REVIEW

Misguided efforts and future challenges for research on “diagnostic tests”

A R Feinstein†

This paper was commissioned by the *Journal of Epidemiology and Community Health*, together with the accompanying commentaries. †Professor Feinstein died while the paper was in press.

At the end of the second world war, the prime scientific goal in clinical medicine was to make a correct diagnosis. Its importance was demonstrated by the attention it received in medical education. Students and house staff regularly went from the hospital ward to the institutional morgue to see the diagnostic findings at necropsy. The clinicopathological conference (CPC) was a popular teaching exercise, which occurred weekly at many medical schools, with reports in each issue of prominent medical journals. In the “detective story” format of a CPC, the clinician would appraise the details of the patient’s clinical course, offer a diagnosis, and then be either happily confirmed or sadly refuted when the pathologist revealed the necropsy results.

During the past half century, however, the importance of both the necropsy and the CPC have sharply declined. The necropsy itself, which was customary for at least 50% of patients who died, is now seldom requested and seldom done. At many leading institutions today, the necropsy percentage has dropped to less than 10%; and many students and house staff can complete their medical education without having attended the procedure. Direct meetings for CPC events have become uncommon, and published results are infrequent. The *New England Journal of Medicine* is one of the few prominent publications that continues to present CPC reports, although no longer at weekly intervals.

Among the many reasons for the reduced attention to necropsy and CPC, one obvious explanation is that the definitive diagnosis that once required postmortem examination is now obtained “in vivo” with the many procedures—biopsy, endoscopy, diverse imagings, exploratory surgery—of modern technology. The pre-mortem diagnoses sometimes turn out to be wrong, but regardless of the effectiveness and accuracy with which these procedures can replace necropsy, they are often regarded as yielding definitive diagnostic evidence.

With diagnosis continuing to be the main focus, major attention has been given to the use and accuracy of “marker tests” that can be less expensive or less invasive indicators of the definitive diagnosis. Although symptoms and physical signs are sometimes appraised as markers, most of the tests are done as technological procedures,

entailing examination of such entities as blood, urine, smears, graphic tracings, and ultrasonograms. A major academic enterprise has been developed to evaluate and teach the performance of the marker tests.

The main points to be made in this essay are: (1) the current methods of evaluating marker tests are unsatisfactory; (2) suitable appraisal of the “definitive” test procedures has been generally overlooked; and (3) the most important contributions of the technological procedures today are for prognostic and therapeutic decisions (rather than for diagnosis alone), but these decisions are seldom specifically evaluated.

DIAGNOSTIC MARKER EVALUATION

As medical technology proliferated after the second world war, mechanisms were needed for its evaluation. Some of the tests, such as packed cell volume and serum sodium concentration, were accepted for their own intrinsic role in general measurement; but others, such as cardiac enzymes and pap smears, were used as diagnostic markers.

Sensitivity and specificity

A method and nomenclature for appraising diagnostic accuracy had been introduced in 1947 by Yerushalmy *et al.*¹ They calculated *sensitivity* as the proportion of diseased “cases” who were correctly identified by a positive test, and *specificity* as the corresponding proportion of non-diseased “controls” with an accurate negative test. Yerushalmy’s group had chosen about equal numbers of cases and controls for their study, so that prevalence of the disease was about 0.5 in the research setting. The results were also arranged in a “double dichotomy” format that classified each test as *positive* or *negative*, and each patients’ disease status as *present* or *absent*.

This format of assembly and evaluation soon became widely accepted despite the many problems it created in both research and clinical practice.

Dichotomous and ordinal divisions

One part of the dichotomy required that each disease under study be unequivocally cited as *present* or *absent*. Thus, the many patients with an uncertain or equivocal diagnostic status were excluded from the research. The other part of the dichotomy required that each test result be either *positive* or *negative*. This demand would also exclude *uncertain* values for pap smears or other tests that might not have a definitive result. Thus, the doubly dichotomous tables assembled for calculating sensitivity and specificity were an inevitable distortion of the real clinical world, which

Author’s affiliation
A R Feinstein, Yale University School of Medicine, New Haven, USA

Correspondence to:
Professor I Hernández Aguado, Medicina Preventiva y Salud Pública, Departamento de Salud Pública, Campus de San Juan, Universidad Miguel Hernández, San Juan, Alicante E-03550, Spain; ihernandez@umh.es

Accepted for publication
20 December 2001

contains many shades of gray, not just pure whites and blacks.

Another undesirable consequence of the doubly dichotomous boundaries was the need to form a binary demarcation for the continuous dimensional values or ordinal grades that occurred in tests such as cardiac enzymes or urinary sugar. To meet this demand, uncertain results were excluded for both markers and diagnoses; and the binary splits were chosen by a complex procedure, derived from engineering models, called *receiver operating curves*. Later on, to avoid the arbitrariness of a single binary demarcation, the marker test results were divided into ordinal zones, and a new mathematical index, called *likelihood ratio*, was applied to denote accuracy of the test in each zone.

CONVERSION TO CLINICAL DIRECTIONS

The investigative results for *sensitivity* and *specificity* or for *likelihood ratios* could be called *nosologic* indices,² because they had come from patients in whom the diagnoses were already known. What practising clinicians needed, however, were truly diagnostic indices that would be applicable to patients for whom a diagnosis was not yet established. Furthermore, the diagnostic application would occur in clinical or public health settings, where the prevalence of disease was much lower (or occasionally higher) than the 0.5 proportion usually used in the research studies.

Accordingly, an intricate mathematical procedure, using the anticipated prevalence or “prior probability” of the disease, was established to convert the *nosologic* indices for sensitivity, specificity, and likelihood ratios into pragmatic *diagnostic* indices. The mathematics used principles of Bayes’s theorem, which led to complex calculational formulas; and the subsequent diagnostic indices or “posterior probabilities”, were called *positive predictive accuracy* and *negative predictive accuracy*. (The calculations could sometimes be avoided by clinicians who carried copies of special graphs, called nomograms, in which a line connecting the “prior” probability and the *nosologic* index would cross the line that showed the “posterior” result).

Erroneous premise about constancy

Further research, however, eventually showed that the entire calculational procedure was based on a wrong premise. The iatromathematicians had assumed that the *nosologic* values of sensitivity and specificity were constant for each disease and for each non-diseased control group, regardless of the spectrum of patients who were tested. This assumption was incorrect. The *nosologic* indices are not constant: they will vary with variations in the clinical, pathological, or comorbid attributes of the patients in different parts of the spectrum for each disease and for the complementary states of non-disease.³⁻⁷

Practitioners’ judgment

Another surprising but often ignored subsequent discovery is that practising clinicians—despite all the grants, publications, academic publicity, and educational instruction—have generally avoided the Bayesian, likelihood, and other complex calculations. When about 300 randomly selected practitioners in different fields were directly interrogated about their practice patterns,⁸ about 95% said that they did not use any of the academically recommended methods. (One practitioner said he used them only when taking examinations for specialty board certification). Instead, the practitioners usually evaluated diagnostic accuracy directly, from the proportions of false positive and false negative results encountered in their own groups of patients.

INADEQUATE APPRAISAL OF OTHER DIAGNOSTIC ROLES

Beyond the cited problems, the conventional “academic” appraisals have almost always regarded the diagnostic marker

tests as “surrogates that could be accurate or inaccurate in demonstrating the presence (or absence) of the selected disease. This monolithic “surrogate” approach ignores the many other diagnostic roles for which the information can be used.

Broad scope

Certain tests—such as chest roentgenograms or an abdominal ultrasonogram—can provide a broad scope of information about different anatomic structures. The general overall value of these tests is not properly appraised if they are evaluated only for their role in diagnosing isolated diseases.

Combinatory

Certain tests, such as cardiac enzymes, do not have a wholly independent role as surrogates. Instead, the results are combined with other information, such as symptoms and electrocardiographic data, to establish diagnoses. The conventional measurements of accuracy (in a single surrogate test) are not appropriate for appraising these combinatory roles.

Specialised function

Even when ordered for purely surrogate purposes, not all tests are expected to have the same function. A “discovery” test, intended for screening asymptomatic persons, will have different requirements from one used in differential diagnosis of patients with symptoms or other manifestations suggesting a disease. In differential diagnosis, a “rule out” test, used to assure that a disease is absent, will require high sensitivity, whereas a “rule in” test, which confirms presence of the disease, must have high specificity.

Spectral markers

Certain tests are used not to diagnose a particular disease (which may already have been demonstrated) but to identify the patient’s location in the spectrum of phenomena associated with the disease. Such tests are used, for example, in “staging” patients with cancer, in determining bacterial sensitivity to antibiotics, or in measuring the pertinent blood levels for anticoagulant or antidepressive treatment. These “spectral-marker” tests cannot be suitably appraised with the customary indices of accuracy for diagnostic markers.

Reassurance

Another important, but usually unmeasured, function is the role of certain tests in offering reassurance to both patient and physician. For example, a CT or MRI scan of the head is probably used most often not to confirm a clinical diagnosis of “stroke”, but to provide reassurance that the stroke is due to a cerebrovascular thrombus, rather than to haemorrhage, brain tumour, or subdural haematoma.

INADEQUATE RESEARCH STUDIES

Regardless of the limited focus and erroneous premises of diagnostic research, and regardless of its academic popularity and publication frequency, many of the studies have not been well done.⁹⁻¹⁰ Tests that involved subjective decisions (by radiologists, pathologists, or other interpreters) have not always been checked for reproducibility. The objectivity needed to avoid “review bias” has seldom been achieved by arranging for “blind readings” when a second procedure is performed after the results of the first procedure are known.

The cases and controls in the academic studies have not always been chosen to be suitable representatives of the corresponding clinical conditions and to cover an appropriately wide spectrum of those conditions. Many studies have had a built in bias, sometimes called “work up” or “verification bias”, produced because the definitive procedure is often ordered mainly for persons with a positive marker result, but not for those with a negative result.

The methodological problems are particularly noteworthy in the new era of molecular biology and genetic testing. In an analysis of 40 publications reporting pertinent DNA research, Bogardus *et al*¹¹ found that only five complied with all of seven appropriate methodological standards, whereas 25 failed to comply with as few as two standards. Unless rigorous principles of clinical epidemiological science are joined to the majestic laboratory advances in molecular genetics, the false positive and other defective results of the new era of "DNA tests" may produce much more human misery in the 21st century than was caused in the 20th by the epidemic of false positive results in the old Wassermann test for syphilis.

APPRAISAL OF "DEFINITIVE" TESTS

The final diagnostic problem to be cited here is the absence of suitable appraisal for the procedures that become the "definitive" or "gold standard" results against which the accuracy of marker tests is compared. Because a "gold standard" can seldom itself be appraised for accuracy, the appraisal must usually be aimed at the quality of the "gold".

If the definitive results come from histopathology, quality depends on observer variability of the pathologists, and on the extent of their agreement about the diagnostic decisions. Although the images obtained from computerised, helical, and magnetic resonance scans should have been validated against pathological anatomy, these validations are seldom performed; and so the images often become the "gold standard". In other circumstances, the macroscopic observations made at endoscopy or exploratory surgery may become the "gold standard" if biopsy specimens are not available for histopathological confirmation.

In all of the foregoing situations, the definitive diagnoses come from human observers. Nevertheless, the personal variability of the observers, and their concordance in diagnostic decisions, is seldom evaluated. Perhaps the main reason for the absence of such evaluations is the distressing amount of intrapersonal and interpersonal disagreement revealed on the few occasions when pertinent evaluations were done.

Yet another type of gold standard comes from an "intellectual" set of criteria, rather than from specific anatomical evidence. Such criteria have been particularly prominent in definitive diagnoses for rheumatic fever, rheumatoid arthritis, Kawasaki's syndrome, and other ailments that lack pathognomonic manifestations. The construction of these criteria entails challenges that are not always well managed in checking for validation as well as testing observer variability. (An excellent example of this process was recently reported by Hertzman *et al*¹² in constructing diagnostic criteria for the eosinophilic myalgia syndrome).

The consequence of the infrequent or inadequate appraisals of "gold standard" procedures is an extensive but often unrecognised uncertainty—not just in determining accuracy for marker tests, but particularly in making decisions about the definitive diagnosis itself.

ADDITIONAL ROLES OF TECHNOLOGICAL DATA

Perhaps the most glaring flaw of the entire appraisal process, however, has been the persistent focus on accuracy of diagnosis. This focus was justified 60 years ago, but is no longer appropriate in the era of modern therapeutic technology, which has greatly changed today's clinical challenges.

In diagnosis, the easy availability of endoscopy, biopsy, and surgical exploration has often produced, during life, the definitive evidence that was formerly found only on postmortem examination. As noted earlier, the advances of imaging have frequently made the various scans replace anatomical pathology as the definitive diagnosis. Consequently, diagnostic marker tests have become less important today than formerly, as clinicians often directly order the "definitive" procedure, without passing through one or more marker tests.

Furthermore, some of the most important roles of technological tests today are in non-diagnostic clinical decisions. The results of the tests are often used in estimating prognosis, in choosing treatment, in appraising post-therapeutic status, and in changing treatment. Yet none of these activities is included in the procedures developed for appraising diagnostic efficacy. If the total clinical contributions of technological tests are to be suitably evaluated now and in the future, this methodological gap will have to be eliminated, with new appraisals developed for the currently unmet challenges.

An important step in the desired progress will require fundamental alterations in nomenclature and in methodology. In nomenclature, the results of the technological procedures should no longer be called "diagnostic tests", because the results are used for much more than diagnosis alone. A name such as technological tests would be more appropriate for the broad array of clinical decisions in which the data can be used.

In methodology, the new approaches will require active, sometimes primary, contributions from knowledgeable clinicians. The appropriate methods cannot be derived from purely mathematical models, and will need major input from clinicians who are familiar with clinical decisions and with the diverse contributions of the technological tests.

The improvements can occur if appropriate clinical investigators will volunteer to collaborate in or actually do the work, and if funding agencies recognise its importance. The investigators, however, may have to contend with two difficult obstacles. One of them is the current dominance of the mathematical models and statistical approaches used to appraise "diagnostic efficacy". The new research may require qualitative methods and a focus on questions that depart from the now conventional "paradigms". The other obstacle will be the general reluctance of funding agencies to support unconventional research, particularly if the proposals are assessed by reviewers whose achievements and reputations depend on maintenance of the status quo.

If the changes do not occur, however, researchers will continue their misguided efforts; technological procedures will continue to be appraised inadequately and often misleadingly; and the medical world will continue spending huge sums of money for research that is often unsatisfactory, and for tests that are often ineffectively evaluated and applied.

REFERENCES

- 1 Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep* 1947;**62**:1432-49.
- 2 Wulff HR. *Rational diagnosis and treatment*. 2nd edn.. Oxford: Blackwell Scientific, 1981.
- 3 Hlatky MA, Pryor DB, Harrell FE Jr, *et al*. Factors affecting the sensitivity and specificity of the exercise electrocardiogram: a multivariable analysis. *Am J Med* 1984;**77**:64-71.
- 4 Rozanski A, Diamond GA, Berman D, *et al*. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med* 1983;**309**:518-22.
- 5 Lachs MS, Nachamkin I, Edelstein P, *et al*. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infections. *Ann Intern Med* 1992;**117**:135-40.
- 6 Feinstein AR. *Clinical epidemiology. The architecture of clinical research*. Philadelphia: WB Saunders, 1985.
- 7 Wagner JM, Straus S, Mcalister F, *et al*. The effect of referral on sensitivity and specificity. [Abstract]. *J Gen Intern Med* 2001;**16**:180.
- 8 Reid CM, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. *Am J Med* 1998;**104**:374-80.
- 9 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. *JAMA* 1995;**274**:645-51.
- 10 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926-30.
- 11 Bogardus ST, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research. The need for methodological standards. *JAMA* 1999;**281**:1919-26.
- 12 Hertzman PA, Clauw DJ, Duffy J, *et al*. Rigorous new approach to constructing a "gold-standard" for validating new diagnostic criteria, as exemplified by the eosinophilia-myalgia syndrome. *Arch Intern Med* (in press).