ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 1205–1219

Journal of Clinical Epidemiology

VARIANCE AND DISSENT

How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design

David F. Ransohoff

Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, CB# 7080, 4103 Bioinformatics Building, Chapel Hill, NC 27599-7080, USA

Accepted 12 April 2007

Abstract

Background and Objective: The search for molecular markers for cancer, using ''discovery-based'' techniques, has resulted in claims of a very high degree of discrimination both for cancer diagnosis (e.g., serum proteomics patterns) and prognosis (e.g., RNA expression genomic signatures). However, many promising initial results have been found to be unreliable or not reproducible, and the larger process of discovery can seem slow and inefficient. To improve the process to develop molecular markers, proposals to use ''phases'' and ''guidelines'' have been made, based on experience with the process of drug development and randomized controlled clinical trials.

The objective is to help improve the reliability and efficiency of development of molecular markers for cancer diagnosis.

Study Design and Setting: The literature was searched to identify important current problems (in serum proteomics for cancer diagnosis and RNA expression genomics for cancer prognosis) are identified, and the roles of tools (''phases,'' ''guidelines,'' and ''study design'') to address those problems are considered. Based on lessons learned, approaches for the future are discussed, some of which may seem ''radical'' compared with drug development.

Results: Phases identify and organize questions to be addressed by individual studies. Guidelines identify features of design and conduct to be reported so that each study's reliability can be judged. Study design involves the myriad details and choices involved in actual planning and conduct of a study. Study design is most important in the sense of determining whether a study is reliable or not. Studies that are unreliable, because of problems from chance and bias, constitute a major current problem leading to inflated expectations, wasted effort, and inefficiency in the larger process of development. By considering fundamental principles, it may be possible to identify approaches that are different than those used in drug development, while preserving reliability and efficiency.

Conclusion: Phases and guidelines have important roles, but issues in study design address the fundamental problems that compromise reliability and efficiency. Tools to study markers are underdeveloped and will evolve over time, perhaps to include seemingly radical approaches.

Keywords: Diagnosis; Proteomics; Genomics; Molecular markers; Screening; Prognosis

1. Introduction

Studies have claimed that molecular markers, identified by discovery-based proteomics and genomics research, provide a high degree of discrimination for cancer diagnosis and prognosis. However, some important claims about markers for diagnosis and prognosis have been unreliable—only weakly reproducible or not reproducible at all—and the process of development seems slow and inefficient. The purpose of this essay is to improve the reliability and efficiency of research about molecular markers.

The first part of this essay describes the problem of how and why strong initial claims have been shown to be unreliable. The second part discusses the relative roles of ''phases,'' ''guidelines,'' and ''study design'' in addressing problems. The third part discusses key lessons and approaches that may usefully be applied in future research about markers.

Problems in marker research discussed previously [1,2] are reviewed here and brought up to date in 2007. Some concepts in this essay have been previously reviewed or will be generally familiar to clinical epidemiologists; they are discussed here in detail for two reasons. The first is to facilitate collaboration and communication among diverse ''audiences'' that include clinical epidemiologists and laboratory-based researchers involved in interdisciplinary research about markers. The second reason is that a detailed discussion of problems and approaches may, for clinical epidemiologists, provide a useful foundation to develop new methodologies to improve future research about molecular markers.

E-mail address: ransohof@med.unc.edu

0895-4356/07/$ – see front matter
doi: 10.1016/j.jclinepi.2007.04.020

1.1. Discovery-based research and molecular markers

The approach of "discovery-based research" has become popular because of the increased availability of high-throughput technologies that can simultaneously measure thousands of genes or proteins. In this approach, "there is no need to identify targets a priori" [3]; instead, large portions of the genome or proteome may be examined for markers that can be used in diagnosis or prognosis. On the basis of this approach, claims have been made that serum proteomics markers for cancer diagnosis are nearly 100% sensitive and specific [4,5]. Similarly, RNA expression signature markers for prognosis [6,7] have been said to be "better than… clinicopathological methods" [8] and to "predict, with 90% accuracy, whether the tumour will remain localized or whether the patient will experience metastases and disease relapse" [9].

However, subsequent research has challenged the reliability of results, including whether some are even reproducible at all [10–13]. Problems involve not one technology or research group but rather entire "omics" fields. Although molecular markers hold great promise for diagnosis and prognosis, the process of discovery and evaluation has been neither reliable nor efficient. Reliability refers to whether a study's results are reproducible and are not explained by chance or bias, so that a study's conclusion can be used as a solid building block or foundation to ask other questions. Efficiency considers, in a roughly quantitative way, the amount of effort and amount of return involved in addressing a series of questions in the process of development. Concepts of reliability and efficiency are related in the sense that, if a study is not reliable, then a process that builds on that study's result will necessarily be inefficient. The concepts of reliability and efficiency may be used to understand and improve the evaluation of molecular markers.

1.2. Markers have many clinical uses

Although the phrase "molecular markers" might suggest the word "markers" has a clear meaning, the reality is that markers have many different clinical uses [14,15] that are studied using many different research strategies, so that a discussion of "marker research" risks being impossibly ambitious. For example, research about markers for the clinical purpose of diagnosis involves cross-sectional design and specific challenges of subject selection, comparison, and bias [16,17]. Studying markers for prognosis involves longitudinal design and a different set of challenges [18], while markers can be used for yet other purposes, like predicting response to therapy or measuring lifetime risk of disease [19]. Such diverse purposes and designs cannot possibly be considered in detail in one essay. At the same time, important overarching problems can be usefully identified and discussed. The main focus of this paper is on the clinical purpose of diagnosis or screening, but other clinical purposes are considered.

2. The problem: strong claims not reliable

2.1. Serum proteomics and cancer diagnosis

In 2002, "patterns" of mass spectroscopy peaks were said to diagnose cancers of the ovary, prostate, and colon with nearly 100% sensitivity and specificity [4,5,20–23]. Because such a high degree of diagnostic discrimination for cancer had never been achieved by a blood test and would be so important clinically, interest predictably followed among investigators, funding agencies, industry, and the public [24]. Plans were made to market a commercial blood test for ovarian cancer by early 2004 but were postponed [24,25]. Skepticism appeared in articles with titles like "New cancer test stirs hope and concern" [24], "Running before we can walk?" [26], and "Hope—and hype—in the cancer war" [27]. Skepticism was based not only on biological plausibility [28–32] but also on whether results had been shown to be reproducible in subjects other than those in whom the "discoveries" had been made [10,33–36]. When data publicly available from the initial ovarian cancer studies were reanalyzed, investigators concluded that "the method… does not establish reproducibility and performs no better than chance" [10]. Bias was also suggested to explain results if cancers and noncancers were analyzed on separate days by a mass spectroscopy machine whose calibration "wanders" over time, introducing signal differentially into the compared groups [10]. In response, the original investigators agreed that "reproducibility has not been demonstrated" and added that such an assessment had not been intended in the first place [37], thus raising questions about whether the initial enthusiasm based on the original reports had been warranted.

In a recent study, bias may similarly explain results when a serum proteomics test was reported to be nearly 100% sensitive and specific to diagnose prostate cancer [38]. On the basis of analysis of peptide products of exoprotease activity after blood collection, the study's results suggested that "future peptide biomarker discovery efforts" should focus on exopeptidases, according to the authors [38], whereas editorialists said that the low molecular weight "biomarker pipeline is surging with potential" [39]. However, a major bias may account for the discrimination observed. The persons with prostate cancer were, on average, 67 years old and of course were males; the comparison group was 30 years younger, and over half were female [40]. Although these differences are briefly discussed and discounted in the text [38] and in a subsequent letter [41], a reader could reasonably wonder whether such a bias, rather than exoprotease activities, explains the results [40]. A reader might also wonder whether other important differences occurred inadvertently in specimen collection and handling.

Although some studies may address chance and bias more persuasively [42,43], the degree of discrimination for cancer diagnosis demonstrated by serum proteomics technology is far from certain. One observer says "I do not understand the surge of activity in the search for biomarkers. The available evidence suggests that proteomics, despite almost a billion dollars investment, has so far failed to deliver any new biomarkers or commensurate returns. Many flagship companies have failed… (and) flawed studies, poor business models and exaggerated expectation will take time to reverse." (Walter Blackstock, University of Sheffield, at Royal Society of Chemistry, London 2006).

In sum, now 5 years after the initial reports, it is not clear that discovery-based serum proteomics approaches provide anywhere near the degree of discrimination initially claimed. Although it is common in science for initial enthusiasm to be tempered, this degree of both initial enthusiasm and subsequent tempering seems extreme.

2.2. RNA expression genomics and cancer prognosis

In genomics, research about transcriptional profiling to predict cancer prognosis seems to involve similar kinds of problems when a high degree of discrimination initially reported is not reproducible. RNA expression levels of thousands of genes are "profiled" to discover patterns, signatures, or specific genes whose up- or down-regulation predict cancer prognosis or response to therapy. In an early report about breast cancer, patients with "good prognosis" signatures had a 90% 5-year survival compared to 50% for the "poor prognosis" group [6]. Because the 40% absolute risk difference indicated greater discrimination than achieved by other prognostic systems, the study was reported in the New England Journal of Medicine [6], on the front page of the New York Times [44], and was said to illustrate a new approach "better than available clinicopathological methods" to predict cancer prognosis [8]. A letter to the editor, however, cautioned that "the degree of prognostic discrimination may have been inflated by the inclusion of patients from the training group" [45] in assessing reproducibility. Because "overfitting" can cause a multivariable model to appear to "discriminate" by chance (as described below), a model must be tested for reproducibility in subjects totally independent of those used to derive the model [2]. When the same investigators tested the original gene panel subsequently on totally independent subjects, the absolute risk difference was 20% [46], much less than the 40% difference originally reported.

Skepticism about transcriptional profiling has appeared in news reports with titles like "Getting the noise out of gene arrays" [47] and "An array of problems" [11]. When investigators tried to reproduce results of seven highly visible studies about cancer prognosis using the studies' original data, they concluded that "Five of the seven studies did not classify patients better than chance" [13]. In an accompanying editorial, provocatively titled "Microarrays and molecular research: noise discovery?," Ioannidis explained that investigators might have taken the original available study groups and divided them repeatedly into training and validation sets, "consciously or unconsciously select[ing] the best-performing pair of training-validation data and analytical mode" [12]. To prevent this problem of multiple testing, the editorial recommended that "…validations should be done by several completely independent teams. I cannot stress 'completely' enough here" [12]. Others have expressed similar concerns [48].

Some studies have addressed problems of chance more effectively [49], and the recent critique of microarray studies [12,13] has been questioned [50]. Nevertheless, the degree of discrimination and the degree of usefulness provided by transcriptional profiling is yet uncertain. The success or failure of the entire field cannot be fully discussed here, much less resolved, nor is that the point: the very fact that such strong challenges are being made about such prominent research is itself extraordinary, suggesting important problems in the process of evaluating molecular markers.

2.3. Reasons for optimism but caution

The message of this essay should not be interpreted as pessimistic. There is reason to be highly optimistic about finding useful markers, because of great advances in understanding biology along with development of powerful technologies to study that biology. It is now possible to identify or measure almost any DNA mutation or protein or peptide, and to do multiple measurements simultaneously. At the same time, new technologies and opportunities may tempt investigators to suspend "rules of evidence" [2] that historically have served critical roles in assuring reliability of research results. A central challenge—and an objective of this essay—is to learn how to explore new fields and technologies in a more reliable and efficient manner.

2.4. Threats to validity of clinical research

Although bias and chance are threats to the validity of all research, the problems are particularly severe for the nonexperimental or observational methods used to study diagnosis and prognosis [1,2]. Understanding details of these threats provides important background to understand and address current problems.

2.4.1. Threat from bias

Erroneous conclusions may be caused by bias when compared groups differ not because of a "signal" produced by cancer but because of an extraneous feature associated with cancer [1]. For example, if blood from cancer subjects is collected in red-top tubes in the oncology clinic while blood from noncancer subjects is collected in purple-top tubes in the screening clinic, then systematic differences could be caused by specimen collection and handling, not

by cancer [51,52]. Bias is the most difficult threat to validity because biases are so numerous, so difficult to avoid, and sometimes even difficult to identify [1], while even one bias can be fatal. Bias cannot be handled simply by ''annotation'' or statistical analysis or adjustment [1]. Bias must be handled by appropriate attention to design, conduct, and interpretation [1].

2.4.2. Threat from chance

Erroneous conclusions may be caused by chance when, for example, ''overfitting'' occurs in the multivariable analysis used in discovery-based research to ''fit'' large numbers of possible predictors (for example, thousands of mass spectroscopy peaks, gene-expression values, or candidate analytes) to a small number of outcomes [2]. Discovery of a ''pattern,'' ''signature,'' or analyte(s) may be caused by ''overfitting'' when powerful computers fit a multivariable analysis to the set of subjects from which the pattern was derived (sometimes called the ''training set''; see the area marked ''hypothesis generation'' in Fig. 1) [53]. If overfitting occurs, a model will not discriminate in an independent group (sometimes called the ''validation set''; see the area marked ''hypothesis testing'' in Fig. 1) [2]. To avoid overfitting, cross-validation is commonly done in ''hypothesis generation,'' in processes that statisticians call ''model selection'' [54,55] and ''prediction error estimation'' [56]. Demonstrating that overfitting has not occurred is accomplished by showing that discrimination is reproducible in subjects *totally independent* of those used in training. This approach is similar to when, in a laboratory, a result generated in one group of animals is considered a weak hypothesis that must be tested for reproducibility using a different set of animals. Nonreproducibility means that a ''discovery'' cannot be relied on as a foundation on which to build future research.

The issue of whether independent reproducibility has been successfully demonstrated is the focus of recent critiques of ''omics'' research [10,12,13]. Demonstration of reproducibility might seem to be an obvious part of the scientific process [2], but it is not often done [57]. One source of confusion is that the word ''validation'' is used to mean different things [2]. For example, the term ''leave-one-out cross-validation'' contains the word ''validation'' and may suggest ''independence'' when subjects are ''left out.'' However, as discussed above, ''cross-validation'' does not achieve the *independence* illustrated in Fig. 1. As shown in the figure, cross-validation may be used during the training step to increase the likelihood that results of training will not be overfitted and so will be confirmed in an independent sample set. But cross-validation does not replace the *totally independent* assessment required. As Simon writes, ''… there is abundant reason to demand… validation based on truly independent data'' [58] because cross-validation, done in generation of a discriminatory pattern, may be done so badly: ''[T]here are many examples of serious errors… of accuracy included in publications in the best journals'' [58].

An investigator can demonstrate independence using unambiguous language, as in a recent article, to describe the date when the methods (the discriminatory pattern, algorithm, or analytes) were ''finalized'' before assessment of a totally independent group: ''The prospectively defined assay methods and end points were finalized in a protocol signed on August 27, 2003. RT-PCR analysis was initiated

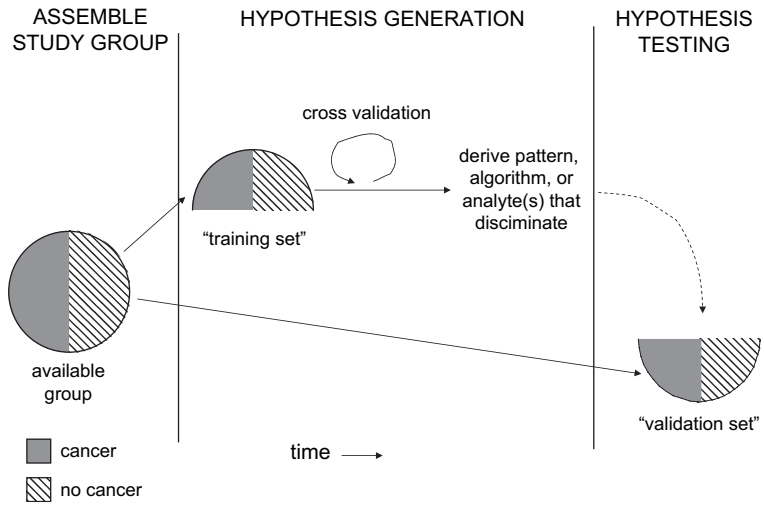

Fig. 1. Assessing reproducibility in independent subjects (modified from [2]): After a hypothesis is generated in ''discovery'' (the hypothesis consists of a discriminatory pattern, algorithm, or analyte[s]), it must be tested for reproducibility in an independent ''validation set.'' The goal is to learn whether the discovery or hypothesis may be explained by the overfitting or chance that commonly occurs in ''discovery-based research.'' Subjects used to test reproducibility must be *totally independent* of those used in hypothesis generation. Methods like ''leave-one out cross-validation'' are commonly used in hypothesis generation but do not constitute or substitute for hypothesis testing using independent subjects. The ''available study group'' may be divided only once into a training set and validation set. After analyzing subjects in the validation set, it is not permissible to return those subjects to the ''available study group'' and divide it again into training and validation sets.

on September 5, 2003, and… data were transferred… for analysis on September 29, 2003'' [49]. Such language should be routinely used by investigators and expected by reviewers and editors.

Demonstrating independent reproducibility is so critical that Ioannidis argues it should not be done by the original investigators because they may do it incorrectly; he says it should be done ''completely'' by others [12]. However, while assessment by others is obviously necessary in the long run, a strong case can be made for the original investigators to assess ''independent reproducibility'' themselves before others even try, using different subjects but exactly the same equipment, algorithms, reagents, and observers. The reason is that, when others assess reproducibility, new variables can make it impossible to interpret results. In other words, if others do not demonstrate reproducibility, the reason is unclear. Is it new variables like different reagents, different observers, or different kinds of subjects? Or is the instrument or approach itself—like the measurement of mass spectroscopy peaks and mathematical analysis—fundamentally flawed, perhaps by overfitting? The cleanest way to learn whether the instrument can be relied on or whether ''discrimination'' is due to overfitting is to test reproducibility in independent subjects while keeping all other things constant, as shown in Fig. 1.

2.5. *''Serial hypothesis generation''—how ''demonstrating reproducibility'' may not establish reliability*

Sometimes an effort to ''demonstrate reproducibility'' does not help establish reliability but rather just perpetuates confusion. The problem occurs when, in high-throughput ''omics'' research, investigators derive multiple different models (based on algorithms, patterns, or specific analytes) and conclude that the overall ''approach''—deriving models—works. The problem is that, until the reproducibility of each model (or any model) is demonstrated in an independent set of subjects, each model must be considered a ''hypothesis'' that has not yet been tested. The problem also occurs when machines or analytes are ''improved'' and evaluated by demonstrating that ''new hypotheses can be generated.'' Deriving multiple models, an approach that might be called ''serial hypothesis generation,'' serves no scientific purpose if overfitting may explain each ''discovery.''

The frequency with which serial hypothesis generation occurs in current research is difficult to determine, but the kind of problem can be illustrated when a statement is made that ''… it is possible to generate not just one, but multiple combinations of proteomic patterns from a single mass-spectral training set'' and that ''multiple combinations of proteomic patterns… are more than 98% sensitive and specific'' [59]; a reader must at least wonder whether results may be explained by serial hypothesis generation. Avoiding the problem is simple: investigators must ''lock down'' one specified pattern or analyte(s) before testing for reproducibility in independent subjects, as described above [49,60] and shown in Fig. 1.

3. Roles of phases, guidelines, and study design in drug development

Both drug and marker research use tools of ''phases,'' ''guidelines,'' and ''study design.'' The roles of these tools are generally familiar to clinical epidemiologists but are discussed in detail here in part to facilitate communication and collaboration among clinical epidemiologists and laboratory-based researchers working in this interdisciplinary area. Some laboratory researchers seem to think that ''phases'' and ''guidelines'' provide thorough prescriptions for conducting clinical research and do not fully understand the importance of study design.

3.1. *Overview of drug development process*

The drug development process includes two major stages. Discovery is conducted in the laboratory and is followed by clinical evaluation in human beings. The overall process is expensive and time consuming (see Fig. 2). Candidate compounds—sometimes numbering in the thousands—are identified and assessed in discovery, based on knowledge from disciplines like physiology and pharmacology and based on data from studies conducted in microbes, cell lines, and animals. Failure is routinely expected. To achieve efficiency and to reduce expense in discovery, one goal is to identify failure (e.g., compounds that are ineffective or toxic) as early as possible. Compounds that survive discovery enter a clinical evaluation stage involving three phases, as discussed below.

The process of discovery and evaluation may be considered ''mature'' based on four decades of experience of investigators and pharmaceutical companies. At the same time, the

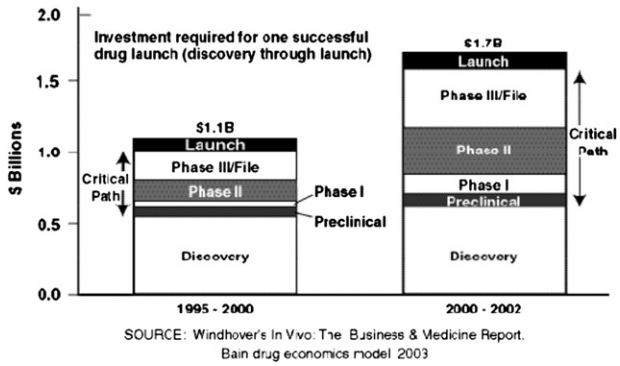

Fig. 2. Overview of the drug development process (from [61]): The drug development process consists of preclinical ''discovery'' followed by ''phases'' of testing in human beings. Although costly and cumbersome, the process has matured over decades of experience and provides the basis for recommendations for using ''phases'' in marker research.

process is widely acknowledged to be inefficient [61–64]. A Food and Drug Administration white paper entitled "Innovation or stagnation" described the dismal fact that "a new medicinal compound entering Phase 1 testing, often representing the culmination of upwards of a decade of preclinical screening and evaluation, is estimated to have only an 8 percent chance of reaching the market" [61]. Even though the process is far from perfect [61–64], it may be close to the best we have. The situation calls to mind Churchill's complaint that democracy is "the worst form of government, except for all those other forms…." (from a speech in the House of Commons, November 11, 1947).

The "phases" and "guidelines" that have evolved for drug research may be examined to learn about their potential application to marker research.

3.2. Phases

The purpose of phases is to identify and prioritize research questions, in order to improve the efficiency of the overall process of drug development [65,66]. Phase I asks about "safety of treatment" (and perhaps about pharmacokinetics for dosing) and involves a small number of human volunteers. If toxicity is found, work does not progress to the next phase. Phase II asks mainly whether there is any effect on "surrogate measurements of the clinical outcome"; if there is not, then work may not progress to the next phase. Phase III asks if "treatment improves the targeted condition," assessing both positive and negative health outcomes [65].

3.3. Guidelines

The purpose of guidelines is to help provide thorough reporting of methods and results so that a reader can judge the reliability or strength of a study. The CONSORT guidelines (Consolidated Standards of Reporting Trials guidelines), widely adopted for randomized controlled trials (RCTs) of drug therapy [67,68], are intended to improve the quality of reporting [69]. Other guidelines may deal with issues beyond reporting and even prescribe some details of study design; for example, the regulations of Good Clinical Practice [70] were developed in part "for the prevention of fraud in pharmaceutical studies" [71]. A Cochrane Collaboration handbook identifies details that can be used to judge "quality" of a clinical trial [72], but its length and complexity make it more of a textbook than a checklist that could be used by an investigator otherwise unfamiliar with study design.

3.4. Study design

If phases list questions and if guidelines prescribe details of reporting, then "study design" encompasses the myriad details that constitute the substance of the actual planning, conduct, and interpretation of a research study. Details involve diverse topics like formulation of the research question, overall strategy of comparison (e.g., cross-sectional study vs. cohort study vs. randomized controlled trial; prospective data collection vs. retrospective), selection of subjects, selection of the outcome, and anticipation of sources of bias and how they can be addressed in study design and conduct [73]. A phase may suggest a few features of study design, and a guideline may suggest details to report, but it is study design—details of planning and conduct—that determines, at the end of the day, the quality and reliability of the science.

3.5. Relationship of phases, guidelines, and study design

Phases, guidelines, and study design are related but have somewhat orthogonal roles. Phases help organize questions. Guidelines provide for detailed reporting of how those questions are addressed. Neither substitutes for study design. Study design cannot be prescribed by rules or checklists; the knowledge and detail involved are found in textbooks, courses, journals, and experience. Learning how to apply principles of study design is an acquired skill, similar to learning "clinical judgment." As useful as "phases" and guidelines may be, they would never be considered sufficient for an investigator to actually do an RCT. As obvious as this is to anyone involved in design and conduct of clinical studies, it may be useful to consider in detail for laboratory researchers unfamiliar with clinical research design, and for clinical epidemiologists who collaborate with them.

4. Roles of phases, guidelines, and study design in marker development

Phases, guidelines, and study design are less evolved for markers than for drugs. Given that limitation, we must consider in 2007 how best to conduct marker research. We also can consider how, in the future, to evolve these tools while improving reliability and efficiency [74]. We should be open minded; in the midst of "underdevelopment" and frustration, there may be unexpected opportunity for perhaps even radical approaches that might make some research about markers easier and more efficient than for drugs.

4.1. Phases

4.1.1. Phases for diagnostic tests

The use of "phases" in diagnostic test research has been discussed for years, even if only recently has it received intense attention. Feinstein proposed sequential stages [75], and Sackett described four "diagnostic research questions" including "Phase I questions: Do patients with the target disorder have different test results from normal individuals?… Phase II questions: Are patients with certain test results more likely to have the target disorder than patients

with other test results?… Phase III questions: Among patients in whom it is clinically sensible to suspect the target disorder, does the test result distinguish those with and without the target disorder?… Phase IV questions: Do patients who undergo the diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not?'' [76]. Knottnerus and Muris explain how complicated just the basic strategy of comparison may be in assessing diagnostic accuracy: ''general design'' can involve ''(1) a survey of the total study population, (2) a case-referent approach, or (3) a test-based enrolment'' [77]. Further, even basic approaches to data collection may vary dramatically: ''The direction of the data should generally be prospective, but ambispective and retrospective approaches are sometimes appropriate'' [77]. The complexity of these choices makes drug evaluation research seem simple and almost pedestrian. The point here is not to describe or to reconcile all complexities of design; it is to illustrate the magnitude of the complexity involved in studies of markers and the lesser ''evolution'' of the field.

4.1.2. Phases for biomarkers—a recent proposal

One proposal for ''phases'' is particularly important because it was specifically intended to be a ''formal structure to guide the process of biomarker development'' [66] and ''has been adopted by the EDRN [the National Cancer Institute's Early Detection Research Network], a number of Specialized Program of Research Excellence (SPORE) consortiums, and other biomarker research projects'' [78].

The proposal's purpose is organization and efficiency. ''The phases of research are generally ordered according to the strength of evidence that each provides in favor of the biomarker, from weakest to strongest. In addition, the results of earlier phases are generally necessary to design later phases'' [66]. ''In general, biomarker evaluation should follow an orderly process wherein one proceeds to the next phase only after meeting prespecified criteria for the current phase'' [78].

In contrast to the three relatively simple phases for drugs, the number of phases proposed for markers is very large. Not only are there five phases, (see Table 1, [66]) but each phase may ask multiple questions, adding up to more than 20 research questions and 20 different studies [66]. For example, secondary aim 1 of phase II involves questions as diverse as optimizing procedures for performing the assay, assessing reproducibility within a laboratory, assessing reproducibility between laboratories, and assessing whether the assay works well on stored clinical specimens [66].

The large number of questions has implications about time and effort and about ''where to start.'' To address only a few of the questions listed in one phase (phase II), research about stool DNA markers to diagnose colorectal cancer (CRC) required over 5 years of work in preparation for a large study of asymptomatic early-stage cancer [60]. Investigators may fairly ask whether phases have to be addressed in order. Although the proposal's authors say, ''[t]he process is not necessarily linear'' [66], it is not clear how other approaches may be done. This issue, discussed below, needs to be explored aggressively in the future by clinical epidemiologists.

4.1.3. Phases for biomarkers—limitations

The proposal provides a thoughtful start to what is likely to be a lengthy and productive dialogue about how to best organize questions. In the meantime it is important, particularly for laboratory-based investigators, to appreciate how the proposal for phases is ''orthogonal'' to study design. Phases may suggest overall strategies of design and comparison (like ''retrospective study'' or ''convenience sample''), but they do not provide the details needed to implement any strategy and to avoid threats from chance and bias. When a 6000-word essay discussing phases use the words ''chance'' and ''bias'' only three times [66], it simply indicates that those topics are orthogonal to the ones addressed in the essay. A person experienced in clinical research methods will understand such distinctions intuitively, but laboratory-based and ''translational'' researchers may mistakenly think that adherence to a ''phase'' or ''guideline'' provides the ''study design'' that leads to reliability of an individual study.

4.2. Guidelines

Guidelines for reporting have recently been developed for studies of diagnostic tests (STARD, Standards for Reporting of Diagnostic Accuracy [79]), for studies of molecular markers for cancer prognosis (REMARK, Reporting recommendations for tumor marker prognostic studies [80]), and for microarray research (MIAME, Minimum Information About a Microarray Experiment [81]). As for RCTs, however, the role of guidelines is to prescribe thoroughness of reporting, and not to prescribe the actual design and conduct of a study. Guidelines may help investigators report details of methods and results in a transparent way, but guidelines cannot prescribe exactly which details need to be made transparent. Choices about details depend on the insight and active involvement of the investigator. ''Guidelines for reporting cannot replace the thoughtful reflection and insight of an investigator who explicitly considers: what are all the possible systematic differences between compared groups that could explain the results; what measurements could be checked to see if those biases occurred; and, based on those measurements, can the direction and magnitude of possible biases be estimated, along with their impact on results and interpretation? This kind of detailed consideration should be provided by authors and expected by reviewers and editors, and it needs to be reflected in every step of research, from design and methods to results, analysis and interpretation'' [1]. Said another way, every possible counterexplanation should be candidly addressed by the investigator.

Table 1
Phases proposed for molecular marker evaluation (abbreviated from [66])
NOTE: The five phases proposed by Sullivan et al. [66] help organize research (primarily about markers of diagnosis) by listing over 20 questions, involving over 20 research studies. Using phases does not directly address threats from chance or bias that can affect each study. Because questions do not need to be addressed in strict order [66], efficiencies may be achieved by imaginative approaches to formulate and answer questions. For details, see text.

Phase I—Preclinical exploratory: promising directions identified
Primary aims
(1) To identify leads for potentially useful biomarkers.
(2) To prioritize identified leads.
Phase II—Clinical assay and validation: clinical assay detects established disease
Primary aim
To estimate the true-positive rate (TPR) and false-positive rate (FPR) or receiver operating characteristic (ROC) curve for…subjects with cancer from subjects without cancer.
Secondary aims
(1) To optimize procedures for performing the assay and to assess the reproducibility of the assay within and between laboratories…. In preparation for phase 3, the assay should also work well on stored clinical specimens.
(2) To determine the relationship between biomarker measurements made on tumor tissue (phase 1) and the biomarker measurements made on the noninvasive clinical specimen (phase 2). For example, one should confirm that patients with high expression of mRNA in tissue are the same patients for whom an associated biomarker protein is measured in serum.
(3) To assess factors, such as sex, age, smoking behavior, etc., that are associated with biomarker status or level in control subjects….
(4) To assess factors associated with biomarker status or level in cancer case subjects—in particular, disease characteristics such as stage, histology, grade, and prognosis….
Phase III—Retrospective longitudinal: biomarker detects disease early before it becomes clinical and a ‘‘screen positive’’ rule is defined
Primary aims
(1) To evaluate, as a function of time before clinical diagnosis, the capacity of the biomarker to detect preclinical disease.
(2) To define criteria for a positive screening test in preparation for phase 4.
Secondary aims
(1) To explore the impact of covariates on the discriminatory abilities of the biomarker before clinical diagnosis, including demographics, disease-related characteristics, and other clinical information about the subject….
(2) To compare markers with a view to selecting those that are most promising.
(3) To develop algorithms for screen positivity based on combinations of markers….
(4) To determine a screening interval for phase 4 if repeated screening is of interest.
Phase IV—Prospective screening: extent and characteristics of disease detected by the test and the false referral rate are identified
Primary aim
To determine the operating characteristics of the biomarker based on screening test in a relevant population by determining the detection rate and the false referral rate.
Secondary aims
(1) To describe the characteristics of tumors detected by the screening test—in particular, with regard to the potential benefit incurred by early detection….
(2) To assess the practical feasibility of implementing the screening program and compliance of screen-positive subjects with work-up and with treatment recommendations….
(3) To make preliminary assessments of the effects of screening on costs and mortality associated with cancer.
(4) To monitor tumors that occur clinically but that are not detected by the screening protocol.
Phase V—Cancer control: impact of screening on reducing the burden of disease on the population is quantified
Primary aim
To estimate the reductions in cancer mortality afforded by the screening test.
Secondary aims
(1) To obtain information about the costs of screening and treatment and the cost per life saved.
(2) To evaluate compliance with screening and work-up in a diverse range of settings.
(3) To compare different screening protocols and/or to compare different approaches to treating screen-detected subjects in regard to effects on mortality and costs.

4.3. Study design

As for drug therapy, it is study design, not phases or guidelines, that determines the reliability of studies about markers. However, the methods of observational epidemiology are underdeveloped in comparison to the experimental methods available in studies of therapy. In an experiment, randomization at baseline—as well as other methods—provides ways to avoid bias in the fundamental comparison and helps assure reliability of results. In contrast, randomization cannot be used in an observational study of a diagnostic or prognostic marker to assign cancer to one group of subjects while ‘‘holding other things equal.’’ It should be no surprise that the ‘‘…the methodology for evaluation of

diagnostics is not much crystallised, in contrast to the deeply rooted consensus regarding the principles of the randomised controlled trial on therapeutic effectiveness and the broad agreement on aetiologic study designs… [and] serious methodological flaws are often found in published diagnostic studies'' [19].

The methodological challenges of marker research are generic, unrelated to whether the marker is a molecule, and have been widely discussed [1,16,17,71,74,82–84]. Studies about molecular markers involve special considerations because biases may be idiosyncratic to the biology and technology being studied [1]. Discussions of molecular marker research include topics such as the order to address questions [66,85]; ways to avoid specific biases and to use retrospective or prospective sample collections [85]; quality of research [86]; and a list of ''do's and don'ts'' for conducting and reporting prognostic marker studies [87]. These discussions provide useful insights but cannot, in 2007, provide a coherent or widely accepted approach in what is necessarily a still-evolving field.

4.4. Relationship of phases, guidelines, and study design

The roles of phases, guidelines, and study design are the same as for drugs, although for studying markers each tool is less evolved than for its counterpart in drugs. Study design has, at the end of the day, the most fundamental role.

5. Fundamental problems in biomarker research

The problems in proteomics and genomics described above are limited in the sense of being ''case reports'' based on a few research studies that received sufficient documentation in the literature to let a reader see enough details of design and conduct to understand what problems may have occurred.

The problems are also limited in the sense that the illustrations address ''early'' phase I [77] or phase II [66] questions that ask, ''Do patients with the target disorder [i.e., cancer or poor prognosis] have different test results'' compared with persons without cancer or with good prognosis? [77] If answered ''yes,'' then other questions would be asked about standardizing the promising technology or about whether an intervention based on test results leads to improved outcome.

Even if studies are ''early,'' they can have substantial consequences if weak studies are considered sufficiently ''reliable'' to provide the foundation for asking other questions. A $100 million initiative in serum proteomics for cancer diagnosis is based on believing that early studies, showing strong discrimination of cancer vs. not, were ''reliable enough'' to warrant asking other questions about reproducibility (ability to identify specified proteins and peptides), computational methods, protein separation techniques, and standard reagents [88]. If that foundation turns out to have been unreliable—if there was no substantial discrimination to begin with—it may be argued that the large investment was premature. One advisor cautioned, in discussing whether to invest, ''Before we move on to big science, the standard is to expect proof of concept…. I just don't see it here'' [89]. It might have been more useful to first ''[demonstrate] proof of principle in a study where everything is very tightly controlled… [to] help us figure out whether the general approach is worthwhile'' [89].

Standardization of terminology, instruments, and reagents of course has an important role in technology development; the issue is ''when'' in the development process to do it. Standardization too early may be counterproductive by drawing effort away from more fundamental discovery [90]. In proteomics research, ''standardization'' might harmonize approaches between laboratories and machines to understand and resolve differences. But the major problem now in much ''omics'' research is not to understand and reconcile methods that different laboratories use to achieve discrimination; the problem is to learn whether *any* method achieves anywhere near the degree of discrimination it claims. Although for proteomics technologies an argument has been made that ''Standardization is absolutely essential if this [proteomic technology] is going to be credible and have a good chance at success'' [88], it is not currently clear ''what'' to standardize. Is it nomenclature, platforms, specimen collection and handling, patient characteristics, or something else? Do we know what portion of the broad dynamic range of proteins and peptides, regarding size or abundance, contains ''diagnostic information'' and where standardization should be focused? The clear demonstration of ''proof of principle'' could help determine where to address subsequent effort like standardization. Premature efforts to standardize much of the field of biochemistry could incur substantial opportunity cost and even be counterproductive if other questions, more fundamental, are not addressed [90].

In another example of possible reliance on weak initial results, a project to standardize the SELDI (surface-enhanced laser desorptionionization) platform showing strong discrimination in early reports in serum proteomics [4,5,20–23] has now incurred years of time, expense, and opportunity cost in a study about prostate cancer detection [91]. If the SELDI platform turns out not to provide any discrimination at all, then an exercise in standardization may be judged to have been premature and wasteful because it was based on unreliable early studies. In retrospect, more effort might have been devoted to determine whether the initial promising results were sufficiently reliable to warrant any effort in standardization.

Some genomics studies about prognosis, discussed above, may similarly have promised a higher degree of discrimination than can be reliably concluded. Time will tell whether costly later-phase studies, involving intervention based on the test result [92,93], were based on unreliable or overinterpreted initial results.

6. Ideas for the future

Understanding current problems may help to evolve approaches that improve the reliability of individual studies and the efficiency of the overall process of development.

6.1. Every study, even "early" ones, must be "reliable"

The idea that "every study should be reliable" sounds so fundamental that it should not need to be stated. Yet, focusing on reliability of individual studies is probably the single most important thing that could be done to improve much current "omics" research, because building on unreliable results can incur such inefficiencies and opportunity costs.

Reliability is determined by proper attention to study design and interpretation. No study can be perfect, but every study must be interpreted fairly. In particular, results must not be overinterpreted. The primary responsibility for interpretation and for specifying a study's limitations belongs to the investigator, although reviewers and editors have some role.

John Platt argues that what distinguishes productive scientific fields from unproductive ones is how a study's limitations—"what might have gone wrong"—are handled. In his brilliant essay "Strong inference" [94], Platt observed the pace of advances in molecular biology in the 1950s and 1960s and asked why some fields of science are "moving forward very much faster than others…."Productive fields, he said, are those where investigators systematically and thoroughly consider "alternative explanations" for results. The strong-inference "attitude is evident just in the style and language in which the papers are written. For example, in analyzing theories of antibody formation, Joshua Lederberg gives a list of nine propositions 'subject to denial,' discussing which ones would be 'most vulnerable to experimental test.'" Or "[o]n any given morning the blackboards of Francis Crick or Sidney Brenner… [will show] the hot new result just up from the laboratory or just in by letter or rumor. On the next line will be two or three alternative explanations, or a little list of 'what he did wrong.' Underneath… a series of suggested experiments or controls that can reduce the number of possibilities" [94]. Platt was saying that progress is based on considering alternative explanations and avoiding overinterpretation.

Identifying "what might have gone wrong" is particularly challenging in molecular marker research about diagnosis and prognosis where research is interdisciplinary and "rules of evidence" span laboratory science and observational epidemiology [1–3,95].

6.2. Rely on principles, not rules

How, then, to determine what could have gone wrong? The answers are in details of study design, not in phases and guidelines. A list of rules or "do's and don'ts" in design [87] can be helpful for some problems, but thoughtful application of principles of study design always has a critical role. Principles are broader than rules and can handle unfamiliar situations of the sort involved in interdisciplinary research about markers.

Instructive lessons about the relative roles of principles and rules can be learned from recent discussions about "standards" in the field of accounting. The Chief Accountant of the U.S. Securities and Exchange Commission noted, "An ideal accounting standard is one that is principle-based" because it requires reflection of substance, not "form." "Rule-based accounting standards provide extremely detailed rules that attempt to contemplate virtually every application of the standard. This encourages a check-the-box mentality… that eliminates judgments…. [making] it more difficult… [to] evaluate whether the overall impact is consistent with the objectives of the standard" [96].

Relying on principles does not mean that rules or checklists have no role. However, as Platt argued, the productive scientist asks, "what could have gone wrong" and then considers all possibilities. That approach requires understanding principles.

7. An example: working from principles

7.1. Overview

Working from principles allows one not only to better apply available tools, but also to consider new approaches, some of which may even seem "radical" or at least very different from currently available approaches like the generally linear progression suggested by "phases." Phases proposed for marker research provide a comprehensive list of questions that eventually must be answered, but the list is long. Although phases do not need to be done in order [66], how is an investigator to think about "where to start"; "what degree of reliability is required"; "what study design features are critical to satisfy that requirement"; and "how might specimens with those features be obtained." Weighing such choices and trade-offs calls on an investigator to be creative, aggressive, and opportunistic, and to work from an understanding of principles.

7.2. Achieving reliability in an "early" study

If the reliability of "early" studies may be weak, a case can be made to turn conventional wisdom on its head and to emphasize high-quality specimens even in early studies. As discussed below, this approach is now being "experimented with." The purpose here is to describe how the approach works, why it is important, and to suggest that it be aggressively pursued wherever possible.

This approach will not be feasible for all diseases or technologies, but it may be for some. In contrast, the approach cannot even be contemplated for drug research. To the extent that marker research is modeled after drug research, opportunities might be missed to find approaches

that achieve reliability while perhaps making some parts of marker research easier and more efficient than for drugs.

7.3. Example: serum proteomics to detect CRC

Suppose an investigator has a proteomics platform like mass spectroscopy that can be used to discover proteins or peptides that discriminate CRC vs. not. An investigator might start with an early question like primary aim of phase II (''To estimate the TPR and FPR… [in] subjects with cancer from subjects without cancer''; see Table 1) [66]. Although the proposal for phases does not prescribe details of study design, it suggests that ''selection based on convenience may be necessary early in phase 2'' [66].

7.4. Choosing specimens: convenience sample vs. strong unbiased sample

After choosing ''what question'' to ask, the selection of specimens involves making decisions, whether deliberately or not, about ''critical features of study design'' and ''degree of reliability.''

7.4.1. Convenience sample

''Convenience samples'' can be chosen in many ways. They provide a basic comparison (like CRC vs. not), but features that assure reliability may be compromised for the sake of convenience. An investigator must consider how to handle compromises, including ones so serious as to make a study not worth doing.

A convenience sample of people without CRC might be collected from the colonoscopy suite, using blood drawn before sedative medications are administered. Finding cancers in that setting, however, is logistically more difficult than for noncancers because the prevalence of cancer is so low—only a few per thousand. A setting with a higher CRC prevalence, like the cancer clinic, would be convenient but might involve differences that distort a proteomics assay: patients would not be fasting and would not have had a recent laxative prep like non-CRC subjects, and they might have had a recent biopsy (that led to referral to the cancer clinic) that could introduce ''signal'' into blood. Any difference might affect proteins in the blood; identifying and weighing all possibly relevant differences are a substantial challenge [1].

7.4.2. Strong unbiased sample

Problems of convenience samples can be avoided by using strong unbiased specimens. The best source is, in general, a ''prospective collection'' in which specimens are collected, processed, and stored soon before the diagnosis is known, because that provides a kind of ''blinding'' that helps assure uniform features of patients and handling. Such a collection is seldom convenient, however. In marker research, however, it may be possible to collect specimens in this manner, or to find appropriate specimens that have been ''already collected'' in the past and banked.

7.5. Are strong unbiased specimens already available?

Strong unbiased specimens sometimes may be already collected and available. Such specimens would be ''prospective'' in the sense of being collected before the diagnosis was established, but ''retrospective'' in the sense that the study may not be planned until after data are collected. (Feinstein uses the term ''retrolective'' to describe this approach [75].) Studying drugs retrospectively is almost never possible.

A typical source of such specimens—but not the only one—is an RCT because RCTs are prospective studies and pay close attention to uniform handling and blinding. RCTs with possible use in marker research include Prostate, Lung, Colon, and Ovary (PLCO) [97] and Women's Health Initiative (WHI) [98]. PLCO is an RCT assessing whether screening for cancers of the prostate, lung, colon, and ovary reduces cancer-specific mortality; approximately 150,000 subjects have been enrolled. Although not needed to learn whether screening reduces mortality, a biorepository was established to address other questions. Bloods were drawn periodically, some near the time before a cancer was diagnosed. The WHI has a similar biorepository. Specimens are also being collected prospectively in this way in non-RCT settings [99]. Whether such specimens ultimately are useful depends on many things, including details like suitability of specimen collection (e.g., time to spinning/freezing) and storage.

7.6. Possible uses of strong unbiased specimens

In marker research, strong unbiased specimens might be used in several ways.

7.6.1. "Validation"

Specimens could be used of course in ''validation,'' to test a ''hypothesis'' that specific analyte(s) or algorithm can discriminate CRC vs. non-CRC [3].

7.6.2. "Discovery" and "validation"

Specimens could be used for both discovery and validation. ''Discovery'' might be done in a training set of CRC and non-CRC specimens (e.g., 100 samples of each; sample size estimation, however, is a separate topic), whereas ''validation'' is done in a separate, blinded set of CRC and non-CRC specimens (again, e.g., 100 of each) [2]. Using the same source for both discovery and validation can incur limitations in generalizability, but it provides reliability by demonstrating that overfitting has been avoided and by providing strengths against bias.

7.6.3. Test multiple technologies simultaneously: the "bakeoff"

Multiple technologies might be tested simultaneously using the same set of samples. Feasibility depends on details of logistics, but the approach can at least be contemplated in evaluating markers. In contrast, drugs almost always must be tested one at a time. To learn whether 100 different technologies could "discriminate" CRC vs. not, each could be assessed using the same specimens, as long as there were enough samples with enough volume. If "discovery" and "validation" could be done on small aliquots of serum (50 μL can be used by some technologies), then constructing 100 sets of samples for training (with 100 CRC and 100 non-CRC) and 100 sets for validation would require 5 mL from 400 subjects. The logistics are challenging but might be feasible; in contrast, testing 100 drugs simultaneously could never be done.

The concept of simultaneously testing multiple technologies is starting to be explored in several settings. The idea is to conduct a kind of "bake-off" in which identical sets of ingredients—strong unbiased specimens—are circulated to laboratories around the country or around the world. Some investigators in NCI's (National Cancer Institute) Early Detection Research Network are assessing whether four different serum proteomics technologies can diagnose colon cancer using the same specimens, (D. Brenner, personal communication), whereas others are conducting a study to learn whether different laboratories using the same proteomics technology can diagnose prostate cancer, using identical sets of specimens [91]. WHI specimens are being used to assess different serum proteomics technologies to diagnose colon cancer (S. Hanash, personal communication), and specimens specially collected from a non-RCT setting are being banked for future use in studies of diagnosis [99].

The ability to assess multiple technologies simultaneously is a unique feature of marker research compared to drug research and needs to be aggressively explored.

7.7. A late phase intervention study might not even need to be done

After an early study demonstrates that a marker can "discriminate" (that a test can diagnose early cancer or can predict poor prognosis), it might not even be necessary in some circumstances to do a "late phase" study. Demonstrating discrimination does not automatically translate to improved outcome, of course; answering that question requires a "late phase" study to assess whether a test result that prompts an intervention (like colonoscopy or treatment) improves an outcome like mortality. For example, the primary aim of phase 5 is "[t]o estimate the reductions in cancer mortality afforded by the screening test" [66].

However, in marker research it may not always be necessary to conduct that later study if proof of principle—that intervention improves outcome—has already been established by an RCT. RCTs have already shown that CRC mortality is reduced by fecal occult blood test screening, presumably by identifying early CRCs and advanced adenomas that, left untreated, will progress to fatal CRC [100–102]. In considering "evaluation of new test technologies" after RCTs have shown that CRC mortality is reduced by fecal occult blood test screening, the authors of practice recommendations concluded, "…it might be appropriate in the future to substitute a newer test… for currently recommended ones if there is convincing evidence that the new test has… comparable performance (e.g., sensitivity and specificity) in detecting cancers or adenomatous polyps at comparable stages…. [and that] it would not be necessary to submit each new technology to the original standard of proof, i.e., a randomized controlled trial with death from CRC as an outcome measure" [103].

8. Conclusion

Although molecular markers for cancer diagnosis and prognosis hold great promise, the methodology to study them is underdeveloped. If promising initial results are unreliable but are used as a foundation, the overall process of development may be inefficient. Although tools of "phases" and "guidelines" are useful, "study design" has the most critical role in addressing fundamental problems of reliability. To improve reliability, it may be useful to turn conventional wisdom on its head and emphasize strong unbiased samples, rather than convenience samples, in early studies. Although it is difficult to achieve "reliability" in marker research, some good news is that, once strong unbiased specimens are available, it may be possible to test multiple technologies simultaneously. In the future, methods to improve the reliability and efficiency of the process of marker development should be aggressively explored.

Acknowledgments

Thanks to colleagues at the University of North Carolina at Chapel Hill, the NCI and elsewhere for reviewing and commenting on earlier versions of the manuscript. Many ideas were developed through participation in activities of the NCI's Early Detection Research Network.

References

[1] Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5(2):142–9.
[2] Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4(4):309–14.
[3] Ransohoff DF. Developing molecular biomarkers for cancer. Science 2003;299(5613):1679–80.
[4] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359(9306):572–7.

[5] Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach JS. Detection of cancer-specific markers amid massive mass spectral data. Proc Natl Acad Sci USA 2003;100(25):14666–71.
[6] van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347(25):1999–2009.
[7] Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361(9369):1590–6.
[8] Ramaswamy S, Perou CM. DNA microarrays in breast cancer: the promise of personalised medicine. Lancet 2003;361(9369):1576–7.
[9] Bernards R, Weinberg RA. A progression puzzle. Nature 2002;418(6900):823.
[10] Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 2005;97(4):307–9.
[11] Frantz S. An array of problems. Nat Rev Drug Discov 2005;4: 362–3.
[12] Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet 2005;365(9458):454–5.
[13] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365(9458):488–92.
[14] Clark GM. Interpreting and integrating risk factors for patients with primary breast cancer. J Natl Cancer Inst Monogr 2001;30:17–21.
[15] Hayes DF, Bast RC, Desch CE, Fritsche H Jr, Kemeny NE, Jessup JM, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst 1996;88(20):1456–66.
[16] Sackett DL, Haynes RB. The architecture of diagnostic research. Br Med J 2002;324(7336):539–41.
[17] Ransohoff DF. Challenges and opportunities in evaluating diagnostic tests. J Clin Epidemiol 2002;55:1178–82.
[18] Concato J. Challenges in prognostic analysis. Cancer 2001;91(8 Suppl):1607–14.
[19] Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. p. 1–17.
[20] Petricoin EF 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, et al. Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2002;94(20):1576–8.
[21] Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 2002;62(13):3609–14.
[22] Vlahou A, Schellhammer PF, Mendrinos S, Patel K, Kondylis FI, Gong L, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001;158:1491–502.
[23] Drake RR, Manne U, Bao-Ling A, Ahn C, Cazares L, Semmes OJ, et al. SELDI-TOF-MS profiling of serum for early detection of colorectal cancer. Gastroenterology 2003;124(Suppl 1):A650.
[24] Pollack A. New cancer test stirs hope and concern. New York Times February 3, 2004;. [D1, D6].
[25] Marcus A. Testing for ovarian cancer is on the way. Wall St J October 1, 2002;. [D1, D2].
[26] Check E. Running before we can walk? Nature 2004;429:496–7.
[27] McCullough M. Hope—and hype—in the cancer war. Phila Inquirer 2005;. [Sect. A1, A12–A13].
[28] Diamandis EP. Re: Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2003;95(6):489–90. [author reply 90–91].
[29] Diamandis EP. Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? Clin Chem 2003;49:1272–5.
[30] Diamandis EP. OvaCheck: doubts voiced soon after publication. Nature 2004;430(7000):611.
[31] Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics 2004;3(4):367–78.
[32] Garber K. Debate rages over proteomic patterns. J Natl Cancer Inst 2004;96(11):816–8.
[33] Baggerly KA, Coombes KR, Morris JS. Bias, randomization, and ovarian proteomic data: a reply to "Producers and consumers". Cancer Inform 2005;1:9–14.
[34] Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20:777–85.
[35] Baggerly KA, Edmonson SR, Morris JS, Coombes KR. High-resolution serum proteomic patterns for ovarian cancer detection. [letter to the editor]. Endocr Relat Cancer 2004;11:583–4.
[36] Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. J Natl Cancer Inst 2005;97(4):315–9.
[37] Liotta LA, Lowenthal M, Mehta A, Conrads TP, Veenstra TD, Fishman DA, et al. Importance of communication between producers and consumers of publicly available experimental data. J Natl Cancer Inst 2005;97(4):310–4.
[38] Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, et al. Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest 2006;116:271–84.
[39] Liotta LA, Petricoin EF. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. J Clin Invest 2006;116(1):26–30.
[40] Diamandis EP, Kulasingam V, Sardana G. Letter to the editor about differential exoprotease activities confer tumor-specific serum peptidome. Accessed February 21, 2006. Available at. http://www.jci.org/cgi/eletters/116/1/271. J Clin Invest 2006.
[41] Tempst P, Villanueva J, Shaffer DR, Lilja H, Scher HI. Response to 'Letter to the Editor' by E.P. Diamandis, V. Kulasingam, G. Sardana. Accessed February 21, 2006. Available at. http://www.jci.org/cgi/eletters/116/1/271. J Clin Invest 2006.
[42] Zhang Z, Bast RC Jr, Yu Y, Li J, Sokoll LJ, Rai AJ, et al. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res 2004;64(16):5882–90.
[43] Mor G, Visintin I, Lai Y, Zhao H, Schwartz P, Rutherford T, et al. Serum protein markers for early detection of ovarian cancer. Proc Natl Acad Sci USA 2005;102(21):7677–82.
[44] Kolata G. Breast cancer: genes are tied to death rates. New York Times December 19, 2002;. [A1].
[45] Ransohoff DF. Gene-expression signatures in breast cancer. N Engl J Med 2003;348:1715–7. [author reply-7].
[46] Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98(17):1183–92.
[47] Marshall E. Getting the noise out of gene arrays. Science 2004;306: 630–1.
[48] Tibshirani R. Immune signatures in follicular lymphoma. N Engl J Med 2005;352:1496–7. [author reply-7].
[49] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2005;351(27):2817–26.
[50] Aliferis CF, Statnikov A, Tsamardinos I, Schildcrout JS, Shepherd BE, Harrell FE Jr. Gene expression microarrays do predict outcome. Technical Report 2007.
[51] Grizzle WE, Semmes OJ, Bigbee W, Zhu L, Malik G, Oelschlager DK, et al. The need for review and understanding of SELDI/MALDI mass spectroscopy data prior to analysis. Cancer Bioinform 2005;1:86–97.
[52] Mitchell BL, Yasui Y, Li CI, Fitzpatrick AL, Lampe PD. Impact of freeze-thaw cycles and storage time on plasma samples used in mass spectroscopy based biomarker discovery projects. Cancer Bioinform 2005;1:25–31.

[53] 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(1):14−8.
[54] Burman P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 1989;76:503−14.
[55] Breiman L, Spector P. Submodel selection and evaluation in regression. The X-random case. Int Stat Rev 1992;60:291−319.
[56] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21(15): 3301−7.
[57] Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362(9394):1439−44.
[58] Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23(29):7332−41.
[59] Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA. Clinical proteomics: translating benchside promise into bedside reality. Nat Rev Drug Discov 2002;1(9):683−95.
[60] Imperiale TF, Ransohoff DF, Itzkowitz SH, Turnbull BA, Ross ME. Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. N Engl J Med 2004;351:2704−14.
[61] Food and Drug Administration. Innovation or stagnation: challenge and opportunity on the critical path to new medical products. Available at. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html 2004. (accessed Aug 25, 2007).
[62] Cohen FJ. Macro trends in pharmaceutical innovation. Nat Rev Drug Discov 2005;4(1):78−84.
[63] Williams D. The critical path to medical innovation. Med Device Technol 2004;15(5):8−10.
[64] Couzin J. The new math of clinical trials. Science 2004;303(5659): 784−6.
[65] Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing clinical research: an epidemiologic approach. 2nd edition. Philadelphia, PA: Lippincott Williams & Wilkins; 2001.
[66] Sullivan Pepe M, 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(14):1054−61.
[67] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134(8):663−94.
[68] Rennie D. How to report randomized controlled trials. The CONSORT statement. JAMA 1996;276(8):649.
[69] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276(8):637−9.
[70] U.S. Food and Drug Administration. FDA regulations relating to good clinical practice and clinical trials. Available at. http://www.fda.gov/oc/gcp/regulations.html. (accessed Aug 25, 2007).
[71] Buntinx F, Knottnerus JA. Are we at the start of a new era in diagnostic research? J Clin Epidemiol 2006;59(4):325−6.
[72] Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 4.2.6;. Updated September 2006. Accessed February 22, 2007. Available at. http://www.cochrane.org/resources/handbook/hbook.htm.
[73] Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford, England: Oxford University Press; 2003.
[74] Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002.
[75] Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia, PA: WB Saunders; 1985.
[76] Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. p. 19−38.
[77] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. p. 39−59.
[78] Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics 2004;5:709−19.
[79] 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. Ann Intern Med 2003;138(1):40−4.
[80] 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):1180−4.
[81] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29(4):365−71.
[82] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. [Review; 32 refs]. J Clin Epidemiol 1995;48:119−30.
[83] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003;56:1118−28.
[84] Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005;58(1):1−12.
[85] Baker SG, Kramer BS, McIntosh M, Patterson BH, Shyr Y, Skates S. Evaluating markers for the early detection of cancer: overview of study designs and methods. Clin trials (London, England) 2006;3(1):43−56.
[86] Bogardus ST, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. JAMA 1999;281:1919−26.
[87] 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(2):147−57.
[88] Hede K. $104 million proteomics initiative gets green light. J Natl Cancer Inst 2005;97:1324−5.
[89] Goldberg KB. Advisors reject NCI's $89 million plan for proteomics as too much, too soon. The Cancer Lett 2005;31(10):1−10.
[90] Martin RL. Breakthrough ideas for 2005: seek validity, not reliability. Harv Bus Rev 2005;83(2). 23−4, 32.
[91] Semmes OJ, Feng Z, Adam BL, Banez LL, Bigbee WL, Campos D, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. Clin Chem 2005;51:102−12.
[92] Bogaerts J, Cardoso F, Buyse M, Braga S, Loi S, Harrison JA, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 2006;3(10): 540−51.
[93] Mauriac L, Debled M, MacGrogan G. When will more useful predictive factors be ready for use? Breast 2005;14:617−23.
[94] Platt JR. Strong inference. Science 1964;146(3642):347−53.
[95] Ransohoff DF. Research opportunity at the interface of molecular biology and clinical epidemiology. Gastroenterology 2002;122:1199.
[96] Herdman RK. Testimony concerning the roles of the SEC and the FASB in establishing GAAP; before the House Subcommittee on Capital Markets, Insurance, and Government Sponsored Enterprises, Committee on Financial Services. Accessed February 11, 2007. Available at. http://www.sec.gov/news/testimony/051402tsrkh.htm 2002.
[97] Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, Crawford ED, et al. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Control Clin Trials 2000;21(6 Suppl):273S−309S.

[98] Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med 2003;349(6):523–34.

[99] Skates SJ, Horick NK, Moy JM, Minihan AM, Seiden MV, Marks JR, et al. Pooling of case specimens to create standard serum sets for screening cancer biomarkers. Cancer Epidemiol Biomarkers Prev 2007;16:334–41.

[100] Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. N Engl J Med 1993;328:1365–71.

[101] Hardcastle JD, Chamberlain JO, Robinson MHE, Moss SM, Amar SS, Balfour TW, et al. Randomised controlled trial of faeca-l-occult-blood screening for colorectal cancer. Lancet 1996;348: 1472–7.

[102] Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. Lancet 1996;348:1467–71.

[103] Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, et al. Colorectal cancer screening: clinical guidelines and rationale. Gastroenterology 1997;112:594–642.