OPINION

Rules of evidence for cancer molecular-marker discovery and validation

David F. Ransohoff

According to some claims, molecular markers are set to revolutionize the process of evaluating prognosis and diagnosis for cancer. Research about cancer markers has, however, been characterized by inflated expectations, followed by disappointment when original results can not be reproduced. Even now, disappointment might be expected, in part because rules of evidence to assess the validity of studies about diagnosis and prognosis are both underdeveloped and not routinely applied. What challenges are involved in assessing studies and how might problems be avoided so as to realize the full potential of this emerging technology?

Claims have been made in recent research studies, editorials and media reports that molecular markers are set to revolutionize the process of prognosis and diagnosis for cancer. RNA expression signatures are thought to predict the prognosis of breast cancer "better than available clinicopathological methods"[1,2], whereas serum proteomic analysis by mass spectroscopy is said to diagnose ovarian cancer with nearly 100% accuracy[3]. However, the RULES OF EVIDENCE for evaluating studies about diagnosis and prognosis[4–6] are less well developed than for studies of therapy, and specific important problems threaten the VALIDITY and reproducibility of such research. This paper focuses in depth on the problem of OVERFITTING that, in DISCOVERY-BASED RESEARCH, could lead to promising but non-reproducible, results. Other problems or challenges to validity, described briefly, should be considered in similar depth.

Molecular markers

Reasons for optimism. Molecular markers hold great promise for refining our ability to establish early diagnosis and prognosis, and to predict response to therapy. Optimism about molecular markers is based on exciting new knowledge and new technology. Knowledge about the molecular biology of cancer allows the identification of candidate target markers, such as the mutations that occur during the evolution of colon tissue from normal to adenoma to invasive cancer[7,8]. Powerful technologies including POLYMERASE CHAIN REACTION, SERIAL ANALYSIS OF GENE EXPRESSION, SINGLE-NUCLEOTIDE-POLYMORPHISM analysis and MICROARRAYS can target almost any DNA or RNA sequence. In addition, new mass-spectroscopy techniques can characterize and quantify thousands of proteins and peptides simultaneously. Technology to perform HIGH-THROUGHPUT ANALYSIS in genomics and proteomics has led to an ability to conduct discovery-based research, in which large quantities of data can be analysed without a hypothesis, to search for patterns that usefully discriminate[9,10] among groups of persons with different diagnosis, prognosis or response to therapy.

Rules of evidence to assess the validity of results, important in any research, are crucial in discovery-based research in which results or patterns might have no clear biological meaning. In a climate of legitimate excitement about the attractiveness of molecular technology and the promise of discovery-based research, a temptation to be casual about adhering to rules of evidence could result in claims that are not reproducible and lead to disappointment.

Reasons for caution. Although molecular markers will undoubtedly provide advances in diagnosis and prognosis, the degree of success claimed at present is extraordinary. Will we look back in 10 years and find that initial results were not reproducible? In an example from a generation ago, carcinoembryonic antigen (CEA) was purported to be nearly 100% sensitive and specific for colorectal cancer screening in initial research[11], whereas subsequent research had very different results. History might not necessarily repeat itself, but it indicates caution before making claims of success[12]. The non-reproducibility of the CEA results was due, in large part, to the fact that individuals who were initially studied had extensive cancer, whereas individuals who were later studied had less extensive asymptomatic cancer in which CEA might not have been increased[13,14]. The fact that test results vary with the 'spectrum' of disease might seem obvious now, but there was little understanding in that era of the concept of spectrum and of the biases that affect research about diagnostic tests. Development of the methods and rules of evidence by which diagnostic tests are judged today occurred in part because of the CEA experience[13,15].

Rules of evidence. One challenge in assessing research about molecular markers is that the rules of evidence for studies of diagnosis and prognosis are not as well developed as for studies of therapy. Therapy is commonly assessed by the experimental method, such as the randomized controlled clinical trial, that elegantly handles threats to validity arising from problems of heterogeneity, complexity and bias. Furthermore, the role of the Food and Drug Administration and other approval bodies, by overseeing the process of drug development and marketing, provides a way of arbitrating results and an incentive to refine methodology and rules. By contrast,

research about diagnosis and prognosis is often conducted using non-experimental methods of observational epidemiology or clinical epidemiology that are less well developed[16–19] and are not closely overseen by regulatory bodies. Although there is a rough hierarchy of steps by which diagnostic tests can be evaluated[17,20,21], the process of validation of these tests is not as well developed or well accepted[12] as for studies of therapy.

Rules of evidence that have developed in the field of clinical epidemiology for studies of diagnosis and prognosis[13,16,22–27] are well suited to improve molecular-marker studies by handling the problems of complexity, heterogeneity and bias that threaten validity in non-experimental research. Investigators are beginning to identify the special challenges of research that is conducted at the interface of epidemiology, genetics and genomics[28], and to discuss specific problems in research about molecular markers[17,19]. This article focuses in depth on one problem — overfitting — that is particularly important in discovery-based research and for which other papers provide little detail or are silent. Overfitting can occur when MULTIVARIABLE MODELS, such as those that are used in discovery-based research, show apparent discrimination (for example, between persons with and without cancer, or with and without good prognosis) that is actually caused by chance and is, therefore, not reproducible. This article is not a technical or statistical treatise, but is intended to describe in depth one fundamental challenge to validity and, by doing so, to illustrate the manner in which other challenges to validity — described briefly below — could be considered.

Overfitting in discovery-based research

In discovery-based research, large amounts of data can be generated and analysed for discriminatory patterns to use in prognosis or diagnosis[9]. For example, RNA expression levels of thousands of genes from a cancer specimen can be analysed for patterns that predict a patient's prognosis or response to therapy. Similarly, thousands of peaks generated by mass spectroscopy of serum can be analysed for protein or peptide patterns that discriminate among persons with and without a cancer diagnosis, allowing cancer to be detected non-invasively at early stages. The first step in assessing validity is simply to learn whether results are reproducible; if not, other issues that are related to validity are irrelevant.

The problem: overfitting causes non-reproducible results. Overfitting can occur when large numbers of potential predictors are used to discriminate among a small number of outcome events. Imagine that for 10 people with cancer and 10 without, 20,000 features with no relation to cancer are assessed, such as the number of films they watched over the past month or the number of times they chew their food. If enough possible predictors are examined, even if nonsensical and random, a pattern could be found that perfectly discriminates among those 20 people with and without cancer. Such a pattern would 'fit' or discriminate among the group of individuals it was derived from in a training set, but it would not discriminate in an independent validation set.

Biostatisticians have empirically assessed how easily overfitting can occur in RNA expression microarray research[29,30]. Simon and colleagues constructed a group of imaginary individuals with (N = 10) and without (N = 10) cancer, along with expression data for 6,000 genes. They then applied different methods of CROSS-VALIDATION, in a manner typical of actual experiments, to discover discriminatory patterns. Using one common method, 98% of the models fit perfectly in the training set[30], indicating how frequently overfitting can occur. Some '-omics' research involves a number of data points (for example, expressed genes or mass-spectroscopy peaks) that far exceeds the number of independent samples, and much discovery-based research would seem to violate the rule applied in clinical epidemiology that states that for predictive models "...to have confidence in the results, there should be ... at least 10 events [for example, cancer recurrence] for each ... variable [for example, a gene overexpressing RNA]..."[31]. Such proportions are obviously unattainable for array research, but the fact that there is a 10:1 rule for this parallel problem indicates the kind of challenge that is involved in analysing such results. Approaches that are used in analysing discovery-based research include artificial neural networks[32], genetic algorithms[33], boosted decision-tree analysis[34] and metagenes[35]. Each approach involves a kind of multivariable analysis for which overfitting might be a substantial problem.

The solution: assess reproducibility. To determine whether overfitting has occurred is easy: the pattern-recognition model that is derived in the 'training set' should be applied to an independent 'validation set' consisting of individuals who are not used in the training set. If discrimination can not be reproduced in the validation set, overfitting is the most likely explanation. The most direct approach is to randomly split the original sample of individuals (sometimes called SPLIT-SAMPLE VALIDATION) into two groups — one for training, the other for validation[36].

An important reason to perform such an assessment early on is that if overfitting or chance explains the ability of a model to discriminate, then further research (that is, additional steps in validation, as discussed below) would be unwarranted and wasteful. A logistical barrier to conducting an early assessment is the practical difficulty of obtaining enough individuals for both training and validation[36]. However, the consequence of not explicitly assessing reproducibility is that non-reproducible results could become the basis for clinical care or future research directions. Indeed, a strong case can be made that if overfitting is not explicitly ruled out and remains a possible explanation, then results should not be published in the first place. To not assess reproducibility in some direct way is arguably as serious a problem as not using a control group in an experiment.

Reproducibility and current research

In recent high-visibility discovery-based research, strong conclusions have been drawn, even though reproducibility has not been demonstrated in independent validation. According to a recent survey, independent validation is carried out in only about 10% of reports about microarray research[37].

Example: validation set not independent. In one report that assessed the ability of RNA expression (transcriptional profiling) of 25,000 genes to predict prognosis of breast cancer, the 295 individuals who were used in the validation set were not independent of the training set, because 61 individuals who were used in validation had come from the training set[38,39]. Because almost all of the published results included training-set individuals, the degree of discrimination shown might have been inflated by overfitting[40]. As noted below, however, these results have been considered to provide definitive conclusions for both clinical care and biology.

Example: validation set too small. Two other reports illustrate problems in the validation step. In one report that assessed RNA expression of 12,000 genes to predict prognosis of breast cancer, only 13 independent individuals were assessed[35], a number that is too small to be meaningful[30,41]. Another study that assessed RNA expression of 24,000 genes to predict therapeutic response to breast cancer used a training set of 24 individuals and a validation set of 6 (REF. 42). In the training step, this study used a method of cross-validation that provides what is technically known as an unbiased estimate[30], so the result could therefore be considered as "not expected by

random chance"[43]. However, even if technically not due to chance, such results might not be reproducible in an independent sample, because cross-validated error estimates have a very large variance when sample size is small (for example, fewer than 50 tumour specimens)[30]. Independent assessment with sufficient numbers is required in such studies to establish confidence intervals for prediction accuracy[30]. Indeed, explicit discussion of confidence intervals should be required for such studies, as is routinely expected for studies of therapy and aetiology.

Example: validation set independent and large. In contrast to the examples above, some RNA expression results have been demonstrated to be reproducible in an independent validation set. From a group of 240 individuals with lymphoma, 160 were used as a training set to derive a prognostic model. The remaining 80 individuals were then used as an independent validation set[44], indicating how rules of evidence from epidemiology can be applied to discovery-based molecular-marker research. Although the success of RNA expression profiling in predicting prognosis of lymphoma is encouraging, success might not similarly occur for other human cancers that lack the unique characteristics of lymphoma. As lymphoma is clonal and fairly homogeneous, quantitative RNA expression profiles might have greater biological meaning than for cancers such as breast cancer, which are more heterogeneous[45] — composed of malignant and non-malignant tissue of several types from glands, stroma and blood vessels. According to one observer, "[it] is not by chance that the first applications of microarrays to the diagnosis of cancer were made on leukaemias and lymphomas"[46].

Implications for practice and biological research. The studies of breast cancer that are discussed above have been cited in editorials and reviews to provide definitive conclusions both for clinical practice and biological research, even though overfitting has not been excluded by demonstrating reproducibility in an independent validation set.

Based in part on these studies[35,38,39], gene-expression patterns of primary tumours have been said to be "...better than available clinicopathological methods for determining the prognosis of individual patients"[1] and approval is being sought to use microarrays for clinical management of breast cancer in The Netherlands[47]. If results are not reproducible, however, such practice might be worthless or even harmful to patients, by leading to inappropriate treatment.

Similarly, the cited research[38] has been judged to settle a controversy about the biology of breast cancer metastasis by providing "…compelling evidence that the genetic program of a cancer cell at diagnosis defines its biologic behavior many years later, refuting a competing hypothesis that the genetic changes driving the development of metastatic disease are acquired in residual cells after adjuvant treatment"[48]. Future research in this direction, if based on non-reproducible results, would probably be misdirected and wasteful.

What does 'validation' mean?

If investigators, reviewers and editors seem unclear about what validation is and when it has been achieved, part of the problem is that the word validation is used in several different ways, leading to substantial confusion.

Numerous meanings of validation. Validation is a word constantly used and seldom defined[49], although in general it "consists of efforts made to confirm the accuracy, precision, or effectiveness of results"[49]. In discussing validity, one text of clinical epidemiology lists five broad topic areas with sub-topics for each[50]. With so many different meanings of the word, it should be no surprise if there is confusion concerning when validation has been achieved (BOX 1).

Split-sample validation. The term split-sample validation might be a particular source of confusion because it is used to refer both to the training step and to the validation step (FIG. 1). As noted above, it can refer to the process of splitting the original sample so that part can be used in an independent validation step. However, it can also be used with an entirely different meaning when it refers to a method in the training step in which split-sample cross-validation[31,49] is done to reduce the possibility of overfitting. This method involves repeatedly removing one or more individuals from the training set — sometimes called the 'leave one out' approach — and constructing a discriminatory model at each iteration. After the entire training set has been examined in this manner, the results are combined into a single model that constitutes the model produced from the training set.

Done properly, split-sample cross-validation can help reduce overfitting[30]. However, the method can be unsuccessful in reducing overfitting if not done properly[30]. In any case, learning whether overfitting has occurred and whether a result is reproducible can not be accomplished solely in the training step; independent validation is required[51]. As Feinstein says, "Just as a scientific hypothesis is not proved with the same data from which it was generated, the validity of the multivariable results cannot be proved with what was

Box 1 | Meanings of the word 'validation'

Examples of meanings of the word 'validation', adapted from three texts.

Feinstein[49]

"Validation is one of those words — like health, normal, probability, and disease — that is constantly used and seldom defined. We can ... simply say that, in data analysis, validation consists of efforts made to confirm the accuracy, precision, or effectiveness of the results."

Hennekens[56]

Evaluation of the presence of a valid statistical association includes the role of chance, the role of confounding and 'generalizability'.

Fletcher[50]

Questions for all research studies

- Does the research design match the clinical question the research is intended to answer?
- Are results generalizable, based on the patients, variables and outcomes that were studied?
- Could findings result from bias?
- Is the magnitude of effect clinically significant (as well as statistically significant)?
- Could the findings result from chance?

Questions for studies of diagnostic tests

- Is the test clearly described?
- Is the true presence or absence of disease established for all individuals?
- Is the spectrum of patients with and without disease adequate?
- Is assessment of test and disease status conducted in an unbiased manner?
- Is test performance summarized by sensitivity and specificity?

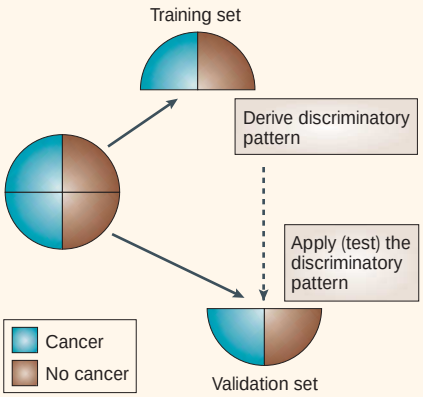

Figure 1 | **Method of dividing original sample to assess reproducibility and overfitting.** A discriminatory pattern or prediction rule is derived from the training set; split-sample cross-validation might be done in this step. The validation set is kept totally independent and is analysed to test the hypothesis (for example, discriminatory pattern or prediction rule) that is derived from the training set. After training is done and the validation set is assessed, further 'splitting' and analysis of the original sample should not be done, to avoid problems of repeated testing.

achieved when they were 'custom-tailored' to fit the 'training set'"[49]. Independent assessment helps determine whether the training step has avoided overfitting and provides a quantitative estimate of the confidence interval for the result[30].

Other topics in validation

Reproducibility is only one issue, albeit fundamental, in validation. Other issues might be similarly important and require consideration in the evaluation of molecular markers.

Bias. Bias in observational epidemiology is addressed in both the design of studies, to try to avoid bias, and in the analysis of studies, to determine whether bias has occurred. Bias in cancer-marker research includes systematic error that is related to a feature associated with, but not caused by, the cancer. For example, results of a study measuring tissue-RNA expression could be biased if cancer specimens were collected differently from non-cancer specimens in a manner that affected RNA expression or degradation. Suppose that cancer specimens were collected in the operating room under general anaesthesia, whereas non-cancer specimens were collected under local anaesthesia. An RNA expression assay might discriminate based on changes caused by general anaesthesia. Such a bias can become hardwired in a study and not detectable in independent validation if, for example, collection procedures similarly differed in an independent validation set. Although bias is routinely discussed in texts about epidemiology, the specific kinds of problems that are involved in biomarker research require detailed consideration of specific technologies, as in the example above.

In proteomics research about diagnosis of ovarian cancer, bias related to unspecified differences in sample handling or processing has been proposed as a possible explanation[52,53] for important current results[33]. However, the authors of the results state that no bias exists, because both cases and controls were handled, stored and processed under the exact same formal standard operating procedure (E. Petricoin and D. Fishman, personal communication). A separate analysis using spectroscopy data from the original study (so that bias, if it existed, could similarly affect results) reported a high degree of discrimination[3]. In an animal model (of pancreatic cancer), serum proteomics patterns have been reported to be reproducible over time[54]; this could be important because, in an animal model, problems of heterogeneity and bias might be more readily controlled or avoided. To summarize, in 2004 there have been conflicting claims in proteomics research about diagnosis of ovarian cancer, and questions of bias and reproducibility remain to be clearly resolved.

Other topics. Beyond overfitting and bias, other topics relating to validation of prognostic or diagnostic markers include confounding, 'generalizability'[18,50,55,56], clinical usefulness, benefit, harm, cost and effort. Detailed consideration of these topics, similar to this consideration of overfitting, could be important both in design and analysis of biomarker research. Although textbooks of clinical epidemiology[18,50,55] routinely discuss such topics, they do not focus on details and nuances of molecular-marker research.

Lessons to be learnt

What can editors, reviewers and investigators do about these problems? In the long run, concepts of validation and rules of evidence will need to be further developed and clarified for molecular-marker research, building on existing efforts[17,20,28,30,41,53]. Increased regulatory oversight could have a useful role. In the meantime, editors and reviewers can ask — as a first question about any claim of discrimination in discovery-based research — whether independent validation has been carried out to assess reproducibility. If overfitting has not been explicitly assessed, then results should be regarded as inconclusive, and that consideration should bear on the

Glossary

CROSS-VALIDATION
A technique used in multivariable analysis that is intended to reduce the possibility of overfitting and of non-reproducible results. The method involves sequentially leaving out parts of the original sample ('split-sample') and conducting a multivariable analysis; the process is repeated until the entire sample has been assessed. The results are combined into a final model that is the product of the training step.

DISCOVERY-BASED RESEARCH
Research in which large amounts of data are examined, without prior hypothesis, to discover markers or patterns that might discriminate among groups of individuals.

HIGH-THROUGHPUT ANALYSIS
Research in which large numbers of variables are analysed simultaneously. RNA expression analysis using microarrays simultaneously examines expression levels of tens of thousands of genes. Proteomic analysis of serum using mass spectroscopy simultaneously examines thousands of peaks related to proteins and peptides.

MICROARRAY
A solid surface on which thousands of specimens, such as synthetic oligonucleotides representing different genes, can be placed in separate locations and used to assess the status of genotype or gene expression for one individual.

MULTIVARIABLE MODELS
Models that simultaneously consider how multiple variables — such as age, gender, co-morbidity, symptoms and gene expression — relate to an outcome such as diagnosis or prognosis.

OVERFITTING
Finding a discriminatory pattern by chance, which can happen when large numbers of variables are assessed for a small number of outcomes.

POLYMERASE CHAIN REACTION
(PCR). A method to replicate or amplify small amounts of DNA into larger amounts that can be used in chemical analysis.

RULES OF EVIDENCE
Rules that are used to evaluate the strength or validity of research results by considering problems such as heterogeneity, complexity, bias and 'generalizeability'. Rules vary depending on the subject or purpose of the study: diagnosis, prognosis, therapy or aetiology.

SERIAL ANALYSIS OF GENE EXPRESSION
(SAGE). A method to estimate numbers of copies of genes.

SINGLE-NULEOTIDE POLYMORPHISM
(SNP). Variations involving a single base.

SPLIT-SAMPLE VALIDATION
Split sample validation is used, confusingly, to mean two different things. It can refer to the method in the training step by which the sample is divided during the process of cross-validation. It can also refer to the method used to divide the original sample of subjects into two groups for use in training and then in independent validation.

VALIDITY
Refers in general to efforts that are made to confirm the accuracy, precision or effectiveness of results, including reproducibility.

decision about whether to publish. Overfitting is analogous to the problem a generation ago when a result of 'no statistical difference was found' (as determined by type I or alpha error testing) was misinterpreted as meaning 'there is no difference'. The problem of course is that studies with small sample size might be 'underpowered' to detect differences that exist, and type II or beta error must be assessed to determine whether a study is sufficiently powered. An article by Chalmers and colleagues[57] that describes this problem alerted editors, investigators and readers to reflexively ask, for any study claiming no difference, whether type II error has been explicitly assessed. In discovery-based research, claims of discrimination should similarly prompt the question of whether reproducibility has been assessed in an independent validation set.

Conclusion

Discovery-based research in genomics, proteomics and other '-omics' fields is likely to become increasingly important as a source of results, claims and expectations, because high-throughput technologies and powerful analytical approaches allow for the generation and analysis of large amounts of data. Important knowledge might emerge when discovery-based research or 'fishing' is done well[58]. However, some current claims will probably not be reproducible and might lead to disappointment among investigators, the larger scientific community and funding agencies. Although hope followed by disappointment (and then further hope and development) is a necessary and expected part of scientific discovery, some current problems might in the future be judged to have been predictable and avoidable. It would be unfortunate if unsustainable claims were to cause discouragement about discovery-based research in '-omics' fields, because the potential of these fields is tremendous. A high priority of researchers, editors and funding agencies should be to avoid unsustainable claims and inflated expectations. The most successful and efficient research about molecular markers will require effective interdisciplinary communication and collaboration involving fields of molecular biology, observational epidemiology and biostatistics. Effective collaboration will be challenging because the underlying languages of the disciplines are intrinsically different, and because the rules of evidence required at this interface have not been fully developed. Strong motivation to develop this interface will be provided by the enormous potential that molecular markers have to improve the assessment of cancer prognosis and diagnosis.

David F. Ransohoff is in the Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, CB# 7080, Bioinformatics Bldg. 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/nrc1322

1. Ramaswamy, S. & Perou, C. M. DNA microarrays in breast cancer: the promise of personalised medicine. *Lancet* **361**, 1576–1577 (2003).
2. Kolata, G. Breast cancer: genes are tied to death rates. *New York Times* A1 (December 19, 2002).
3. Zhu, W. *et al.* Detection of cancer-specific markers amid massive mass spectral data. *Proc. Natl Acad. Sci. USA* **100**, 14666–14671 (2003).
4. US Preventive Services Task Force. *Guide to clinical preventive services* 2nd edn Ch. 2 (US Government Prinitng Office, 1996).
5. Woolf, S. H. Practice guidelines, a new reality in medicine. II. Methods of developing guidelines. *Arch. Intern. Med.* **152**, 946–952 (1992).
6. Tannock, I. F. & Warr, D. G. Unconventional therapies for cancer: a refuge from the rules of evidence? *CMAJ* **159**, 801–802 (1998).
7. Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N. Engl J. Med.* **319**, 525–532 (1988).
8. Ahlquist, D. A. *et al.* Colorectal cancer screening by detection of altered human DNA in stool: feasibility of a multitarget assay panel. *Gastroenterology* **119**, 1219–1227 (2000).
9. Stears, R. L., Martinsky, T. & Schena, M. Trends in microarray analysis. *Nature Med.* **9**, 140–145 (2003).
10. Ransohoff, D. F. Developing molecular biomarkers for cancer. *Science* **299**, 1679–1680 (2003).
11. Thomson, D. M., Krupey, J., Freedman, S. O. & Gold, P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc. Natl Acad. Sci. USA* **64**, 161–167 (1969).
12. Reid, M. C., Lachs, M. S. & Feinstein, A. R. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* **274**, 645–651 (1995).
13. Ransohoff, D. F. & Feinstein, A. R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N. Engl. J. Med.* **299**, 926–930 (1978).
14. Sackett, D. L. Zlinkoff honor lecture: basic research, clinical research, clinical epidemiology, and general internal medicine. *J. Gen. Intern. Med.* **2**, 40–47 (1987).
15. Feinstein, A. R. Clinical biostatistics XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests. *Clin. Pharmacol. Ther.* **17**, 104–116 (1975).
16. Ransohoff, D. F. Challenges and opportunities in evaluating diagnostic tests. *J. Clin. Epid.* **55**, 1178–1182 (2002).
17. Sullivan Pepe, M. *et al.* Phases of biomarker development for early detection of cancer. *J. Natl Cancer Inst.* **93**, 1054–1061 (2001).
18. Sackett, D. L., Haynes, R. B., Tugwell, P. & Guyatt, G. H. *Clinical Epidemiology: a Basic Science for Clinical Medicine* (Little, Brown and Company, Boston, 1991).
19. Bogardus, S. T., Concato, J. & Feinstein, A. R. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. *JAMA* **281**, 1919–1926 (1999).
20. Deyo, R. A. & Jarvik, J. J. New diagnostic tests: breakthrough approaches or expensive add-ons? *Ann. Intern. Med.* **139**, 950–951 (2003).
21. Simon, R. & Altman, D. G. Statistical aspects of prognostic factor studies in oncology. *Br. J. Cancer* **69**, 979–985 (1994).
22. Wasson, J. H., Sox, H. C., Neff, R. K. & Goldman, L. Clinical prediction rules. Applications and methodological standards. *N. Engl. J. Med.* **313**, 793–799 (1985).
23. Lachs, M. S. *et al.* Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann. Intern. Med.* **117**, 135–140 (1992).
24. Jaeschke, R., Guyatt, G. & Sackett, D. L. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* **271**, 389–391 (1994).
25. Sackett, D. L. & Haynes, R. B. The architecture of diagnostic research. *BMJ* **324**, 539–541 (2002).
26. 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).
27. Bossuyt, P. M. *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. Intern. Med.* **138**, W1–W12 (2003).
28. Potter, J. D. At the interfaces of epidemiology, genetics and genomics. *Nature Rev. Genet.* **2**, 142–147 (2001).
29. 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).
30. 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).
31. Katz, M. H. Multivariable analysis: a primer for readers of medical research. *Ann. Intern. Med.* **138**, 644–650 (2003).
32. Selaru, F. M. *et al.* Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology* **122**, 606–613 (2002).
33. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
34. Qu, Y. *et al.* Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* **48**, 1835–1843 (2002).
35. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596 (2003).
36. Harrell, F. E. Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, And Survival Analysis* (Springer–Verlag, New York, 2001).
37. Ntzani, E. E. & Ioannidis, J. P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–1444 (2003).
38. 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).
39. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
40. Ransohoff, D. F. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **348**, 1715–1717 (2003).
41. Baker, S. G., Kramer, B. S. & Srivastava, S. Markers for early detection of cancer: statistical guidelines for nested case–control studies. *BMC Med. Res. Methodol.* **2**, 4 (2002).
42. Chang, J. C. *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362**, 362–369 (2003).
43. Brenton, J. D. & Caldas, C. Predictive cancer genomics: what do we need? *Lancet* **362**, 340–341 (2003).
44. Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1947 (2002).
45. Hunter, K. W. Allelic diversity in the host genetic background may be an important determinant in tumor metastatic dissemination. *Cancer Lett.* **200**, 97–105 (2003).
46. Masters, J. R. & Lakhani, S. R. How diagnosis with microarrays can help cancer patients. *Nature* **404**, 921 (2000).
47. Pharmalicensing. *Agenda to develop microarray-based breast cancer test using Agilent Technologies' gene expression platform* [online], (cited 22 Sept. 2003) <http://atlas.pharmalicensing.com/news/headlines/1061478551_3f44e0970c520> (2003).
48. Wooster, R. & Weber, B. L. Breast and ovarian cancer. *N. Engl. J. Med.* **348**, 2339–2347 (2003).
49. Feinstein, A. R. *Multivariable Analysis: An Introduction* (Yale University Press, New Haven, 1996).
50. Fletcher, R. H., Fletcher, S. W. & Wagner, E. H. *Clinical Epidemiology: The Essentials* 3rd edn (Williams & Wilkins, Baltimore, 1996).
51. Bleeker, S. E. *et al.* External validation is necessary in prediction research: a clinical example. *J. Clin. Epidemiol.* **56**, 826–832 (2003).
52. Sorace, J. M. & Zhan, M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **4**, 24 (2003).
53. Baggerly, K. A., Morris, J. S. & Coombes, K. R. Reproducibility of SELDI–TOF protein patterns in serum: comparing data sets from different experiments. *Bioinformatics*. 29 Jan 2004. (doi:10.1093/bioinformatics/btg484)
54. Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437–450 (2003).

55. Feinstein, A. R. *Clinical Epidemiology: The Architecture of Clinical Research* (WB Saunders, Philadelphia, 1985).
56. Hennekens, C. H. & Buring, J. E. *Epidemiology in Medicine* (Little, Brown and Company, Boston, 1987).
57. Freiman, J. A., Chalmers, T. C., Smith, H. Jr & Kuebler, R. R. The importance of β, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 'negative' trials. *N. Engl. J. Med.* **299**, 690–694 (1978).
58. Ransohoff, D. F. Discovery-based research and fishing. *Gastroenterology* **125**, 290 (2003).

Acknowledgements
Thanks to many colleagues at the National Cancer Institute, The University of North Carolina at Chapel Hill and elsewhere for reviewing and commenting on earlier versions of the manuscript.

Competing interests statement
The author declares that he has no competing financial interests.

Online links

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

FURTHER INFORMATION
Research about cancer molecular markers: http://www3.cancer.gov/prevention/cbrg/edrn
Access to this interactive links box is free online.

TIMELINE

The evolution of thalidomide and its IMiD derivatives as anticancer agents

J. Blake Bartlett, Keith Dredge and Angus G. Dalgleish

Thalidomide was originally used to treat morning sickness, but was banned in the 1960s for causing serious congenital birth defects. Remarkably, thalidomide was subsequently discovered to have anti-inflammatory and anti-angiogenic properties, and was identified as an effective treatment for multiple myeloma. A series of immunomodulatory drugs — created by chemical modification of thalidomide — have been developed to overcome the original devastating side effects. Their powerful anticancer properties mean that these drugs are now emerging from thalidomide's shadow as useful anticancer agents.

Thalidomide (α-(*N*-phthalimido)glutarimide) — a synthetic glutamic-acid derivative — was manufactured and marketed by the German pharmaceutical company Chemie Grunenthal during the mid-1950s (BOX 1; TIMELINE). It is a non-barbiturate drug with sedative and anti-emetic activity and was found to be useful because of an apparent lack of toxicity in human volunteers. These properties led to it being marketed as the safest available sedative of its time. It rapidly became popular as a drug to counter the effects of morning sickness in Europe, Australia, Asia and South America, although it did not receive Food and Drug Administration (FDA) approval in the United States because of concerns about neuropathy — tingling hands and feet after long-term administration — that were associated with its use. It was withdrawn from the other markets in early 1961 after two clinicians — William McBride in Australia and Widukind Lenz in Germany — reported independently that thalidomide use was associated with birth defects[1,2]. A report associating thalidomide use with neuropathies was also reported at around this time[3]. Unfortunately, this withdrawal was too late to prevent the birth of between 8,000 and 12,000 babies with severe developmental deformities, which include the stunted-limb development that is characteristic of 'thalidomide babies'.

In 1965, following a serendipitous discovery by Israeli dermatologist Jacob Sheskin, it was reported that thalidomide was remarkably effective at improving lesions, fever and night sweats in patients with erythema nodosum leprosum (ENL) — a potentially life-threatening inflammatory complication of lepromatous leprosy[4]. After finding thalidomide in the clinic and remembering that it was a sedative, Sheskin administered it to a patient who was having trouble sleeping and — remarkably — the next morning the patient's inflammation was significantly reduced. This discovery was investigated in a study that was coordinated by the World Health Organization in thousands of men who had ENL and showed that a vast majority had complete remission within a couple of weeks of starting thalidomide treatment[5]. This was the catalyst that eventually led to the use of thalidomide as an immunomodulatory and anti-inflammatory drug[6–8]. However, thalidomide was only given FDA approval for the treatment of acute ENL in 1998, after further investigations found an immunological basis for this effect[9]. Even then, its use was limited by very strict guidelines.

It is now clear that despite its teratogenicity (BOX 1), which caused the birth defects, thalidomide is useful in treating several clinical conditions for which there are few or no alternative treatment options. An early appreciation of the immunosuppressive properties of thalidomide in several animal models led to its use in various conditions that are associated with immune activation. Initial, but mainly anecdotal, reports from the early 1980s onwards indicated that thalidomide was effective in the treatment of several autoimmune disorders. However, because the use of thalidomide was necessarily restricted, large-scale studies were not undertaken until much later. Instead, the results of various small uncontrolled studies were published and these seemed to demonstrate the efficacy of thalidomide in the treatment of patients with autoimmune disorders such as rheumatoid arthritis[10], cutaneous lesions of systemic lupus erythematosus and Behcet's disease[11,12]. The immunosuppressive properties of thalidomide also led to its use in the treatment of chronic graft-versus-host disease associated with allogeneic bone-marrow transplantation[13–15].

As thalidomide initially seemed to show promise for the treatment of these conditions, it was quickly used in further studies in small cohorts of patients with various untreatable ailments. From these investigations, it has become apparent that thalidomide is not merely an immunosuppressant, but that it has other clinically useful properties. Each new property that has been discovered has led to thalidomide being used in different spectra of disease. As a result, thalidomide is now an option for a diverse range of clinical applications and is again a profitable drug, with sales that amount to over $200 million per year in the United States and rising.

Mechanisms of thalidomide action

Thalidomide inhibits monocyte-derived TNF-α. The key finding that explained, at least in part, the potent anti-inflammatory activity of thalidomide came in 1991, when it was discovered that thalidomide inhibited the synthesis of tumour-necrosis factor-α (TNF-α) by activated monocytes[16] — the mRNA becomes less stable. TNF-α is a pro-inflammatory cytokine that is an important regulator of the inflammatory cascade and is a useful therapeutic target in inflammatory disease, particularly if activated monocytes