which conditions vary.

4. ***Inappropriate statistical and inferential methods may be used.*** A poor choice of method might have been made, the method itself may be a bad one, or a method may be misapplied. The statistical method(s) may be in flux, owing to a protocol having not been locked down before substantial data analysis. The conclusions drawn may reflect logical fallacies.

5. ***Normal variation may dominate the observations.*** The findings may reflect false-positive results of the type that can emerge from random patterns, i.e., patterns lacking informational content. Their emergence can be virtually guaranteed when voluminous data are "overfitted" to a relatively smaller number of subjects. False positives can evade recognition when there is a failure to correct for multiple statistical comparisons conducted (multiplicity), when one fails to respect the often-unfamiliar logic of discoveries arising from data mining, and from a lack of strenuous testing of independent validation set(s) of samples to attempt to verify the results arising in a discovery set of initial samples. These problems are accentuated when there are known and strong "false-positive" signals, and, alas, when there actually exists no "true-positive" marker to find.

6. ***The marker may be implausible.*** The 5 phases of marker development (9) are, in reality, insufficient, for they are preceded in all instances of biomarker development by an initial phase, which we can term "phase zero." Phase zero is an intellectual process to identify plausible sets of markers before conducting the phase I laboratory investigations of marker discovery. Phase zero can be understood in Bayesian theory as the recognition of a "prior probability" of success: the success or failure of a marker in phase I investigations will act upon and change this prior probability to a posterior (postphase I) probability. In phase zero, various marker categories can be assigned as "impossible," "improbable," or "plausible" [for example. Kuk and colleagues referred to the phase zero process as a filter (25)], although Ransohoff referred to a step of development that asks, "Might it work?" as contrasted with later steps where one might ask, "Does it work?" or "How does it work?" (26). Bayesian theory teaches that impossible markers, those having a prior probability of zero, cannot succeed, regardless of the accumulated empiric data supporting them. For improbable markers (those having a prior probability not much greater than zero), one would need much higher levels of supportive data as compared with initially plausible markers to achieve attractive posterior probabilities. Markers can be viewed as implausible on the basis of facts concerning innate properties (e.g., the mass of the marker produced in early disease may be very low), facts known about external features (e.g., there may be extreme variability in the reference "normal" values), deductive reasoning based on theoretical grounds (e.g., stochastic risks may not be predictable owing to being independent of prior conditions), inductive reasoning from historical patterns (e.g., the marker may belong to a generally unsuccessful marker category), or for other compelling reasons.

7. ***Deficiencies may exist in the studied populations.*** General defects would include, for example, small sample size. As an example of a more specific defect, a poor choice of the control tissue or population could, for example, create the straw-man argument (a too-easy comparison, a comparison of apples and oranges). Also, the subject population may lack the interesting disease states, or its data about outcomes may be ambiguous.

8. ***Deficiencies may exist in the investigator system.*** This system comprises the principal investigators, coinvestigators, and advisors; the research institution, colleagues, and the peer environment; the funding agencies and their preferences; journals and their review systems; and fads affecting the premises of the participants. The recent report of the Institute of Medicine (IOM) concerning the Duke biomarker scandal conveys examples of deficiencies in all of these aspects of the investigator system, in a series of vignettes revolving essentially around the procedural choices made by a very few principal investigators (11). About the full Duke biomarker episode, involving 162 coauthors across 40 publications, two thirds of which were at the time being partially or completely retracted, lead investigator and IOM editor Omenn noted that "not a single one of those investigators had raised a peep about the transparent flaws in these original papers" (27). In the absence of scandal or misconduct, the investigator system is likewise often at the heart of biomarker failure, as was emphasized by the IOM report and conveyed in its suggestions for improvements. The journals and funding agencies further impair research in their reluctance to publish or sponsor replication studies or experiments generating negative data (11, 28–30).

9. ***The marker, since its discovery, may be disproved or abandoned.*** Markers can be disproved by additional empiric evidence, by reanalysis of the original data, or by deductive reasoning. The initial investigators may be slow to accept the retirement of a marker. Conversely, the original authors may abandon a marker for just cause although the field continues to keep it in full usage for a (perhaps erroneous) diagnostic, predictive, or presumed causal role. Undeserved marker immortality can arise when there is reluctance to set numerical performance goals for success or to set "kill rules" for abandoning a marker.

Simple errors are commonly encountered during biomarker development. They may doggedly persist owing to an underappreciation of the immense diversity of such flaws and of the quiet erosive power they carry. An improved recognition of these patterns of failure could improve future biomarker development as well as biomarker retirement.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

This work has been supported by NIH grant CA62924 and the Everett and Marjorie Kovler Professorship in Pancreas Cancer Research.

Received August 15, 2012; revised September 10, 2012; accepted September 12, 2012; published OnlineFirst November 19, 2012.

References

1. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359–66.
2. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
3. Diamandis EP. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? BMC Med 2012;10:87.
4. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 2004;96:353–6.
5. Diamandis EP. Cancer biomarkers: can we turn recent failures into success? J Natl Cancer Inst 2010;102:1462–7.
6. Diamandis EP. Is early detection of cancer with serum biomarkers or proteomics profiling feasible? AACR Education Book; Philadelphia (PA): AACR; 2007. p. 129–32.
7. Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. Brief Bioinform 2011;12: 189–202.
8. Cramer DW, Bast RC Jr, Berg CD, Diamandis EP, Godwin AK, Hartge P, et al. Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. Cancer Prev Res (Phila) 2011;4:365–74.
9. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61.
10. Mischak H, Ioannidis JP, Argiles A, Attwood TK, Bongcam-Rudloff E, Broenstrup M, et al. Implementation of proteomic biomarkers: making it work. Eur J Clin Invest 2012;42:1027–36.
11. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials. Evolution of translational omics: lessons learned and the path forward. Washington, DC: Institute of Medicine of the National Academies; 2012.
12. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies. J Clin Oncol 2005;23:9067–72.
13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.
14. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative—Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003;49:1–6.
15. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 2009;55: 611–22.
16. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147–57.
17. Deutsch EW, Ball CA, Berman JJ, Bova GS, Brazma A, Bumgarner RE, et al. Minimum information specification for *in situ* hybridization and immunohistochemistry experiments (MISFISHIE). Nat Biotechnol 2008;26:305–12.
18. Goldstein NS, Hewitt SM, Taylor CR, Yaziji H, Hicks DG. Recommendations for improved standardization of immunohistochemistry. Appl Immunohistochem Mol Morphol 2007;15:124–33.
19. Rimm DL, Nielsen TO, Jewell SD, Rohrer DC, Broadwater G, Waldman F, et al. Cancer and Leukemia Group B Pathology Committee guidelines for tissue microarray construction representing multicenter prospective clinical trial tissues. J Clin Oncol 2011;29:2282–90.
20. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142–9.
21. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
22. Berry DA. Multiplicities in cancer research: ubiquitous and necessary evils. J Natl Cancer Inst 2012;104:1125–33.
23. Berry DA. Biomarker studies and other difficult inferential problems: statistical caveats. Semin Oncol 2007;34:S17–22.
24. Welch HG, Black WC. Overdiagnosis in cancer. J Natl Cancer Inst 2010;102:605–13.
25. Kuk C, Kulasingam V, Gunawardana CG, Smith CR, Batruch I, Diamandis EP. Mining the ovarian cancer ascites proteome for potential ovarian cancer biomarkers. Mol Cell Proteomics 2009;8: 661–9.
26. Ransohoff DF. Evaluating discovery-based research: when biologic reasoning cannot work. Gastroenterology 2004;127:1028.
27. Kaiser J. Clinical medicine. Biomarker tests need closer scrutiny, IOM concludes. Science 2012;335:1554.
28. Kern SE. No news is still news: Publishing negative results. J NIH Res 1997;9:39–41.
29. Brody JR, Kern SE. Stagnation and herd mentality in the biomedical sciences. Cancer Biol Ther 2004;3:903–10.
30. Carpenter S. Psychology research. Psychology's bold initiative. Science 2012;335:1558–61.