Evidence base of clinical diagnosis

The architecture of diagnostic research

D L Sackett, R B Haynes

Considerable effort has been expended at the interface between clinical medicine and scientific methods to achieve the maximum validity and usefulness of diagnostic tests. This article focuses on the specific kinds of questions that arise in diagnostic research and the study architectures (the conversions of these clinical questions into appropriate research designs) used to answer them. As an example we shall take shall take assessment of the value of the plasma concentration of B-type natriuretic peptide (BNP) in the diagnosis of left ventricular dysfunction.[1] Randomised controlled trials are dealt with elsewhere.

As in other forms of clinical research, there are several different ways studying the potential or real diagnostic value of a physical sign or laboratory test, and each is appropriate to one kind of question and inappropriate for others. Among the possible questions about the relation between a putative diagnostic test and a target disorder (for example, the concentration of BNP and left ventricular dysfunction), four are most relevant.

Summary points

Diagnostic studies should match methods to diagnostic questions

- Do test results in affected patients differ from those in normal individuals?
- Are patients with certain test results more likely to have the target disorder?
- Do test results distinguish patients with and without the target disorder among those in whom it is clinically sensible to suspect the disorder?
- Do patients undergoing the diagnostic test fare better than similar untested patients?

The keys to validity in diagnostic test studies are

- independent, blind comparison of test results with a reference standard among a consecutive series of patients suspected (but not known) to have the target disorder
- inclusion of missing and indeterminate results
- replication of studies in other settings

Both specificity and sensitivity may change as the same diagnostic test is applied in primary, secondary, and tertiary care

Types of question

Phase I questions

Do test results in patients with the target disorder differ from those in normal people? Table 1 shows the architecture of this question.

For example, investigators at a British university hospital measured concentrations of BNP precursor in non-systematic ("convenience") samples from normal controls and from patients who had various combinations of hypertension, ventricular hypertrophy, and left ventricular dysfunction.[2] They found large differences in median concentrations of BNP precursors between the two groups, and no overlap between the ranges. They therefore concluded that testing for BNP concentration was "a useful diagnostic aid for left ventricular dysfunction."

Phase I studies are typically conducted among a group of patients known to have the disease and a group of people definitely known not to have it, rather than patients merely suspected to have it. As a result, this phase of evaluation of a diagnostic test cannot be translated into diagnostic action, but such studies add to our biological insights into mechanisms of disease, and they may serve later research into treatment as well as diagnosis. This kind of study is also quick and relatively cheap to carry out, and a negative result saves having to proceed to the tougher, more time consuming, and costlier questions of phases II-IV.

Table 1 Answering a phase I question: do patients with left ventricular dysfunction have higher concentrations of B-type natriuretic peptide (BNP) precursor than normal individuals?

	Patients known to have disorder	Normal controls
Median (range) concentration of BNP precursor (pg/ml)	493.5 (248.9-909.0)	129.4 (53.6-159.7)

Phase II questions

Are patients with certain test results more likely to have the target disorder than patients with other test results?

Once a phase I question has received a positive answer, it is logical to ask a phase II question, this time changing the direction of interpretation so that it runs from diagnostic test result forward to diagnosis. Although answers to phase II questions often can be

This is the second in a series of five articles

Trout Research and Education Centre at Irish Lake, RR1, Markdale, ON, Canada N0C 1H0
D L Sackett
professor

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada L8N 3Z5
R B Haynes
director

Correspondence to:
D L Sackett
sackett@bmts.com

BMJ 2002;324:539–41

P+

The full list of DLS's extensive potential conflicts of interest appears on the BMJ's website

obtained from the same dataset that provided the phase I answer, the methods of asking and answering phase II questions differ. For example, a second group of investigators, at a Belgian university hospital, measured BNP concentrations in normal controls and in three groups of patients with coronary artery disease and varying degrees of left ventricular dysfunction.[3] Among the analyses they performed was a simple plot of individual BNP results, which generated the results shown in table 2 by picking the cut-off point that best distinguished patients with severe left ventricular dysfunction from normal controls.

The results in table 2 are extremely encouraging. Whether the test is used to rule out left ventricular dysfunction by its high sensitivity (SnNout) or "rule it in" with its high specificity (SpPin),[4] the BNP concentration looks useful, so it is no wonder that the authors concluded that "BNP concentrations are good indicators of the severity and prognosis of congestive heart failure." But is table 2 too encouraging? It compares test results of groups of patients who already have established diagnoses (rather than patients who are merely suspected of the target disorder), and contrasts an extreme group of normal people with a group with severe disease. Thus, it tells us whether the test shows diagnostic promise under ideal conditions. As long as the writers and readers of the report of a phase II explanatory study make no pragmatic claims about its usefulness in routine clinical practice, no harm is done.

Phase III questions

Does the test result distinguish patients with and without the target disorder among patients in whom it is clinically reasonable to suspect that the disease is present?

Given its promise in phase I and II studies, it is understandable that the BNP concentration should be tested in a phase III study to determine whether it is really useful among patients clinically suspected of having left ventricular dysfunction. As we were writing this article, a UK group of clinical investigators reported having done exactly this by inviting general practitioners in their area "to refer patients with suspected heart failure to our clinic."[5] These referred patients (n = 126) underwent independent, blind BNP measurements and echocardiography. Their results are summarised in table 3.

About a third of the patients referred had echocardiographic evidence of left ventricular dysfunction. The investigators reported that measurements of BNP concentration did not look nearly as promising when tested in a phase III study in the real world setting of routine clinical practice and concluded that "introducing routine measurement [of BNP] would be unlikely to improve the diagnosis of symptomatic [left ventricular dysfunction] in the community."

Table 2 Answering a phase II question: are patients with higher concentrations of B-type natriuretic peptide (BNP) more likely to have left ventricular dysfunction than patients with lower concentrations?

	Patients known to have target disorder	Normal controls
High BNP concentration	39	2
Normal BNP concentration	1	25

Test characteristics (95% CI):
- Sensitivity=98% (87% to 100%)
- Specificity=92% (77% to 98%)
- Positive predictive value=95% (84% to 99%)
- Negative predictive value=96% (81% to 100%)
- Likelihood ratio for an abnormal test result=13 (3.5 to 50.0)
- Likelihood ratio for a normal test result=0.03 (0.0003 to 0.19)

Table 3 Answering a phase III question: among patients in whom it is clinically sensible to suspect left ventricular dysfunction (LVD), does the concentration of B-type natriuretic peptide (BNP) distinguish patients with and without left ventricular dysfunction?

	Patients with LVD on echocardiography	Patients with normal results on echocardiography
Concentration of BNP:		
High (>17.9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29
Prevalence (pretest probability) of LVD	40/126=32%	

Test characteristics (95% CI):
- Sensitivity=88% (74% to 94%)
- Specificity=34% (25% to 44%)
- Positive predictive value=38% (29% to 48%)
- Negative predictive value=85% (70% to 94%)
- Likelihood ratio for an abnormal test result=1.3 (1.1 to 1.6)
- Likelihood ratio for a normal test result=0.4 (0.2 to 0.9)

Several threats to the validity of phase III studies can distort their estimates of the accuracy of a diagnostic test. The first is violation of the old guide to critical appraisal, "Has there been an independent, blind comparison with a gold standard of diagnosis?"[4] By "independent" is meant that all study patients have undergone both the diagnostic test and the reference ("gold") standard evaluation and, more specifically, that the reference standard has been applied regardless of the result of the diagnostic test. "Blind" means that the reference standard has been applied and interpreted in total ignorance of the diagnostic test result, and vice versa. Anticipating these threats at the initial, question forming phase of a study allows them to be avoided or minimised.

Another threat to the validity of estimates of accuracy generated in phase III studies arises whenever the selection of the "upper limit of normal" or cut-off point for the diagnostic test is under the control of the investigators. When the investigators are free to place the cut-off point wherever they wish, it is natural for them to place it where it maximises sensitivity, specificity, or the total number of patients correctly classified in that particular "training" set of patients. If the study were repeated in a second, independent "test" set of patients, using that same cut-off point, the diagnostic test would be found to function a little or a lot worse. Thus, the true accuracy of a promising diagnostic test is not known until it has been evaluated in one or more independent studies.

Threats to validity

These threats to validity apply whether the diagnostic test comprises a single measurement of a single phenomenon or a multivariate combination of several phenomena—for example, Wells et al determined the diagnostic accuracy of the combination of several items from the medical history, physical examination, and non-invasive testing in the diagnosis of deep vein thrombosis.[6] Although their study generated similar results in two centres in Canada and one in Italy, they recommended further prospective testing before widespread use of such a combination.

Limits to the applicability of phase III studies

Introductory courses in epidemiology introduce the concept that predictive values change as we move back and forth between screening or primary care settings (with their low prevalence or pretest probability of the target disorder) to secondary and tertiary care (with their higher probability of the target disorder). This point is usually based on the assumption that sensitivity and specificity remain constant across all settings. However, the mix (or spectrum) of patients also varies between these locations—for example, screening is applied to asymptomatic people with early disease, whereas in tertiary care patients have advanced or florid disease.

Because primary care patients with positive diagnostic test results (which comprise false positive as well as true positive results) are referred forward to secondary and tertiary care, we might expect specificity to fall as we move along the referral pathway. Wagner showed this effect in over 2000 patients with clinically suspected appendicitis seen in primary care and on inpatient surgical wards (J Wagner, personal communication, 2000). The diagnostic tests were the clinical signs that are sought when clinicians suspect appendicitis, and the reference standard was a combination of pathology reports on appendices when operations were performed and a benign clinical course when they were not. A comparison of the results in primary and tertiary care showed that the proportion of patients with appendicitis rose from 14% in patients in primary care to 63% in patients in tertiary care, but, of course, this increase in prevalence occurred partly because patients with right lower quadrant tenderness (regardless of whether this was a true positive or false positive finding) tended to be referred to the next level of care, whereas patients without this sign tended not to be referred onward; this is confirmed by the rise in the incidence of this sign from 21% of patients in primary care to 82% of patients in tertiary care. Although this kind of increase in the proportion of positive diagnostic test results is widely recognised, its effect on the accuracy of the test is not. The forward referral of patients with false positive test results leads to a fall in specificity, in this case dramatically from 89% to 16%. As a result, a diagnostic sign of real value in primary care (positive likelihood ratio 8, negative likelihood ratio 0.2) is useless in tertiary care (positive and negative likelihood ratios both 1); in other words, its diagnostic value has been "used up" along the way.

This phenomenon can place major limitations on the applicability of phase III studies carried out in one kind of setting to another setting where the mix of test results may differ. Replicating a promising phase III study in a second "test" setting with patients of the kind whom the test is claimed to benefit avoids this problem. Specificity does not always decrease between primary care and tertiary care settings, and so this feature cannot be used to "adjust" for differences between the two settings.

Clinicians who wish to apply the bayesian properties of diagnostic tests require accurate estimates of the pretest probability of target disorders in their area and setting. These estimates can come from five sources—personal experience, population prevalence figures, practice databases, the publication that described the test, or one of a growing number of primary studies of pretest probability in different settings.[7]

Phase IV questions

Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who are not tested?

The ultimate value of a diagnostic test is measured in the health outcomes that follow from the further diagnostic and therapeutic interventions that the test results precipitate. Sometimes this benefit is self evident, as in the correct diagnosis of patients with life threatening disorders who thereby receive life saving treatments. On other occasions phase III studies may hint at these outcomes if the reference standard for the absence of the target disorder is a benign clinical course without any active treatment. More often, however, when dealing with tests for the early detection of asymptomatic disease, phase IV questions can only be answered by the follow-up of patients randomised to undergo the diagnostic test of interest or some other (or no) test.

Competing interests: DS has been wined, dined, supported, transported, and paid to speak by countless pharmaceutical firms for over 40 years (see bmj.com).

1 Hobbs R. Can heart failure be diagnosed in primary care? *BMJ* 2000;321:188-9.
2 Talwar S, Siebenhofer A, Williams B, Ng L. Influence of hypertension, left ventricular hypertrophy, and left ventricular systolic dysfunction on plasma N terminal pre-BNP. *Heart* 2000;83:278-82.
3 Selvais PL, Donickier JE, Robert A, Laloux O, van Linden F, Ahn S, et al. Cardiac natriuretic peptides for diagnosis and risk stratification in heart failure. *Eur J Clin Invest* 1998;28:636-42.
4 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology, a basic science for clinical medicine*. 2nd ed. Boston: Little Brown, 1991:83.
5 Landray MJ, Lehman R, Arnold I. Measuring brain natriuretic peptide in suspected left ventricular systolic dysfunction in general practice: cross sectional study. *BMJ* 2000;320:985-6.
6 Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process. *J Intern Med* 1998;243:15-23.
7 Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine; how to practise and teach EBM*. 2nd ed. Edinburgh: Churchill Livingstone, 2000:82-4.

One hundred years ago

A new statue to Pasteur

Seven cities in antiquity contended for the honour of being regarded as the birthplace of Homer, and probably as many places in our tight little island could quote expressions used by the late Mr. Gladstone which might be taken to mean that he first saw the light there. As far as we are aware, there can be no such dispute about the birthplace of Pasteur, but in the matter of statues he already leaves both Mr. Gladstone and Homer far behind. The cities, towns, and villages of the pleasant land of France appear to be vying with each other in erecting sculptured memorials of the great investigator. The latest is Marnes, where Pasteur resided during the closing years of his life. The municipality of that town has formed a strong Committee to secure the erection of a monument to his labours. Subscriptions on an international basis are being invited. (*BMJ* 1902;i:609)