Improving standards of medical and public health research

G A Colditz

The need for application and dissemination of best practices

In his essay, Feinstein reiterates many of the themes he pursued over the past 20 years as he called for greater scientific rigour in the evaluation of clinical medicine. His call for the application of standard clinical epidemiological research rigour to the evolving field of molecular markers in clinical medicine emphasises his prior writing in this area,¹ and similar calls to improve standards for evaluation of diagnostic tests research.² This follows on the tradition of content analysis of published studies to improve the quality of research. Studies that evaluate the methods used in the conduct and reporting of studies have contributed over the years to improved standards for clinical research in surgery and medicine.³⁻⁴ For example, DerSimonian *et al* surveyed all 67 clinical trials published in the *New England Journal of Medicine*, the *Lancet*, and the *British Medical Journal* from July to December 1979 and in the *Journal of the American Medical Association* from July 1979 to June 1980 to determine the frequency of reporting 11 important aspects of design and analysis.⁴ Based on their review they concluded that reporting could be substantially improved. Subsequently, guidelines for reporting of randomised trials have been adopted⁵⁻⁶ and the quality of reporting has improved.⁷ This approach to documenting current practice in medical research across a wide range of topic areas has been implemented by Feinstein and his colleagues, as well as numerous other groups working to improve the quality of medical research and its application.

Is the cup half full or half empty? While Feinstein sets forth a series of methodological challenges, it is important to note that much is already being done to address several of the issues that are raised. For example, the notion that many diagnostic tests do not have a definitive or "gold standard" has been addressed by methodologists who now set forth a strategy to combine results for tests when the standard is not perfect.⁸ Importantly, the role of the evaluation of diagnostic tests is primarily an evaluation of the value of the information that is used in clinical decision making.⁹ In fact, the very emphasis on statistical

methods that Feinstein criticises¹⁰ forms the basis for many of the evolving approaches to the evaluation and application of diagnostic tests. The Radiology Diagnostic Oncology Group has been in place in the United States for over a decade with a goal to provide timely and generalisable clinical evaluations of imaging technologies.¹¹

Also of note is the call by Feinstein to reconsider the allocation of resources to avoid research that is "often unnecessary". This is consistent with the refrain by many that the value of information from testing should be evaluated before the test is ordered, and if no new action will be taken based on the possible results then the test is not indicated. The clear handbook on evidence based clinical medicine that grew out of the informative series in the *JAMA* explicitly sets forth the steps in evaluating research studies and applying the results to individual patients.¹² It also addresses such issues as the measurement of agreement between observers and the grading of treatment guidelines or recommendations. Thus, while Feinstein has criticised the "use of reductive calculation of sensitivity and specificity, and other mathematical indices that directly connect manifestation, marker, or test result with a specific disease" the field has evolved to also consider the steps in the application of results from group level data (as seen in the results of clinical trials) to individuals.¹³⁻¹⁴ The implementation of these approaches remains to be effectively disseminated.¹⁵ Furthermore, there is an urgent need for research on the relative effectiveness and efficiency of different strategies to speed dissemination and implementation.

A theme of Feinstein's writing has been the limited quality of scientific studies and the menace that this can cause—either in clinical settings through bias and lack of consideration of error in evaluations, or in epidemiological studies where causal inference may be incorrectly applied.¹⁶ His critiques have not always been correct. His erroneous conclusion that alcohol is not causally related to breast cancer has been clearly rebutted by the more than 40

studies published on this relation,¹⁷ consistent results from numerous prospective studies,¹⁸ and a series of mechanistic studies.¹⁹

Despite these concerns, the need for rigour in the evaluation of medical interventions is imperative. The priority for this is exemplified by reports from the US Institute of Medicine, National Academy of Science,²⁰⁻²¹ and more recently the establishment in the United Kingdom of the National Coordinating Centre for Health Technology Assessment, which coordinates the health technology assessment programme for the department of health. In summary, much exciting work is ongoing to address the concerns identified by Feinstein. The field requires continuing methods development, application, and dissemination of best practices. The challenge is before us and we must endeavour to move all aspects of this research agenda forward.

J Epidemiol Community Health
2002;56:333-334

Author's affiliations

G A Colditz, Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA, USA

Correspondence to: Professor G A Colditz, Harvard Medical School, Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115-5889, USA; graham.colditz@channing.harvard.edu

REFERENCES

- 1 Bogardus ST, Cancato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodologic standards. *JAMA* 1999;281:1919-26.
- 2 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. *JAMA* 1995;274:645-51.
- 3 Bunker J, Barnes B, Mosteller F. *Costs risks and benefits of surgery*. Oxford: Oxford University Press, 1977.
- 4 DerSimonian R, Charette LJ, McPeck B, *et al*. Reporting on methods in clinical trials. *N Engl J Med* 1982;306:1332-7.
- 5 Begg C, Cho M, Eastwood S, *et al*. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:637-9.
- 6 Rennie D. How to report randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:649.
- 7 Moher D, Jones A, Lepage L, CONSORT Group [Consolidated Standards for Reporting of Trials]. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001;285:1992-5.
- 8 Walter SD, Irwig L, Glasziou P. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;52:943-51.
- 9 Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. *Acad Radiol* 2000;7:681-3.
- 10 Feinstein AR. Clinical judgment revisited: the distraction of quantitative models. *Ann Intern Med* 1994;120:799-805.
- 11 Gatsonis C, McNeil BJ. Collaborative evaluations of diagnostic tests: experience of the Radiology Diagnostic Oncology Group. *Radiology* 1990;175:571-5.
- 12 Guyatt G, Rennie D. *Users' guides to clinical medical literature. A manual for*

- evidence-based clinical practice. Chicago: AMA Press, 2002.
- 13 Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ* 1995;**311**:1356–9.
 - 14 Dans A, McAlister F, Dans L, *et al.* Therapy and applying the results: applying results to individual patients. In: Guyatt G, Rennie D, eds. *Users' guide to the medical literature. A manual for evidence-based medicine.* Chicago: AMA Press, 2002:369–84.
 - 15 Bero LA, Grilli R, Grimshaw JM, *et al.* Getting research findings into practice: closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. *BMJ* 1998;**317**:465–8.
 - 16 Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 1988;**242**:1257–63.
 - 17 Longnecker MP. Alcoholic beverage consumption in relation to risk of breast cancer: meta-analysis and review. *Cancer Causes Control* 1994;**5**:73–82.
 - 18 Smith-Warner SA, Spiegelman D, Yaun S-S, *et al.* Alcohol and breast cancer in women: a pooled analysis of cohort studies. *JAMA* 1998;**279**:535–40.
 - 19 Dorgan JF, Baer DJ, Albert PS, *et al.* Serum hormones and the alcohol-breast cancer association in postmenopausal women. *J Natl Cancer Inst* 2001;**93**:710–15.
 - 20 Committee for evaluating medical technologies in clinical use. *Assessing medical technologies.* Washington, DC: National Academy Press, 1985.
 - 21 Sox H, Stern S, Owens D, *et al.* *Assessment of diagnostic technology in health care: rationale, methods, problems, and directions.* Washington, DC: National Academy Press, 1989.

Diagnostic research

Future challenges for diagnostic research: striking a balance between simplicity and complexity

B C K Choi

The bridge from complex models to the clinicians' practice

In this issue of the journal, Feinstein has provided great insight and correctly pointed out a number of problem areas in current research on diagnostic tests¹: dichotomising a disease state into yes and no, and a test result into positive and negative, do not represent the real clinical situation. Sensitivity, specificity, and likelihood ratios are calculated from patients with known disease status and therefore do not help practising physicians to make a diagnosis. Many research studies on diagnostic tests assume a disease prevalence (prior probability) of 0.5, which is unrealistic. The premise of diagnostic research, that sensitivity and specificity of a diagnostic test is constant, is now known to be wrong. Practising physicians often want to avoid the correct and academically recommended methods because of complex computations. Many diagnostic tests can produce additional information, but it is usually disregarded in the assessment of the tests. Tests in combination should be evaluated differently than a single test. Tests used to “rule out” or to “rule in” a disease should be evaluated differently. Tests used to identify the stage of a disease, or to offer reassurance rather than diagnosis, cannot be evaluated with conventional indices of accuracy. Tests based on subjective decisions are difficult to assess. Many research studies fail to comply with appropriate methodological standards and are often affected by “work up” bias. Current research seldom questions the accuracy

of the “gold standard”. Human errors, such as intraobserver and interobserver variability and concordance, are seldom evaluated. Finally, Feinstein pointed out that the persistent focus on accuracy of diagnosis, while justified 60 years ago, is no longer appropriate, and that the current dominance of mathematical models and statistical approaches can preclude involvement of clinicians in diagnostic research.

In the past several decades, in diagnostic research that involves both clinicians and mathematicians, we have seen both a quest for simplicity and a quest for complexity.

Efforts to achieve simplicity started when Yerushalmy proposed in 1947 the indicators “sensitivity” and “specificity” for a dichotomous test, which are supposed to be simple constants for patients with and without a disease, respectively.² Youden in 1950 went a step further in simplifying and created a single index by combining together sensitivity and specificity.³ The Youden index, which is “sensitivity plus specificity minus 1”, has been shown to be the excess test performance rate.⁴ (It turns out that the likelihood ratios are relative test performance rates.⁴) One successful example of simplicity is the development of simple graphical techniques (“nomograms”) to replace the complex Bayes's theorem for calculating posterior probabilities from prior probabilities.¹ Clinicians can visualise and use and are, therefore, in favour of simple approaches. A number of years ago, a group of knowledgeable clinicians in

Toronto questioned why the tests ordered for the diagnosis of leukaemia that came back from different laboratories seemed to frequently contradict each other. A simple interrater agreement study was conducted and the κ statistics confirmed the clinicians' suspicion. Results from four diagnostic methods (routine morphology, electron microscopy, cell surface marker, and cancer cytogenetics) correlated poorly for cell type identification in leukaemia.⁵

While simple concepts and calculations are welcome by clinicians, who are the primary users of diagnostic tests, mathematicians continue to find problems in the simple models and develop complex models to correct such problems. In this regard, we see an evolution from the use of “sensitivity” and “specificity” for a dichotomous diagnostic test with known disease status,² to “predictive values” (posterior probabilities) for a dichotomous test with unknown disease status,⁴ “likelihood ratio” for an ordinal test with known disease status,¹ and the use of the slope of the receiver operating characteristic (ROC) curve to estimate the likelihood ratio for a continuous test.⁶ It has even been pointed out that there are at least three likelihood ratios and three corresponding slopes on an ROC curve.⁶ (To put things into perspective, for a dichotomous test, positive likelihood ratio is simply the ratio of “sensitivity” to “1 minus specificity”, and negative likelihood ratio is the ratio of “1 minus sensitivity” to “specificity”.⁶ Furthermore, the positive likelihood ratio is the ratio of the “posterior odds” of disease to the “prior odds” of disease, while the negative likelihood ratio is the ratio of the “posterior odds” of no disease to the “prior odds” of no disease.⁴)

The quest for complexity is sometimes needed. For example, a mass screening for rectal neoplasm was found to be affected by work up bias.⁷ When a mathematical procedure was developed to correct for the bias,⁸ results were found to be completely reversed—the sensitivity of faecal occult blood test for rectal neoplasm dropped from 91.6% (without

the correct procedure) to 27.8% (with the correct procedure).⁷ The new procedure resulted in a heated debate over which correction procedure is more correct for work up bias.⁹⁻¹⁰ This has subsequently led to the development of three sets of complex mathematical models, using causal modelling, to estimate sensitivity and specificity of three types of tests.¹¹

In the search for ultimate perfection, however, mathematicians often get carried away and create complex models that become alienated from the clinicians. Feinstein's call for new unconventional research using qualitative methods and paradigm shift is very timely and welcome. The question is how to strike a balance between simplicity and complexity. The solution seems to once again lie jointly in the hands of clinicians and mathematicians.

One way to provide a balance is to create complex models with simple model-user interface. In the olden days, if one wants to create or use a machine, such as a horse drawn cart or a sail boat, one must understand the six simple machines (lever, pulley, wheel and axle, inclined plane, wedge, and screw). With modern technology, however, such thorough knowledge is no longer possible or necessary. Modern automobiles and ocean liners are created by teams of experts, who themselves can understand and contribute to only a small part

of the complex problems. Most users do not have a clue as to what is going on inside the machines, because using such modern machines has become simpler, such as with the introduction of cruise controls and auto-pilots. In terms of diagnostic research, it could mean development of automated computer software, with specified assumptions and restrictions for users, which requires only simple inputs, to do complex calculations, and to generate simple outputs that users can easily interpret and apply.

To quote from Major Greenwood, "The scientific purist, who will wait for medical statistics until they are nosologically exact, is no wiser than Horace's rustic waiting for the river to flow away" (page IX).¹² Diagnostic tests are imperfect but improving, as is research on diagnostic tests.

J Epidemiol Community Health
2002;**56**:334-335

Author's affiliations

B C K Choi, Population and Public Health Branch, Health Canada, AL#1918C3, Tunney's Pasture, Ottawa, Ontario K1A 0K9, Canada; Department of Public Health Sciences, University of Toronto; Department of Epidemiology and Community Medicine, University of Ottawa, Canada

Correspondence to: Dr B C K Choi
(Bernard_Choi@hc-sc.gc.ca)

REFERENCES

- 1 **Feinstein AR**. Misguided efforts and future challenges for research on "diagnostic tests". *J Epidemiol Community Health* 2002;**56**:330-2.
- 2 **Yerushalmy J**. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep* 1947;**62**:1432-49.
- 3 **Youden WJ**. Index for rating diagnostic tests. *Cancer* 1950;**3**:32-5.
- 4 **Choi BCK**. Index for rating predictive accuracy of screening tests. *Meth Inform Med* 1982;**21**:149-53.
- 5 **Choi BCK**, de Harven E, Bailey DJ, *et al*. Cell type identification in leukemia: the level of agreement among four independent diagnostic methods. *Cancer Detect Prev* 1994;**18**:383-91.
- 6 **Choi BCK**. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *Am J Epidemiol* 1998;**148**:1127-32.
- 7 **Zheng G**, Choi BCK, Yu X, *et al*. Mass screening for rectal neoplasm in Jiashan County, China. *J Clin Epidemiol* 1991;**44**:1379-85.
- 8 **Choi BCK**. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992;**45**:581-6.
- 9 **Diamond GA**. "Work-up bias". [Letter]. *J Clin Epidemiol* 1993;**46**:207.
- 10 **Choi BCK**. "Work-up bias": response. [Letter]. *J Clin Epidemiol* 1993;**46**:208-9.
- 11 **Choi BCK**. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. *Epidemiology* 1997;**8**:80-6.
- 12 **WHO**. *International Classification of Diseases: Manual of the international statistical classification of diseases, injuries, and causes of death*. Ninth revision. Vol 1. Geneva: World Health Organisation, 1977.

Diagnostic research

Future challenges for research on diagnostic tests: genetic tests and disease prevention

S S Coughlin

Evidence based assessments for new diagnostic strategies

In his paper about possible ways to improve research on diagnostic testing,¹ Dr Feinstein observed that "the methodological problems are particularly noteworthy in the new era of molecular biology and genetic testing." Methodological issues surrounding research on genetic testing have been addressed in several recent articles and reports from expert advisory panels.²⁻³ Although some of these reports have focused on technological and quality assurance challenges

in large scale genetic testing and screening, many of their recommendations and conclusions are applicable to genetic testing conducted in a clinical setting or as part of research.

The National Institutes of Health-Department of Energy Working Group on Ethical, Legal, and Social Implications of Human Genome Research, Task Force on Genetic Testing, noted that the clinical use of a genetic test must be based on evidence that the gene being examined is

associated with the disease in question, that the test results will be useful to the people being tested, and that the test itself has analytical and clinical validity.⁴ Whereas clinical validity refers to the accuracy with which a test predicts the presence or absence of a clinical condition or predisposition, analytical validity is an indicator of how well a test performs in the laboratory.⁵ For DNA based tests, analytical validity requires establishing the probability that a test will be positive when a particular sequence (analyte) is present (analytical sensitivity) and the probability that the test will be negative when the sequence is absent (analytical specificity).⁴ Analytical validation of a new genetic test includes comparing it with the most definitive or "gold standard" method, performing replicate determinations to ensure that a single observation is not spurious, and "blind" testing of coded positive samples (from patients with the disease in whom the genetic alteration is known to be present) and negative samples (from controls).⁴ In most instances, the gold standard entails gene sequencing to detect mutations.

As noted by the Task Force, clinical validation involves establishing several measures of clinical performance including the probability that the test will be positive in people with the disease (clinical sensitivity), the probability that the test will be negative in people without the disease (clinical specificity), and the probability that people with positive test results will get the disease (positive predictive value) and that people with negative results will not get the disease (negative predictive value), topics also considered by Dr Feinstein. For many predictive genetic tests, knowledge of the test's clinical validity may be incomplete for years after the test is developed, which requires that the potential harms of the test be considered more carefully.³

The heterogeneity of genetic diseases and their penetrance (the probability that disease will appear when a disease related genotype is present) affect clinical validity. The same genetic disease might result from the presence of any of several different variants (alleles) of the same gene (allelic diversity) or of different genes (locus heterogeneity).⁴ With current technology, all disease related alleles cannot always be identified, and this failure to detect all disease related mutations reduces a test's clinical sensitivity.⁴ When the penetrance is incomplete, the positive predictive value is reduced.⁴

Dr Feinstein noted the challenges of evaluating diagnostic marker tests when the disease cannot always be determined to be present or absent, and test results are not always positive or negative. Even with DNA based testing, test results can be indeterminate and genetic variants are sometimes detected that are of unknown clinical significance. For example, studies of BRCA1 and BRCA2 gene mutations have sometimes found genetic variants that are of uncertain significance with respect to risk of breast and ovarian cancer.⁶ Information about the limitations of genetic testing should be conveyed to research participants as part of informed consent and genetic counselling.

Dr Feinstein noted that some tests "are used not to diagnose a particular condition but to identify the patient's location in the spectrum of phenomena associated with the disease." Genetic tests may have similar applications in the future. Persons identified on the basis of phenotype or a pronounced family history of disease may undergo genetic testing in research protocols to determine whether they carry certain mutations. For example, persons with

hereditary non-polyposis colorectal cancer (HNPCC) may undergo genetic testing to determine whether they carry hMLH1 or hMSH2 gene mutations.⁷ Genetic studies have also examined whether disease risks vary by the location of genetic mutations. For example, risk of breast and ovarian cancer has been reported to vary according to the specific location of BRCA2 gene mutations.⁸

In addition to studies of test validity, scientifically rigorous evaluations of the effectiveness of genetic testing and screening are needed.^{9,10} Evidence that screening for a genetic trait or mutation and providing early intervention or treatment results in improved prognosis and favourable health outcomes should ideally come from randomised trials. Genetic testing and screening sometimes allow for the identification of a genetic condition or predisposition before the onset of clinically recognised, irreversible disease. Screening for phenylketonuria (PKU) in newborns, for example, can detect the condition for which a preventive intervention is available.

Genetic screening for adult onset disorders is not recommended at present, outside of high risk families and research protocols. However, studies are examining the genetic basis of common, adult onset disorders that cause substantial morbidity and mortality. Examples include mutations to the hMSH2 and hMLH1 genes, which are associated with increased susceptibility to colorectal cancer and associations between factor V Leiden and thromboembolic disorders.⁷

In the future, it may be feasible to prevent diseases in adults by identifying and modifying environmental risk factors among genetically susceptible persons.^{10,11} Genetic testing might allow for the identification of persons truly at increased risk for an illness and for targeted medical interventions. For example, genetic testing may allow for the identification of subgroups of patients who are more likely to benefit from preventive strategies such as the use of cholesterol lowering drugs. Pharmacogenetics may allow for tailored drug therapy and disease prevention based upon genetic variation in effectiveness and side effects.^{12,13}

Rigorous research methods for assessing the validity and effectiveness of genetic tests are needed to obtain evidence based assessments of their clinical and public health utility. Both conventional methods and future refinements in research methods may have a role.

J Epidemiol Community Health
2002;**56**:335–6

Author's affiliations

S S Coughlin, Epidemiology and Health Services Research Branch, Division of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA

Correspondence to: Dr S S Coughlin, Epidemiology and Health Services Research Branch, Division of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, 4770 Buford Hwy, NE (K-55), Atlanta, GA 30341, USA; sic9@cdc.gov

REFERENCES

- 1 **Feinstein AR**. Misguided efforts and future challenges for research on "diagnostic tests". *J Epidemiol Community Health* 2002;**56**:330–2.
- 2 **Khoury MJ**, Wagener DK. Epidemiologic evaluation of the use of genetics to improve the predictive value of disease risk factors. *Am J Hum Genet* 1995;**56**:835–44.
- 3 **Watson MS**. Medical and public health strategies for ensuring the quality of genetic testing. In: Khoury MJ, Burke W, Thomson EJ. *Genetics and public health in the 21st century. Using genetic information to improve health and prevent disease*. New York: Oxford University Press, 2000:223–41.
- 4 **NIH-DOE Working Group on Ethical, Legal, and Social Implications of Human Genome Research**, Task Force on Genetic Testing. *Promoting safe and effective genetic testing in the United States*. (<http://www2.med.jhu.edu/ftgtelsi/promoting>)
- 5 **Secretary's Advisory Committee on Genetic Testing**. *Enhancing the oversight of genetic tests: recommendations of the SAGT*. Bethesda: National Institutes of Health. (<http://www4.od.nih.gov/oba/sacgt.htm>)
- 6 **Coughlin SS**, Khoury MJ, Steinberg KK. BRCA1 and BRCA2 gene mutations and risk of breast cancer. Public health perspectives. *Am J Prev Med* 1999;**16**:91–8.
- 7 **Coughlin SS**, Miller DS. Public health perspectives on testing for colorectal cancer susceptibility genes. *Am J Prev Med* 1999;**16**:99–104.
- 8 **Thompson D**, Eason D. Variation in cancer risks, by mutation position, in BRCA2 mutation carriers. *Am J Hum Genet* 2001;**68**:410–19.
- 9 **Burke W**, Coughlin SS, Lee NC, et al. Screening for adult onset conditions. *Genetic Testing* 2001;**5**:201–11.
- 10 **Coughlin SS**. The intersection of genetics, public health, and preventive medicine. *Am J Prev Med* 1999;**16**:89–90.
- 11 **Khoury MJ**. From genes to public health: the applications of genetic technology in disease prevention. *Am J Public Health* 1996;**86**:1717–22.
- 12 **Khoury MJ**, Morris J. *Pharmacogenomics and public health: the promise of targeted disease prevention*. (<http://www.cdd.gov/genetics/info/factsheets/pharmacofacts.htm>)
- 13 **Omenn GS**. Prospects for pharmacogenetics and ecogenetics in the new millennium. *Drug Metab Dispos* 2001;**29**(4 pt 2):611–14.

Diagnostic studies as multivariable, prediction research

K G M Moons, D E Grobbee

Patient outcomes in diagnostic research

Dr Feinstein provides a topical overview of the history and current status of diagnostic research.¹ It addresses the various forms of bias that may occur in research aiming to evaluate the accuracy of diagnostic tests. Diagnostic accuracy is defined by the extent to which a test correctly indicates the ("true") presence or absence of the disease at issue as determined by a particular reference. We largely agree with this overview. However, we would like to discuss in more detail a certain issue raised by Dr Feinstein. Notably, the use of quantitative statistical models in diagnostic research and related to this the practitioners' judgement, and the role of test results in offering prognostic information.

MATHEMATICAL MODELS TO ESTIMATE THE TRUE (ADDED) DIAGNOSTIC ACCURACY OF A TEST

Dr Feinstein argues that diagnostic research is currently unfortunately dictated by mathematical, quantitative models, often ignoring the practitioners' judgement. We, however, believe that these quantitative models are *necessary* in order to estimate the true, independent (or added) value of a test, just as they are necessary in aetiological studies to estimate the independent association of a particular exposure and the occurrence of a particular outcome.

In medical practice, the diagnostic investigation starts with a patient presenting with a particular symptom or sign indicative for the presence of a particular disease, the so called target disease.² The diagnostic investigation is a consecutive (hierarchical) process always starting from patient history and physical examination, followed by more invasive, time consuming and costly tests such as imaging. It amounts to the estimation of the probability of presence of the target disease given all test results, further referred to as the diagnostic probability.³ In practice no diagnosis is set by a single test result and each test result is judged together with other (previous) test results. Given this consecutive investigation, the fact that different tests provide to varying extents the same

information and that each test may be more or less burdening for the patient, time consuming and costly, physicians need to know which tests additionally or independently contribute to the diagnostic probability estimation.¹⁻¹¹ For example, to determine whether an exercise stress test in patients suspected of coronary artery disease has diagnostic value, one needs to quantify whether the test changes the diagnostic probability based on previous test results (for example, patient history and physical examination) that are commonly available anyway before such test is applied.

In our view, to quantify whether a particular test result additionally increases or decreases the diagnostic probability *requires* multivariable (prediction) modelling of the occurrence (prevalence) of the disease at issue as function of the different diagnostic test results. It requires a comparison of the probability (prevalence) of disease presence before and after the test, using, for example, stepwise logistic regression modelling according to the chronology in practice. The result of such analyses is the definition of one or more diagnostic prediction models including the relevant (contributing) tests. If needed, these models can be simplified to obtain easy applicable diagnostic decision rules for use in practice.¹²⁻¹⁵

Knowledge on the independent (added) diagnostic value can, as Dr Feinstein also suggests, not be inferred from the still widely used singular test parameters—that is, a test's sensitivity, specificity, and likelihood ratio.¹¹ Similarly to the variability of these parameters across patient characteristics, the independent value of a test can also not be validly quantified by the use of Bayes's theorem.³ The need for multivariable models in diagnostic research is not different from other types of medical research such as aetiological and prognostic research. Not the singular association (odds ratio) between a particular exposure and the occurrence of the outcome is informative, but their association independent of other (co) factors.

PRACTITIONERS' JUDGEMENT TO ASSESS DIAGNOSTIC ACCURACY

Dr Feinstein argues that physicians are generally reluctant to use quantitative

prediction models in their diagnostic practice but rather infer a test's accuracy from their own experience. In our view, this should be no reason to withdraw from the use of quantitative prediction models in diagnostic research.

For the evaluation of efficacy of preventive and therapeutic devices we have left the era of "experience or eminence based medicine" in favour of "evidence based medicine". This should also apply to diagnostic technologies.

Resistance to use prediction models in diagnostic practice is most likely because they are still seldomly used and reported in diagnostic studies, in contrast with aetiological and prognostic, including therapeutic, research. For the latter types of research a proper methodological framework encouraging the use of these methods can be found in all epidemiological textbooks. This is not the case for the few textbooks addressing methods of diagnostic research. Most diagnostic research still concentrates on the estimation of singular test parameters in the analyses. To enable diagnostic research to better serve practice the prevailing framework of methods for both design and analyses must be improved.

We agree, however, that the physician's judgement or experience is often an important "test" in the diagnostic investigation, which should not be ignored in research. However, we believe it is preferable to express this experience in more or less objective test results, or at least to quantify whether this "experience test" provides added information to the more objective test results (as described above).

THE VALUE OF TEST RESULTS IN TERMS OF PATIENT OUTCOME

To set a diagnosis is fundamental in medical care as it offers an indication of the patient's prognosis and directs therapeutic management. As Dr Feinstein suggests, a diagnostic technology may not contribute to the assessment of the final diagnosis, but rather provide information that could be relevant to the patient's prognosis or to therapeutic decisions.

We agree that diagnostic technologies should not only be evaluated on their diagnostic accuracy, for example, their ability to determine the presence or absence of the disease but sometimes also on their ability to change patient outcome.¹⁶⁻¹⁹ These assessments, however, call for different types of outcomes and thus different study designs. Evaluation of a test on its (added) diagnostic accuracy requires a multivariable, cross sectional prediction study: of each subject the result of the test under study is measured as well as the simultaneous presence or absence of the disease as determined by a reference. Evaluating a test on patient outcome comprises the

evaluation of diagnostic tests plus all possible administered therapies combined. This enters the realm of aetiological and therapeutic research and requires follow up studies or trials rather than cross sectional, prediction studies.^{19, 20} We believe that evaluation of diagnostic tests on patient outcome is not always necessary. In general, we think that follow up studies are not necessary and the (beneficial) effect of a diagnostic test for patient outcome may be considered as established if (1) diagnostic (cross sectional) studies have shown the test's ability to detect a particular disease and (2) therapeutic studies provided evidence on efficacy of the management of this disease.²⁰ Follow up studies to quantify the effect of a diagnostic technology on patient outcome are necessary (1) if the disease at issue lacks a specific reference to determine its presence or absence (such as heart failure) or (2) if the (new) diagnostic technology provides other therapeutic information than the reference, potentially leading to other treatment choices or (3) if the technology itself may have therapeutic properties such as salphingography to determine patency of the uterine tubes.

J Epidemiol Community Health
2002;**56**:337–338

.....

Authors' affiliations

K G M Moons, D E Grobbee, Julius Centre for Health Sciences and Primary Care, University Medical Centre, Utrecht, Netherlands

Correspondence to: Dr D E Grobbee, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, PO Box 80035, 3508 TA Utrecht, Netherlands; www.juliuscenter.nl

REFERENCES

- 1 **Feinstein AR**. Misguided efforts and future challenges for research on "diagnostic tests". *J Epidemiol Community Health* 2002;**56**:330–2.
- 2 **Sackett DL**, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown, 1985.
- 3 **Moons KGM**, van Es GA, Deckers JW, Habbema JDF, et al. Limitations of sensitivity, specificity, likelihood ratio and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* 1997;**8**:12–17.
- 4 **Feinstein AR**. *Clinical epidemiology: the architecture of clinical research*. Philadelphia: WB Saunders, 1985.
- 5 **Harrell FE**, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst* 1988;**80**:1198–202.
- 6 **Begg CB**. Advances in statistical methodology for diagnostic medicine in the 1980s. *Stat Med* 1991;**10**:1887–95.
- 7 **Knottnerus JA**. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. *Med Decis Making* 1992;**12**:93–108.
- 8 **Jaeschke R**, Guyatt GH, Sackett DL, et al. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients. *JAMA* 1994;**271**:703–7.
- 9 **Grobbee DE**, Miettinen OS. Clinical epidemiology: introduction to the discipline. *Neth J Med* 1995;**47**:2–5.
- 10 **Miettinen OS**, Henschke CI, Yankelevitz DF. Evaluation of diagnostic imaging tests: diagnostic probability estimation. *J Clin Epidemiol* 1998;**51**:1293–8.
- 11 **Moons KGM**, van Es GA, Michel BC, et al. Redundancy of single diagnostic test evaluation. *Epidemiology* 1999;**134**:276–81.
- 12 **Sox H Jr**. Clinical prediction rules in practice. *Hosp Pract* 1986;**21**:100–3.
- 13 **Wasson JH**, Sox HC, Neff RK, et al. Clinical prediction rules. Applications and methodological standards. *N Engl J Med* 1985;**313**:793–9.
- 14 **Harrell FE**, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87.
- 15 **Laupacis A**, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;**277**:488–94.
- 16 **Begg CB**. Experimental design of medical imaging trials. Issues and options. *Invest Radiol* 1989;**24**:934–6.
- 17 **Fryback D**, Thornbury J. The efficacy of diagnostic imaging. *Med Decis Making* 1991;**11**:88–94.
- 18 **Dixon AK**. Evidence-based diagnostic radiology. *Lancet* 1997;**350**:509–12.
- 19 **Bossuyt PPM**, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;**356**:1844–7.
- 20 **Moons KGM**, Ackerstaff RGA, Moll FL, et al. Association of intraoperative transcranial doppler monitoring variables with stroke from carotid endarterectomy. *Stroke* 2001;**32**:813.

Diagnostic research

The need for expanding and re-focusing of statistical approaches in diagnostic research

H Brenner, T Stürmer, O Gefeller

Fitting statistical methodology to the need of diagnostic research

In his contribution "Misguided efforts and future challenges for research on diagnostic tests"¹ Dr Feinstein has identified major gaps and shortcomings in previous and current diagnostic research. While we fully agree with most of his criticisms, we would like to take issue with him over the role of statistics and mathematical formalisation in diagnostic research. In particular, we would like to emphasise the need and potential of expanding and re-focusing rather than abandoning statistical approaches in diagnostic research.

The traditional concepts of sensitivity and specificity as well as of the "posterior probabilities" have certainly been useful as a methodological framework for structuring efforts to quantify accuracy of diagnostic markers in well defined, very special settings in the past, and they may continue to be useful as such in the future. The major limitation of these concepts does not so much lie in their intrinsic properties, but in the uncritical adoption of these concepts to a wide range of different, usually more complex settings. This misapplication along with

some misconceptions outlined below have often been severely misleading indeed.

MISCONCEPTIONS AND LIMITATIONS IN DIAGNOSTIC MARKER EVALUATION

An important example is the dogma still found in most textbooks of clinical epidemiology and biostatistics, that the sensitivity and specificity as well as the likelihood ratios are constant benchmarks of test performance, which, in contrast to the posterior probabilities, are independent of disease prevalence in the population studied. As pointed out by Dr Feinstein, this dogma has repeatedly been challenged in various settings by empirical counter-evidence. Furthermore, it has been demonstrated by more general methodological work that in situations commonly encountered in practice, in which diagnostic tests are based on dichotomisation of inherently continuous traits rather than on inherently dichotomous traits, major departures from this dogma are expected to be the rule rather than the exception.² In particular, it has been shown that for tests based on dichotomisation of inherently continuous traits, variation with

disease prevalence is typically expected to be strong for sensitivity and specificity, and even more so for the likelihood ratios. Although positive and negative predictive values also strongly vary with disease prevalence, this variation is usually much less pronounced than one would expect under the (incorrect) dogma of independence of sensitivity and specificity of disease prevalence.

NEED FOR EXPANSION AND RE-FOCUSING OF STATISTICAL APPROACHES

These findings have clear implications for diagnostic marker evaluations and their interpretations. The performance of diagnostic tests must clearly be related to the population characteristics in which it is evaluated. Simply estimating overall values of sensitivity and specificity, or of measures derived from these parameters, should no longer be sufficient to characterise test performance. Furthermore, diagnostic studies should be designed in such a way that they also allow to estimate dependence of test performance on characteristics of the study population and on the study setting.

In particular, it should become common practice in diagnostic studies to evaluate the role of covariates, such as sex, age, comorbidity, and medication, in addition to disease prevalence in the study population, on the outcome of a diagnostic test. Multivariate approaches that have surprisingly much less tradition in the evaluation of diagnostic tests than in other fields of epidemiology, may help to figure out the independent contribution of different factors in this context. Multivariate approaches may also help to improve the validity of diagnostic tests by taking covariates into account. For example, some suitable function of the result of the diagnostic test itself and of the values of the covariates may often be more predictive of the true disease status than the result of the diagnostic test alone. In this context, the framework of generalised linear models offers an appropriate statistical modelling approach to select the relevant covariates and to find a suitable function for predicting the true disease status from the available information, thereby potentially enhancing the predictive ability of the diagnostic test.

INADEQUATE APPRAISAL OF OTHER DIAGNOSTIC ROLES: EXAMPLES AND IMPLICATIONS

We agree with Dr Feinstein's notion that previous predominant or even exclusive focus on diagnostic accuracy does not adequately capture the challenges for research on diagnostic tests. A particularly important example is discrimination between two or more alternative diagnoses, which often is a more relevant diagnostic task than distinguishing between absence or presence of a certain disease in most practical situations. For example, the most challenging part of the diagnosis of acute stroke is the differentiation of cerebral infarction and cerebral haemorrhage.^{3,4} Other well known examples include the challenging differentiation between Crohn's disease and ulcerative colitis^{5,6} or between benign prostatic hyperplasia and prostate cancer among patients with intermediate levels of prostate specific antigen.⁷

Weinstein and Fineberg have briefly addressed situations, in which several diseases are under diagnostic consideration. They provided a generalisation of Bayes's theorem that yields individual probabilities for each alternative diagnosis given a specific test result.⁸ More recently, specific statistical approaches to quantify the differential diagnostic value of diagnostic procedures to distinguish between two among several alternative diagnoses have been proposed,⁹ which may be more meaningful in this context than measures of diagnostic accuracy. However, these approaches are only occasionally applied in practice and followed up in diagnostic research so far, and they need to be expanded to more complex situations of diagnostic decisions.

The approaches outlined above illustrate, that overcoming the deficits outlined by Dr Feinstein do not imply the need to leave mathematical models and statistical approaches. The opposite may be true: we have to strive for a deeper understanding of the necessary statistical methodology to find better formal strategies of conceptualising what really matters in diagnostic research. A reformulation of the mathematical models and statistical approaches in a way that better reflects the actual challenges encountered in clinical practice may be a

promising avenue to pursue. Pertinent collaborative efforts of clinical investigators and methodologists who are open to depart from conventional "paradigms" seem to be most promising in this context and should deserve particular encouragement by funding agencies and editorial boards.

J Epidemiol Community Health
2002;**56**:338–9

Authors' affiliations

H Brenner, T Stürmer, Department of Epidemiology, German Centre for Research on Ageing, Heidelberg, Germany
O Gefeller, Department of Medical Informatics, Biometry and Epidemiology, University of Erlangen-Nuremberg, Germany

Correspondence to: Dr H Brenner, Department of Epidemiology, German Centre for Research on Ageing, Bergheimer Str 20, D-69115 Heidelberg, Germany;
brenner@dzfa.uni-heidelberg.de

REFERENCES

- 1 Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests". *J Epidemiol Community Health* 2002;**56**:330–2.
- 2 Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997;**16**:981–91.
- 3 Weir CJ, Murray GD, Adams FG, et al. Poor accuracy of stroke scoring systems for differential clinical diagnosis of intracranial haemorrhage and infarction. *Lancet* 1994;**344**:999–1002.
- 4 Stürmer T, Schindwein G, Kleiser B, et al. Clinical diagnosis of ischemic versus hemorrhagic stroke - applicability of existing scores in the emergency situation and proposal of a new score. *Neuroepidemiology* 2002;**21**:8–17.
- 5 Ogorek CP, Fisher RS. Differentiation between Crohn's disease and ulcerative colitis. *Med Clin North Am* 1994;**78**:1249–58.
- 6 Bezabeh T, Somorjai RL, Smith IC, et al. The use of 1H magnetic resonance spectroscopy in inflammatory bowel disease: distinguishing ulcerative colitis from Crohn's disease. *Am J Gastroenterol* 2001;**96**:442–8.
- 7 Matsuyama H, Baba Y, Yamakawa G, et al. Diagnostic value of prostate-specific antigen-related parameters in discriminating prostate cancer. *Int J Urol* 2000;**7**:409–14.
- 8 Weinstein MC, Fineberg HV. *Clinical decision analysis*. Philadelphia: WB Saunders, 1980:109–12.
- 9 Brenner H. Measures of differential diagnostic value of diagnostic procedures. *J Clin Epidemiol* 1996;**49**:1435–9.

Challenges in dia-prognostic research

J A Knottnerus

The relevance of design issues and new methods

It is sad to comment on work of a founding father of clinical epidemiology, who so profoundly enjoyed the scientific debate, without him being able to respond. But it is not only sad, it is also a great honour to have the opportunity to discuss one of the last pieces of this great architect of clinical research.¹

The points made by Alvan Feinstein are clear and I agree with most of his analysis, but some additional challenges should be emphasised.

LIMITATIONS OF "NOSOLOGICAL INDICES"

Indeed, apart from interpretative problems, "nosological indices" such as sensitivity and specificity are problematic as they will vary with the spectrum and selection of patients.²⁻⁴ However, also "diagnostic indices" (predictive accuracy) are not constants. Taking account of all relevant covariables is, if at all useful, generally impossible because many of these are not clear, unmeasurable, or too particularistic for a useful general evaluation. So, we must not overemphasise the value of diagnostic indices. And where we use them, we must be very specific as to objective, context, and setting. We should also be keen on setting specific external validation when we want to generalise observed index values for clinical purposes. Acknowledging all this, it cannot be denied that for clinical and epidemiological purposes there is often a need to use "nosological indices" to communicate general characteristics of a test. Also Feinstein did not escape from speaking about characteristics required for a "rule out" test (high sensitivity) compared with a "rule in" test (high specificity). The question what index is "the best" is too general, and therefore a non-issue. While in principle all indices (and combinatory roles, using multivariable analysis) can be calculated from data on the association between test outcome and disease, which index is actually used depends on objective and context. For example, knowing about sensitivity and specificity is important for the selection of the most appropriate screening test, while a high predictive value is important when evaluating the diagnostic value of a pre-mastectomy cancer diagnosis.

Bayes's theorem, ROC curves with varying cut off values, likelihood ratios,

and logistic regression can be unattractive for many clinicians and also these entities are not insensitive for spectrum and selection.⁵ But for various purposes they cannot be missed, for example, in data analysis, computer algorithms, and in finding optimal cut off values for screening tests. At the same time, many challenges in the context of data analysis remain to be adequately dealt with (such as: assessing additional diagnostic value of a test rather than accuracy in itself⁶; differences between multivariable analysis for diagnostic and aetiological purposes (with major emphasis on subgroup analysis for the former); meta-analysis of research data and making them tailor-made at the same time; and bridging the gap between clinical reasoning in practice and products of data analysis.

DESIGN ISSUES MORE IMPORTANT

We should not forget that mathematical indices are just ways to summarise collected research data. For the quality of research, defining the research question, and methodological challenges in study design, are far more important. The high degree of difficulty of diagnostic research studies is reflected by various overviews over the past decades, repeatedly showing important flaws in a high proportion of studies,⁷⁻⁹ and indicating only slow improvement.

Diagnostic research should be improved and refined with respect to strategy (where phase I to phase IV studies should be subsequently designed¹⁰), spectrum and selection effects, diagnostic and prognostic reference standards, and the assessment of the clinical impact of testing. Better methods to warrant and evaluate external clinical validity are required. Furthermore, proposed innovations must always be compared with achievements of experienced clinicians, before they are recommended as better. We need more understanding of the "doctor's black box" of diagnostic decision making, jumping between observations on groups and caring for individuals, including reassurance. This will facilitate more efficient diagnostic reasoning, and developing custom-made support systems. Efficiency in evaluation of the impact of diagnostic procedures can be gained if new data on specific aspects

(for example, a diagnostic test) can be inserted in the mosaic of available evidence on a clinical problem, rather than studying the whole problem again whenever one element has changed. For this purpose, flexible scenario models of current clinical knowledge are needed. This can be strongly supported by ongoing prospective systematic review and meta-analysis of diagnostic studies and developing clinical databases.

DIA-PROGNOSTIC SIGNIFICANCE

As Feinstein points out, significance for prognosis and medical decision making is the justification of diagnosis. This insight is already often implemented and evaluated, for example in the process of staging tumours or in the clinical classification of heart failure. But we must recognise that (evaluation of) prognostic significance needs much more attention, especially where test technology advances without concomitant therapeutic improvements. A well known example is detailed, three dimensional MRI observations of the brain of which clinical relevance is not (yet) clear.

Innovation of biomedical knowledge and understanding pathophysiological processes is decisive for the development of tests with better "dia-prognostic" impact. DNA diagnostics is in fact prognostics as it touches the basic functional level of our biomedical nature. DNA testing will not only be supportive of genetic counselling and reproductive applications, but also for diagnostic and prognostic purposes. Moreover, DNA testing is expected to improve the targeting and dosing of interventions. Much work in this field is being done, for example, in cardiovascular medicine, oncology, and psychiatry. However, it will take much time before these promises will have impact in daily patient care. Considerable efforts are still needed not only in the laboratory but also in clinical epidemiological research¹¹ including long term follow up to clinically validate diagnostic and prognostic predictions.

INDIVIDUALISATION AND DEALING WITH (UN)CERTAINTY

In view of the ambition to develop a more tailor-made, perhaps even individualised, "dia-prognostic" process, study population oriented validations will get increasingly under pressure. In this context, $n=1$ research, focused on individual patients, represents great methodological challenges. In addition, the moment of testing and the occurrence of a clinical problem will not necessarily be always related. The DNA profile with its predictive potentials is there from the beginning. Ethical questions on when to test, how to deal with privacy of genetic information, and the right of (not) knowing have to be addressed. Doctors

and patients, traditionally fighting to reduce uncertainty, must also learn to cope with possibly reaching the goal of getting certainty.

FUNDING AND FORMAL STANDARDS

In contrast with therapeutic research, and despite the clinical fact that accurate dia-prognosis is the key to effective management, the funding possibilities for dia-prognostic research are generally poor. Feinstein's plea for improvement is hitting the target. Such improvement would be stimulated by adopting formal standards for evaluation of dia-prognostic procedures, to control acceptance, maintenance, and substitution in the healthcare market. This also requires high quality and transparency of evaluation reports. The recent initiative taken by the Amsterdam Academic Medical Centre to reach international agreement

on Standards for Reporting Diagnostic Accuracy (STARD) deserves therefore full support from the scientific and healthcare community.

J Epidemiol Community Health
2002;**56**:340–341

Author's affiliation

J A Knottnerus, Netherlands School of Primary Care Research, University of Maastricht, PO Box 616, 6200 MD Maastricht, Netherlands

Correspondence to: Professor J A Knottnerus;
andre.knottnerus@hag.unimaas.nl

REFERENCES

- 1 Feinstein AR. *Clinical epidemiology: the architecture of clinical research*. Philadelphia: WB Saunders, 1985.
- 2 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.
- 3 Knottnerus JA. The effects of disease verification and referral on the relationship

between symptoms and diseases. *Med Decis Making* 1987;**7**:139–48.

- 4 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;**6**:411–23.
- 5 Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;**45**:1143–54.
- 6 Moons KGM, van Es GA, Deckers JW, et al. Limitations of sensitivity, specificity, likelihood ratio, and Bayes's theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* 1997;**8**:12–17.
- 7 Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. *JAMA* 1984;**252**:2318–22.
- 8 Reid ML, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic research. Getting better but still not good. *JAMA* 1995;**274**:645–51.
- 9 Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
- 10 Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, ed. *The evidence base of clinical diagnosis*. London: BMJ Books, 2002.
- 11 Bogardus ST Jr, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. *JAMA* 1999;**281**:1919–26.