OPINION

Bias as a threat to the validity of cancer molecular-marker research

David F. Ransohoff

Abstract | Claims that molecular markers can accurately diagnose cancer have recently been disputed; some prominent results have not been reproduced and bias has been proposed to explain the original observations. As new '-omics' fields are explored to assess molecular markers for cancer, bias will increasingly be recognized as the most important 'threat to validity' that must be addressed in the design, conduct and interpretation of such research.

Finding a non-invasive marker for cancer has been an important goal of cancer diagnosis research for three decades[1]. Recent advances in molecular biology promise markers that will establish diagnosis, predict prognosis and provide insights into the aetiology of cancer. New technologies allow measurements of DNA and proteins to be carried out in great detail and in large numbers ('high-throughput'). These advances have made 'discovery-based' research — in which measurements of proteins, mutations and gene expression are made without a specific hypothesis[2] — feasible and popular. The data generated in this way can be used to identify specific proteins or expressed genes that can then be assessed as potential markers[1]. Alternatively, the data might be analyzed to derive signatures or 'patterns' that are then used as the test or marker, without the need to understand exactly which proteins or genes account for the pattern[1,2].

The stakes are high for successful non-invasive cancer tests, particularly for ovarian and other cancers for which no good screening method exists. In one prominent example, a discovery-based 'pattern-recognition' serum-proteomics approach was reported to discriminate, with nearly 100% sensitivity and specificity, people with and without cancer of the ovary[3,4]. The same approach has been used to identify different patterns of protein markers for other cancers, including breast[5] and prostate cancers[6,7]. Based on these promising initial results[3,8], commercial groups announced plans to market a blood test for ovarian cancer in the first quarter of 2004[9,10]. However, plans have been delayed by the United States Food and Drug Administration[11–13]. A number of concerns have been raised in scientific journals and the lay press[10,15–22] about whether results are reproducible and effective[10,14].

In the meantime, some observers have suggested that the pattern-recognition serum-proteomics approach is not biologically plausible because some proteins or peptides might be too small to be biologically informative[16,18] or because the original results might be due to bias[19,20]. Bias can occur if the cancer and non-cancer groups are handled in systematically different ways, introducing an apparent 'signal' into one group but not the other. Such differences might be introduced at several stages, including specimen collection, handling and storage, or during mass spectroscopy[17,20–22]. Similarly, in discovery-based genomics research, RNA expression patterns have been reported to predict the prognosis of breast cancer[23,24] "…better than any available techniques"[25]; however, results may have been distorted by problems in design and analysis[26], and recently the reproducibility of RNA expression array results has been questioned[27]. Indeed, the conduct, reporting and interpretation of high-throughput discovery-based '-omics' research is complicated enough to have created "...a breed of 'forensic' statisticians, who doggedly detect and correct" (REF. 30; see also REFS 28,29) possible errors in published reports.

The controversy over serum proteomics and ovarian cancer raises questions not only about whether discovery-based pattern-recognition serum proteomics can diagnose ovarian cancer (or any cancer), but, more importantly, about the process through which discovery-based research is designed, conducted and interpreted[1,31,32]. Beyond proteomics and genomics, multiple new '-omics' fields, with names like transcriptomics, metabolomics, epigenomics and ribonomics, will similarly be explored for molecular markers for the diagnosis and prognosis of cancer and of other diseases.

A previous article discussed how chance — specifically, the problem of overfitting (BOX 1) — can threaten the validity of molecular-marker research[31]. This Perspective article considers the even more important problems caused by bias.

Experimental and observational design

As summarized by Hulley and colleagues, a fundamental decision when designing studies for scientific research is "…whether to take a passive role in the events taking place in the study subjects in an observational study, or to apply an intervention and examine its effects on those events in a [randomized] clinical trial."[33] The experimental (intervention) method provides more effective ways to deal with bias than the observational method. In clinical research, the heterogeneity of groups studied might provide particularly problematic sources of bias when groups of participants differ in ways that can affect outcome. By contrast, in a laboratory setting, the subjects might be genetically-identical cell lines or animals, and it is often possible to tightly monitor and control the conditions that affect outcome. Although the experimental method cannot be used to address many kinds of research questions, including those concerning diagnosis and prognosis, the methods used to avoid bias in an experiment or randomized controlled trial (RCT) can provide useful lessons about how to avoid bias in observational research.

The central principle of an experiment, like a RCT, involves arranging a fair and unbiased comparison by the "...creation of duplicate sets of circumstances in which only one factor that is relevant to the outcome varies, making it possible to observe the effect of variation on that factor."[34] This principle — which can be difficult or impossible to apply in observational research — involves keeping all variables 'equal' between groups, except for the agent being examined. This principle is applied at every step of research — design, conduct and interpretation — so that any difference between the groups can be attributed to the agent, not to bias. The agent in an experiment or RCT in humans is typically a therapy, and in animals it might be an infectious organism or an induced mutation. Although observational research, such as that used to study molecular markers for diagnosis or prognosis, differs from an experiment in purpose and design, the general approach to address bias is similar, and lessons learned from RCTs can be applied directly to non-experimental observational research.

Threats to validity in clinical research

The threats to validity that affect clinical research[35] fall into three broad areas: chance, bias and 'generalizability'. The threat from chance, referred to as overfitting, is discussed in REF. 31 and summarized in BOX 1. The threat from generalizability concerns to whom the results of a comparison can be applied and is described in BOX 1.

Bias refers to a completely different kind of threat. Compared with threats from chance or generalizability, bias is more difficult to address in study design, conduct and interpretation. Bias is unintentional and unconscious. It is defined broadly as the systematic erroneous association of some characteristic with a group in a way that distorts a comparison with another group. For example, in a study designed to assess a test's ability to discriminate groups with and without cancer, if people in the cancer group are 70 years old and those in the comparison group are 25 years old and if the test result depends on age, then bias could account for an ability to discriminate between the two groups. Although this bias is obvious, many others are not. Bias can be so powerful in non-experimental observational research that a study should be presumed 'guilty' — or biased — until proven innocent. As noted by Cole, bias is a "…plague upon the house of epidemiology"[36] and even a single bias might result in errors sufficiently large to invalidate results. The case for a research study's innocence is made through the process of design, conduct, interpretation and reporting, as discussed below.

Biases might pose a special challenge for laboratory researchers who are used to biological reasoning and the tightly controlled conditions of experimental research. Such researchers unwittingly become non-experimental observational epidemiologists when they apply molecular assays in studies of diagnosis and prognosis[37], for which the experimental method is not available and for which biological reasoning might have limited usefulness[38].

The potential for bias to affect results and interpretation cannot be addressed by a simple process in the way that adjusting sample size can address type I or II errors. The process is more complicated and involves making everything equal during the design, conduct and interpretation of a study, and reporting those steps in an explicit and transparent way.

Importance and difficulty of avoiding the three threats. Chance and bias are more important threats to validity than generalizability because they affect the fundamental comparison — or the internal validity — of the study, and so must be addressed in every study. In contrast, generalizability concerns to whom the comparison fairly applies. For example, an experiment that compares the results from two groups of rats that receive a drug versus a placebo might be conducted in a way that successfully avoids problems from chance and bias (that is, it has internal validity), but the results might have limited generalizability and not be applicable to other strains of rats, to mice or to humans.

Problems caused by chance and generalizability can be addressed in a conceptually straightforward way; although logistically it can be difficult to find sufficient numbers of appropriate participants, at least investigators (and reviewers and editors) understand exactly what needs to be done. By contrast, problems caused by bias are more difficult to address both conceptually and logistically. First, it might be difficult to identify which biases are important and need to be addressed. Next, it might be difficult to design a study in a way that minimizes biases, to conduct measurements that check whether biases have occurred, and to interpret results by explicitly considering the potential magnitude and impact of each bias on results[35]. If this process sounds challenging, it is; large portions of epidemiology texts and courses are devoted to understanding

Box 1 | **Threats to validity: chance and 'generalizability'**

Chance

Chance can threaten validity by leading to erroneous conclusions in several ways. Investigators are probably familiar with how type I error can cause the erroneous or false-positive conclusion that there is a difference between compared groups when no difference exists. Similarly, type II error can result in the false-negative conclusion that there is no difference when a difference does exist. While neither error can ever entirely be avoided, a simple method to decrease their likelihood is to increase sample size[33].

Less-familiar to investigators is a problem caused by chance that can occur in discovery-based '-omics' research to identify molecular markers. Called overfitting, the problem can occur when, for example, a multivariable model designed to discover a 'pattern' that discriminates among individuals with and without cancer, is made to perfectly 'fit' a set of data. For example, the analysis of thousands of peaks from mass spectroscopy or of thousands of genes expressed in a microarray can result in a model or algorithm that appears to have perfect discrimination[31,58]. A problem occurs when, in assessing a large number of possible predictors, a pattern is found that fits perfectly, but by chance. Such a model has no discriminatory ability that can be reproduced in individuals different from those used to derive the model. Overfitting is not inherent in molecular-marker research; it can occur in any discovery-based research that uses multivariable analysis to assess associations between large numbers of possible predictors and an outcome. Overfitting can be easily checked for by assessing reproducibility in a completely independent group of individuals[31]. Overfitting remains a problem in discovery-based research mainly because of the failure to adequately carry out such checking; only about 10% of studies involving pattern-recognition analysis of RNA expression arrays report the independent assessment of reproducibility necessary to check for overfitting[59].

Generalizability

Generalizability, sometimes called 'external validity', is a separate problem and concerns to whom the results of the comparison (for example, comparison of test results in people with cancer and people without cancer) can be applied. The generalizability of a study depends on the characteristics of participants and how they are selected, regarding age, gender, comorbidity, symptom status and so on.

Typically, in the course of studying a problem over time, initial studies have limited generalizability but are perfectly satisfactory to establish a 'proof of principle' and provide the basis for later (and usually larger and more expensive) studies that assess broader generalizability. Strong internal validity is critically important for initial studies, to avoid wasted effort and cost in later ones. The 'phases' of research routinely used in drug development reflect a step-wise consideration of several issues including generalizability[60]. Phase I drug studies typically assess dose, while Phase II studies evaluate biological activity and adverse events, often in people who are very sick. Phase III studies involve individuals and outcomes that are more representative of those for whom the therapy would be used[60]. The proposal that molecular markers should be studied in 'phases' similarly addresses issues such as generalizability. This proposal suggests that initial studies should involve tissues and animals, whereas later ones involve symptomatic and then asymptomatic people[61]. The proposal does not discuss bias in detail for any of the phases; bias is simply a different topic.

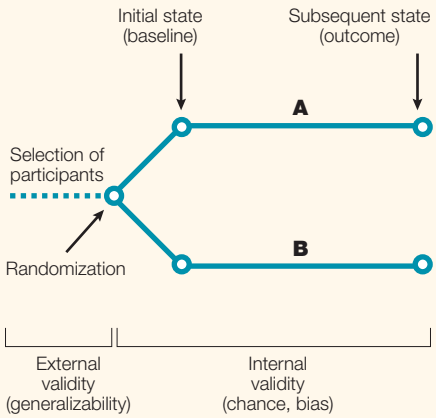

Figure 1 | **Research structure of an experiment.** The research structure of the experiment or randomized controlled trial (RCT) illustrates sources of bias and can provide lessons for non-experimental observational research such as that used to study molecular markers for cancer. The experiment or RCT is the 'gold standard' research structure for which problems caused by bias are well understood and can be clearly addressed. The purpose of an experiment is to determine whether an agent administered to group A causes a different outcome compared with group B. After participants have been selected for inclusion, allocation to the compared groups is done by randomization, to try to assure baseline equality of the compared groups. The groups are followed over time to an outcome (subsequent state), when results are compared. Later in a study, randomization might also be used to avoid a different bias; for example, randomizing the order in which specimens are analysed on a spectroscopy machine could help avoid bias from signal that could be inadvertently introduced by the machine.

and addressing bias. This Perspective article, necessarily a brief primer, advocates that laboratory investigators who study markers for diagnosis and prognosis and use methods of observational epidemiology should, like clinical investigators, arrange appropriate expertise or collaboration to address these challenges.

Principles for addressing bias in clinical research. The schematic shown in FIG. 1 shows the structure of a RCT or experiment, and illustrates sources of bias and how a RCT deals with them. Participants start at a baseline state and are followed over time to a state at which results or outcomes are compared. One group (A) receives the agent while the other (B) receives a comparative agent or placebo. Bias occurs if one group is handled differently in any way other than whether or not it receives the agent. 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. Instead, the process requires that investigators consider the larger questions of 'what biases could occur and how they might be addressed?' Useful lessons can be learned by considering how such problems would be handled in a RCT.

To illustrate those lessons and how the process works, this Perspective article describes (below) two common kinds of bias in observational research — inequality at baseline and unequal assessment of results or outcomes. This general description is followed by examples of the details that need to be considered in studies that assess molecular markers.

Bias of inequality at baseline. The bias of baseline inequality is addressed during the design of an experiment by using the powerful method of randomization to assign individuals to the compared groups. Because randomization sometimes does not work, investigators involved in the conduct of the study typically measure and report results of the randomization, commonly in the first table in a publication of a RCT (FIG. 2). The characteristics described in this first table include not only demographic features such as age and gender but also others that, if unequal, might account for different results. Deciding which characteristics need to be reported requires thoughtful consideration by investigators about the details of the problem and technology being studied. As the study shown in FIG. 2 concerns the treatment of cancer, the baseline characteristics chosen for the first table in this paper include lymph-node status, histology, oestrogen-receptor status and previous therapy, because differences at baseline might account for different results. A great strength of randomization is that it addresses not only those baseline characteristics that can be identified, measured and described in the first table, but it also should address others that cannot be identified or measured. Because randomization usually is successful, baseline inequality is seldom a problem that requires investigators' attention. For the example shown in FIG. 2, the investigators reported that the "...two groups were balanced."[39] Even if inequality occurs, it is not necessarily important; after investigators find an inequality and "...consider...the most likely direction and magnitude of...impact"[35], it might turn out that a bias is judged to be present but not important because of its direction or magnitude.

This process illustrates the importance of reporting results to check whether inequality results by chance during randomization. Thorough reporting is even more important in observational research, where biases cannot be addressed by powerful methods such as randomization.

Bias of unequal assessment of results. Randomization of groups at baseline addresses only the bias of baseline inequality; biases that occur elsewhere in a study must be addressed in different ways. For example, results could be biased if data are collected or interpreted differently in the groups being compared. If the outcome is death from prostate cancer (as opposed to death with prostate cancer), based on interpreting clinical and autopsy data, the interpretation of cause of death might be unequal or biased if investigators know an individual's treatment group. To try to avoid this bias, investigators might be kept 'blind' to, or unaware of, an individual's treatment group during data collection, analysis and interpretation. In a RCT, this process too is reported; the Consolidated Standards of Reporting Trials (CONSORT) guidelines for reporting conduct of RCTs suggest describing "Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated."[40] Not all biases are important in all studies. For example, if an outcome is assessed in a totally objective manner, such as the status of alive or dead, then biased ascertainment is less of a problem than if the outcome were cause of death, which involves the subjective interpretation of clinical and laboratory data. In observational research, making an effort to avoid this kind of bias where possible, and reporting that effort, is even more important than in a RCT, where bias might be more easily avoided.

Prospective design might help to minimize bias. Because RCTs are 'prospective' — the study is planned before occurrence of the phenomena that become the data — it is possible to apply methods such as randomization and blinding in order to avoid bias. In observational research, randomization to assure baseline equality cannot be carried out; however, other methods can sometimes be used to address bias in design and conduct. It is beyond the scope of this article to discuss details of experimental versus observational and prospective versus retrospective research. The point is that the prospective, uniform, blinded collection of data might be unfeasible or even impossible, so that investigators must routinely conduct research that is less than ideal.

Said another way, methods like randomization and blinding are not ends in themselves but rather are a means to an end: minimizing erroneous conclusions based on biased results. That end may be reached in other ways. When randomization at baseline cannot be done, inequality might be addressed by inclusion and exclusion criteria that select otherwise homogeneous groups, or by using stratification to minimize heterogeneity between compared groups. Similarly, if blinding cannot be done during the interpretation of an assay, then rigorous explicit operating procedures might help to minimize bias. In a retrospective study, where the phenomena that form the data occur before the study is planned, it might still be possible to apply randomization and blinding to some aspects of the study. For example, suppose a new type of mass-spectroscopy machine is applied to analyse a sample of previously collected stored specimens. If the spectroscopy pattern 'wanders' over time so that the machine can inadvertently introduce a signal into the data, it might be possible to avoid bias from this source by randomizing cancer and non-cancer specimens with respect to the time of analysis. Randomization at this step addresses bias in specimen analysis but does not address other biases. If the cancer specimens had been stored 10 years longer than non-cancer specimens, then a bias resulting from serum changes caused by storage will be 'hard-wired' into the data and not addressed by efforts to avoid bias in application of a new assay. In other words, randomization at such a point in a study might be useful (and the study might be called 'randomized') but would deal with only one bias. Biases do not have to be managed perfectly, but they must be considered explicitly and the process reported clearly, enabling investigators, reviewers, editors and readers to interpret the strength of a study's results.

Table 1. Base-Line Characteristics of the Patients and Tumors and Primary Treatment.*

Variable	Exemestane (N=2362)	Tamoxifen (N=2380)
Demographic characteristics		
Age — yr	64.3±8.1	64.2±8.2
White race — no. (%)	2308 (97.7)	2325 (97.7)
Nodal status — no. (%)		
Negative	1211 (51.3)	1211 (50.9)
1–3 Positive nodes	715 (30.3)	706 (29.7)
≥4 Positive nodes	321 (13.6)	330 (13.9)
Positive, but no. of nodes missing	5 (0.2)	9 (0.4)
Unknown	84 (3.6)	96 (4.0)
Missing data	26 (1.1)	28 (1.2)
Histologic type — no. (%)		
Infiltrating ductal	1814 (76.8)	1871 (78.6)
Infiltrating lobular	346 (14.6)	327 (13.7)
Other	172 (7.3)	156 (6.6)
Unknown	3 (0.1)	1 (<0.1)
Missing data	27 (1.1)	25 (1.1)
Estrogen-receptor status — no. (%)†		
Positive	1917 (81.2)	1936 (81.3)
Progesterone-receptor positive	1312 (55.6)	1307 (54.9)
Progesterone-receptor negative	351 (14.9)	384 (16.1)
Progesterone-receptor status unknown or missing	254 (10.8)	245 (10.3)
Negative	26 (1.1)	33 (1.4)
Unknown	398 (16.9)	392 (16.5)
Missing data	21 (0.9)	19 (0.8)
Progesterone-receptor status — no. (%)		
Positive	1320 (55.9)	1313 (55.2)
Negative	360 (15.2)	395 (16.6)
Unknown	659 (27.9)	653 (27.4)
Missing data	23 (1.0)	19 (0.8)
Type of surgery — no. (%)		
Mastectomy	1222 (51.7)	1235 (51.9)
Breast-conserving	1116 (47.2)	1123 (47.2)
Unknown	3 (0.1)	2 (0.1)
Missing data	21 (0.9)	20 (0.8)
Previous chemotherapy — no. (%)		
Yes	766 (32.4)	765 (32.1)
No	1575 (66.7)	1596 (67.1)
Missing data	21 (0.9)	19 (0.8)
Previous hormone-replacement therapy — no. (%)		
Yes	567 (24.0)	557 (23.4)
No	1723 (72.9)	1747 (73.4)
Unknown	51 (2.2)	54 (2.3)
Missing data	21 (0.9)	22 (0.9)
Duration of tamoxifen therapy at randomization — yr		
Median	2.4	2.4
Interquartile range	2.1–2.7	2.1–2.7
Tamoxifen dose — no. (%)		
20 mg	2243 (95.0)	2270 (95.4)
30 mg	77 (3.3)	76 (3.2)
Missing data	42 (1.8)	34 (1.4)

* Plus–minus values are means ±SD. Patients with missing data had no value reported for a given variable; for patients in the "unknown" category, data were reported as unknown.
† Data for positive and negative estrogen-receptor status include retrospectively ascertained status for some patients whose status was unknown at randomization.

Figure 2 | **Checking to see whether randomization was successful.** Investigators report the results of randomization to see whether baseline characteristics of the compared groups were made equal by randomization. Typically, these results become the first table in a report of a randomized controlled clinical trial. The table above includes demographic, clinical and other features that, if unequal, might affect the outcome. For this study, the authors wrote: "The two groups were balanced with regard to base-line characteristics." Analogous detail is seldom provided in observational studies of molecular markers for diagnosis and prognosis, even though problems due to bias in such studies are much more difficult to manage than in randomized controlled trials. Table reproduced with permission from REF. 39 © (2004) Massachusetts Medical Society.

Challenges to addressing bias. Although the word 'bias' is singular, there are many potential biases, and each might require specific consideration during study design and conduct. Some biases are straightforward to identify and address, but others are subtle. The identification and management of biases that occur in RCTs have taken decades and have been based on trial and error. In an early effort to understand bias in case–control research — a particularly problematic kind of observational research — an international conference[41] catalogued over 50 different biases[42].

In addition, while some biases are straightforward to identify and measure, others can be identified but not measured. As one observer has noted, "Even when such biases can be identified, their magnitude — and sometimes even their direction — can be nearly impossible to assess."[36] (The process of trying to 'adjust for' bias in analysis is an entirely different subject, fraught with major difficulties.) For example, in a non-randomized observational study of a therapy (with the same structure as the study outlined in FIG. 1, except there is no randomization), the assessment of baseline

Table 1 | **How bias is addressed in experimental and observational studies**

	Involving people	Involving specimens
Experimental study (for example, randomized controlled trial)		
Design	Randomize allocation to compared groups at baseline	Arrange for uniform (and, if possible, blinded) collection, handling and analysis of specimens
Conduct	Measure and report baseline characteristics of groups	Check to see whether uniform handling occurred and whether blinding was successful
Interpretation	If groups are unequal, discuss direction, magnitude and potential impact of bias	If groups are unequal, discuss direction, magnitude and potential impact of bias
Observational study		
Design	Avoid heterogeneity in selection; or stratify subjects in a way that minimizes differences between groups	Find specimen groups that have minimal differences; or, where possible (and it is usually not), arrange for uniform and blinded collection, handling and analysis of specimens
Conduct	Measure and report baseline characteristics of groups	Measure and report details of how specimens in each group were collected, handled and analyzed
Interpretation	Discuss possible biases and their direction, magnitude and potential impact	Discuss possible biases and their direction, magnitude and potential impact
Example	Subjects in one group are old and have multiple illnesses; subjects in the comparison group are young and healthy	Collection: blood specimens for the cancer group, from clinic number 1, sit for 6 hours before being separated and frozen; specimens for the non-cancer group, from clinic number 2, are immediately separated and frozen Handling: cancer specimens have been thawed and refrozen five times; the non-cancer specimens only once Analysis: cancer and non-cancer groups are analysed on different days; if the machine 'wanders' over time, 'signal' may inadvertently become introduced into the data

The table illustrates a general approach to the problem of bias and is not comprehensive. For description of the kinds of details that investigators must consider, see text.

equality ideally requires recording and categorising the reason for 'assignment' to one treatment group instead of another. Physicians typically make the decision to assign a patient to one group or another, but the reason is rarely explicit in a patient's chart. In a non-randomized study of cancer therapy, the observation that patients receiving radiation therapy have worse survival than those receiving surgery might be accounted for by the physicians' systematic assignment of very sick patients, who have little chance for cure, to less-invasive radiation therapy. In a more subtle and high-profile example, the observation that patients who received transurethral prostatectomy for benign prostatic hypertrophy had worse long-term survival than those who had open prostatectomy[43] could be accounted for by the selection of sicker patients for the less invasive procedure. In that study, the 'reason' for assignment could not be directly assessed, and the question of whether baseline characteristics indicating degree of 'sickness' had been successfully measured and adjusted for was the subject of discussion in the original report[43], an editorial[44] and subsequent analysis[45]. An entire field of research now concerns how to conduct 'comorbidity adjustment' — which accounts for the effects of concomitant but unrelated disease — and whether that adjustment can be successful, reflecting the difficulty in identifying, measuring and adjusting for inequality at baseline in observational research[46,47].

An entirely different bias, occurring later while ascertaining the results of a study, might explain why oestrogens appeared to lower the risk of coronary artery disease in observational cohort studies, whereas subsequent RCTs showed that oestrogens raise the risk[48,49]. Although several explanations could account for the difference in results of these observational studies and RCTs — including a biological explanation that the use of hormones induces early coronary artery disease in persons who, in an observational study, would not be included in the study population — one possibility is that biased ascertainment of cause of death might have occurred if investigators preferentially attributed coronary artery disease as the cause of death for people not taking oestrogens[50].

Bias in molecular-marker research

Addressing bias during the investigation of molecular markers involves the same process as the experimental method shown in FIG. 1 and discussed above, but is not as straightforward. For example, in order to select participants for studies of diagnostic tests, several different approaches can be used (and these cannot be as readily displayed as those in the figure), each of which might be associated with different problems. Individuals might be selected for study before they receive the test that is being assessed and before it is known whether they have the disease, and every participant receives the same evaluation for the test and disease. Alternatively, participants might be selected once they are known to have the disease. Different kinds of biases can occur in these different situations. However, in spite of these differences, the same principle — assuring equality — can be applied to the design, conduct and interpretation of the studies. The examples shown in TABLE 1 do not provide a comprehensive list of biases but illustrate how the same principle is used to address bias in experimental and observational studies.

The details of biology and technology affect which biases might be important. The kinds of bias that can occur in molecular-marker research depend, as for other research, on specific details of the biology and technology that is being assessed. For example, in a study designed to evaluate the ability of pattern-recognition serum proteomics to diagnose cancer, an investigator's knowledge of the biology of cancer and proteins will suggest specific possible important biases. Specifically, proteins are known to vary widely in different individuals because of variables such as age, gender, medications and the presence of various diseases; or they might be affected by methods of specimen collection or handling, such as the length of time until blood separation and freezing.

Table. **STARD Checklist for the Reporting of Studies of Diagnostic Accuracy***

Section and Topic	Item #		On page #
TITLE/ABSTRACT/KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS		Describe	
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cutoffs, and/or categories of the results of the index tests and the reference standard.	
	10	The number, training, and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).	
	22	How indeterminate results, missing responses, and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

* MeSH = Medical Subject Heading; STARD = Standards for Reporting of Diagnostic Accuracy.

Figure 3 | **Guidelines for reporting studies of diagnostic accuracy.** Guidelines for reporting the methods and results of studies of diagnosis help to provide transparency but do not specify the details of design and conduct that must be addressed in any particular study. Such details, often specific to features of the biology and technology studied, must be identified and addressed by the investigator. Table reproduced with permission from REF. 52 © (2003) American College of Physicians.

Not all kinds of bias are equally important for every cancer and every technology. Expected biases might be different for different cancers or technologies. In a study to predict cancer prognosis using RNA expression, results might be affected by the process of specimen collection, for example if cancer tissue comes from patients under general anesthesia while non-cancer tissue comes from outpatients. In contrast, in a study measuring DNA mutations, specimen handling might be less important as a source of bias because DNA is much more stable than proteins or RNA.

Status of current efforts to avoid bias. As problems of bias in observational studies of molecular markers are more difficult than experimental studies, it might be expected that investigators routinely provide (and that reviewers and editors routinely expect) detailed description and discussion of the possible biases that could seriously or fatally compromise a study's results and conclusions. However, the amount of detail typically provided in Methods, Results and Discussion sections is often meagre and perfunctory, along the lines of 'subjects consisted of 100 consecutive cancers and 100 controls' and 'all specimens were stored at minus 80 degrees.' It is rare that the level of detail that is reported approaches the level provided in the first table of reports of RCTs (FIG. 2). Considering bias in such detail might be a new experience for many laboratory investigators, because heterogeneity and bias are far less problematic when the groups compared comprise genetically-identical animals or cell lines, or when the ascertainment of outcome in a laboratory setting can easily be carried out in a uniform and blinded manner. By contrast, studies of "free-living human beings"[35] necessarily involve multiple sources of bias.

Role of guidelines for reporting research. Guidelines for research, such as those outlined by the CONSORT for RCTs[40,51] and the Standards for Reporting of Diagnostic Accuracy (STARD) initiative for studies of diagnostic tests[52,53], do not concern details of design but instead relate to the thoroughness and transparency of reporting. The Minimum Information About a Microarray Experiment (MIAME) guidelines address the reporting of technical details of how assays are done[54], and guidelines proposed for research about tumour markers for prognosis also address reporting (L. M. McShane *et al.*, manuscript in preparation).

Guidelines for reporting studies about diagnosis, such as the STARD guidelines shown in FIG. 3, are useful because they help to provide transparency, but they cannot prescribe exactly which details need to be made transparent. Instead, they suggest places for researchers to 'look' for potential bias. It is then up to the researcher to determine the details, which are often specific to the biology and technology of the study, that must be addressed during design and conduct. In the STARD guidelines, bias could be related to features reported in items 3, 4, 6, 8, 11 and 15. Item 8, for example, suggests that investigators report "…how and when measurements were taken." However, this guideline does not, nor could it be expected to, specify that the measurements relevant to a pattern-recognition serum-proteomics assay might include details of timing from blood collection to separation and freezing, the duration that specimens have been stored or the numbers of thaw–freeze cycles. 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.

Approaches that do not avoid bias

When considering approaches to address bias, it is also worthwhile to identify approaches that, while they might be good working practices for other purposes, do not address bias.

Large sample size. Using a large sample size does not directly address bias, although it can reduce statistical uncertainty by providing a smaller confidence interval around a result. Small studies done well can effectively answer important questions and demonstrate a 'proof of principle' about a molecular marker[55]. The essential feature of such a study is design that minimizes problems from chance and bias and discussion that appropriately considers possible shortcomings.

Demonstrating reproducibility. Demonstrating reproducibility does not assure against bias. In a study of serum proteomics, if a group of specimens is split into training and validation sets, a difference in the collection or handling of specimens could cause signal or bias to become hard-wired into the data in a way that would be reproduced in the validation set. Demonstrating reproducibility would exclude chance (that is, overfitting) as a cause of discrimination but would not exclude bias. Even demonstrating reproducibility in multiple centres or multiple studies might not address bias. The reduction of coronary artery-disease risk by oestrogen was reproduced in several large high-quality non-randomized observational studies before RCTs showed that oestrogens increase risk. As Dave Sackett famously said, "Bias times 12 is still bias"[36]. Non-reproducibility might indicate that bias could be a problem if there is no alternative explanation such as differences between participants or technical details, but assessing reproducibility does not itself provide a direct, efficient or reliable means to address bias.

Prospective design. Although prospective design does not itself address bias, it does provide the opportunity to apply methods that do address bias, such as the blinding of investigators and the uniform handling of participants, specimens and data. In reality however, prospective studies with the uniform and blinded handling of specimens are expensive and can be unfeasible. In a study to assess the ability of stool DNA to detect colorectal cancer and in which prospective design provided uniform, blinded handling of specimens and assays, over 5,000 participants were enrolled to find 31 invasive colon cancers, at a cost of over 10 million US$ (REF. 56). A large RCT designed to assess an intervention can sometimes be used as a basis for additional studies about diagnosis, aetiology or prognosis. The prostate, lung, colon, ovary (PLCO) clinical trial of cancer screening, which is being conducted by the National Cancer Institute in the USA, involves a quarter-century of work, over 150,000 individuals enrolled at 10 screening sites, a central field collection coordinating centre and a data-analysis and support centre[57]. Even with such an infrastructure, the addition of a high-quality biorepository for serum and tissue has involved substantial further effort and expense. The PLCO biorepository is likely to provide an extraordinarily valuable resource to assess molecular markers for diagnosis and prognosis, but the effort to create this, or any high-quality prospective blinded 'database,' requires huge resources and commitment. Such efforts will typically be feasible only in 'late-phase' clinical trials and in focused research that addresses specific questions based on promising preliminary data, similar to efforts in research about drug therapy.

Statistical analysis. Statistical analysis of collected data cannot solve fundamental problems of design; if bias is hard-wired into data by faulty design or conduct, then statistical analysis or data-mining cannot eliminate bias. Yet scientists sometimes seem infatuated by the power of such analysis. The issue, as Breslow notes, concerns "...the fundamental quality of the data, and to what extent are there biases in the data that cannot be controlled by statistical analysis[.] One of the dangers of having all these fancy mathematical techniques is people will think they have been able to control for things that are inherently not controllable."[36]

Conclusion

Of the three threats to validity in clinical research, the problems caused by chance and bias must be minimized in every study because they compromise internal validity — the fundamental comparison between the groups. Of the three threats, bias presents the greatest difficulty at every step of design, conduct and interpretation. Some biases are relatively well understood, but many are subtle. The presence of even one bias, whether clear or subtle, recognized or not, can be fatal. In current molecular-marker research, problems related to bias are widely ignored by investigators, reviewers and editors. Simple and straightforward processes — such as the consideration of equality regarding characteristics of individuals, specimen collection, handling and storage — could go a long way towards improving the situation. Without such attention, molecular-marker research is likely to generate erroneous conclusions, many of which could be avoided by using appropriate design and interpretation. Guidelines for reporting research results provide a positive first step by raising general awareness and encouraging the transparency of reporting. However, no guideline can replace an investigator's insight and reflection in considering and addressing possible sources of bias in every step of research, from design and methods to results, analysis and interpretation.

David F. Ransohoff is in the Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, CB# 7080, Bioinformatics Building, 4103 Chapel Hill, North Carolina 27599–7080, USA and at the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, 20892–7354, USA.
e-mail: ransohof@med.unc.edu

doi:10.1038/nrc1550

1. Ransohoff, D. F. Developing molecular biomarkers for cancer. *Science* **299**, 1679–1680 (2003).
2. Stears, R. L., Martinsky, T. & Schena, M. Trends in microarray analysis. *Nature Med.* **9**, 140–145 (2003).
3. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
4. Zhu, W. *et al.* Detection of cancer-specific markers amid massive mass spectral data. *Proc. Natl Acad. Sci. USA* **100**, 14666–14671 (2003).
5. Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. Y. & Chan, D. W. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin. Chem.* **48**, 1296–1304 (2002).
6. Adam, B. L. *et al.* Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609–3614 (2002).
7. Petricoin, E. F. *et al.* Serum proteomic patterns for detection of prostate cancer. *J. Natl Cancer Inst.* **94**, 1576–1578 (2002).
8. FDA Press Office. *Protein patterns may identify ovarian cancer* [online], <http://www.fda.gov/bbs/topics/NEWS/2002/NEW00797.html> (2002).
9. Marcus, A. Testing for ovarian cancer is on the way. *Wall Street Journal* (*New York*) D1, D2 (1 Oct 2002).
10. Pollack, A. New cancer test stirs hope and concern. *New York Times* (*New York*) D1, D6 (3 Feb 2004).
11. Food and Drug Administration. *Letter to correlogic systems, Inc.* [online], <http://www.fda.gov/cdrh/oivd/letters/021804-correlogic.html> (2004).
12. Food and Drug Administration. *Letter to laboratory corporation of America* [online], <http://www.fda.gov/cdrh/oivd/letters/030204-labcorp.html> (2004).
13. Food and Drug Administration. *Letter to quest diagnostics* [online], <http://www.fda.gov/cdrh/oivd/letters/030204-quest.html> (2004).
14. Society of Gynecologic Oncologists. *Society of gynecologic oncologists statement regarding OvaCheck* [online], <http://www.sgo.org/images/pdfs/policy/OvaCheck_statement.pdf> (2004).

15. Wagner, L. A test before its time? FDA stalls distribution process of proteomic test. *J. Natl Cancer Inst.* **96**, 500–501 (2004).
16. Garber, K. Debate rages over proteomic patterns. *J. Natl Cancer Inst.* **96**, 816–818 (2004).
17. Check, E. Running before we can walk? *Nature* **429**, 496–497 (2004).
18. Diamandis, E. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J. Natl Cancer Inst.* **96**, 353–356 (2004).
19. Baggerly, K. A., Morris, J. S. & Coombes, K. R. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20**, 777–785 (2004).
20. Sorace, J. M. & Zhan, M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **4**, 24 (2003).
21. Baggerly, K. A., Coombes, K. R. & Morris, J. S. Are the NCI/FDA ovarian proteomic data biased? A reply to 'producers and consumers'. *Cancer Informatics* (in the press).
22. Baggerly, K. A., Edmonson, S. R., Morris, J. S. & Coombes, K. R. High-resolution serum proteomic patterns for ovarian cancer detection. *Endo. Relat. Cancer* **11**, 583–584 (2004).
23. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
24. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596 (2003).
25. Ramaswamy, S. & Perou, C. M. DNA microarrays in breast cancer: the promise of personalised medicine. *Lancet* **361**, 1576–1577 (2003).
26. Ransohoff, D. F. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **348**, 1715–1717; author reply 1715–1717 (2003).
27. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630–631 (2004).
28. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99**, 6562–6566 (2002).
29. Baggerly, K. A. *et al.* A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3**, 1667–1672 (2003).
30. Mehta, T., Tanik, M. & Allison, D. B. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature Genet.* **36**, 943–947 (2004).
31. Ransohoff, D. F. Rules of evidence for cancer molecular-marker discovery and validation. *Nature Rev. Cancer* **4**, 309–314 (2004).
32. Ransohoff, D. F. Discovery-based research and fishing. *Gastroenterology* **125**, 290 (2003).
33. Hulley, S. B. *et al. Designing Clinical Research: An Epidemiologic Approach* (Lippincott Williams & Wilkins, Philadelphia, 2001).
34. Rothman, K. J. *Modern Epidemiology* (Little, Brown and Company, Boston/Toronto, 1986).
35. Hennekens, C. H. & Buring, J. E. *Epidemiology in Medicine* (Little, Brown and Company, Boston, 1987).
36. Taubes, G. Epidemiology faces its limits. *Science* **269**, 164–169 (1995).
37. Ransohoff, D. F. Research opportunity at the interface of molecular biology and clinical epidemiology. *Gastroenterology* **122**, 1199 (2002).
38. Ransohoff, D. F. Evaluating discovery-based research: when biologic reasoning cannot work. *Gastroenterology* **127**, 1028 (2004).
39. Coombes, R. C. *et al.* A randomized trial of exemestane after two to three years of tamoxifen therapy in postmenopausal women with primary breast cancer. *N. Engl. J. Med.* **350**, 1081–1092 (2004).
40. Altman, D. G. *et al.* The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann. Intern. Med.* **134**, 663–694 (2001).
41. The case-control study: consensus and controversy. *J. Chronic. Dis.* **32**, 1–144 (1979).
42. Sackett, D. L. Bias in analytic research. *J. Chronic. Dis.* **32**, 51–63 (1979).
43. Roos, N. P. *et al.* Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *N. Engl. J. Med.* **320**, 1120–1124 (1989).
44. Greenfield, S. The state of outcome research: are we on target? *N. Engl. J. Med.* **320**, 1142–1143 (1989).
45. Concato, J., Horwitz, R. I., Feinstein, A. R., Elmore, J. G. & Schiff, S. F. Problems of comorbidity in mortality after prostatectomy. *JAMA* **267**, 1077–1082 (1992).
46. Iezzoni, L. I. *et al.* Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? *JAMA* **267**, 2197–2203 (1992).
47. Iezzoni, L. I. The risks of risk adjustment. *JAMA* **278**, 1600–1607 (1997).
48. Hulley, S. *et al.* Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA* **280**, 605–613 (1998).
49. Manson, J. E. *et al.* Estrogen plus progestin and the risk of coronary heart disease. *N. Engl. J. Med.* **349**, 523–534 (2003).
50. Col, N. F. & Pauker, S. G. The discrepancy between observational studies and randomized trials of menopausal hormone therapy: did expectations shape experience? *Ann. Intern. Med.* **139**, 923–929 (2003).
51. Rennie, D. How to report randomized controlled trials. The CONSORT statement. *JAMA* **276**, 649 (1996).
52. Bossuyt, P. M. *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann. Intern. Med.* **138**, 40–44 (2003).
53. Bossuyt, P. M. *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. Intern. Med.* **138**, W1–12 (2003).
54. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet.* **29**, 365–371 (2001).
55. Ransohoff, D. F. Lessons from controversy: ovarian cancer screening and serum proteomics. *J. Natl Cancer Inst.* (in the press).
56. Imperiale, T. F. *et al.* Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. *N. Engl. J. Med.* **351**, 2704–2714 (2004).
57. Prorok, P. C. *et al.* Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. *Control Clin. Trials* **21**, 273S–309S (2000).
58. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst.* **95**, 14–18 (2003).
59. Ntzani, E. E. & Ioannidis, J. P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–1444 (2003).
60. Friedman, L. M., Furberg, C. D. & De Mets, D. L. *Fundamentals of Clinical Trials,* 3rd edn (Springer, New York, 1998).
61. Sullivan Pepe, M. *et al.* Phases of biomarker development for early detection of cancer. *J. Natl Cancer Inst.* **93**, 1054–1061 (2001).

Acknowledgements

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

Competing interests statement

The author declares no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:
National Cancer Institute: http://cancer.gov/
breast cancer | colorectal cancer | ovarian cancer | prostate cancer

FURTHER INFORMATION

Early Detection Research Network:
http://www3.cancer.gov/prevention/cbrg/edrn
Access to this interactive links box is free online.

ONLINE CORRESPONDENCE

Nature Reviews Cancer publishes items of correspondence online. Such contributions are published at the discretion of the Editors and can be subject to peer review. Correspondence should be no longer than 500 words with up to 15 references and should represent a scholarly attempt to comment on a specific Review or Perspective article that has been published in the journal. To view correspondence, please go to our homepage and select the link to New Correspondence, or use the URL indicated below.
The following correspondence has recently been published:

T cells with potential to target metastatic cells
Wolpert, E. Z. and Dammeyer, P.
http://www.nature.com/nrc/archive/correspondence.html

This correspondence relates to the article:

THE PROMISE OF CANCER VACCINES
Gilboa, E.
Nature Rev. Cancer **4**, 401–411 (2004)