2 The architecture of diagnostic research

DAVID L SACKETT, R BRIAN HAYNES

Summary box

- Because diagnostic testing aims to discriminate between clinically "normal" and "abnormal", the definition of "normal" and "the normal range" is a basic issue in diagnostic research. Although the "gaussian" definition is traditionally common, the "therapeutic definition" of normal is the most clinically relevant.
- The diagnostic research question to be answered has to be carefully formulated, and determines the appropriate research approach. The four most relevant types of question are:
- **Phase I questions: Do patients with the target disorder have different test results from normal individuals?** The answer requires a comparison of the distribution of test results among patients known to have the disease and people known not to have the disease.
- **Phase II questions: Are patients with certain test results more likely to have the target disorder than patients with other test results?** This can be studied in the same dataset that generated the Phase I answer, but now test characteristics such as sensitivity and specificity are estimated.
- Only if Phase I and Phase II studies, performed in "ideal circumstances", are sufficiently promising as to possible discrimination between diseased and non-diseased subjects, it is worth evaluating the test under "usual" circumstances. Phase III and IV questions must then be answered.
- **Phase III questions: Among patients in whom it is clinically sensible to suspect the target disorder, does the test result**

distinguish those with and without the target disorder? To get the appropriate answer, a consecutive series of such patients should be studied.
- The validity of Phase III studies is threatened when cases where the reference standard or diagnostic test is lost, not performed, or indeterminate, are frequent or inappropriately dealt with.
- Because of a varying patient mix, test characteristics such as sensitivity, specificity and likelihood ratios may vary between different healthcare settings.
- **Phase IV questions: Do patients who undergo the diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not?** These questions have to be answered by randomising patients to undergo the test of interest or some other (or no) test.

Introduction

When making a diagnosis, clinicians seldom have access to reference or "gold" standard tests for the target disorders they suspect, and often wish to avoid the risks or costs of these reference standards, especially when they are invasive, painful, or dangerous. No wonder, then, that clinical researchers examine relationships between a wide range of more easily measured phenomena and final diagnoses. These phenomena include elements of the patient's history, physical examination, images from all sorts of penetrating waves, and the levels of myriad constituents of body fluids and tissues. Alas, even the most promising phenomena, when nominated as diagnostic tests, almost never exhibit a one-to-one relationship with their respective target disorders, and several different diagnostic tests may compete for primacy in diagnosing the same target disorder. As a result, considerable effort has been expended at the interface between clinical medicine and scientific methods in an effort to maximise the validity and usefulness of diagnostic tests. This book describes the result of those efforts, and this chapter focuses on the specific sorts of questions posed in diagnostic research and the study architectures used to answer them.

At the time that this book was being written, considerable interest was being directed to questions about the usefulness of the plasma concentration of B-type natriuretic peptide in diagnosing left ventricular dysfunction.[1] These questions were justified on two grounds: first, left ventricular dysfunction is difficult to diagnose on clinical examination; and second, randomised trials have shown that treating it (with angiotensin

converting enzyme inhibitors) reduces its morbidity and mortality. Because real examples are far better than hypothetical ones in illustrating not just the overall strategies but also the down-to-earth tactics of clinical research, we will employ this one in the following paragraphs. To save space and tongue twisting we will refer to the diagnostic test, B-type natriuretic peptide, as BNP and the target disorder it is intended to diagnose, left ventricular dysfunction, as LVD. The starting point in evaluating this or any other promising diagnostic test is to decide how we will define its normal range.

What do you mean by "normal" and "the normal range"?

This chapter deals with the strategies (a lot) and tactics (a little) of research that attempts to distinguish patients who are "normal" from those who have a specific target disorder. Before we begin, however, we need to acknowledge that several different definitions of normal are used in clinical medicine, and we confuse them at our (and patients') peril. We know six of them[2] and credit Tony Murphy for pointing out five.[3] A common "*gaussian*" definition (fortunately falling into disuse) assumes that the diagnostic test results for BNP (or some arithmetic manipulation of them) for everyone, or for a group of presumably normal people, or for a carefully characterised "reference" population, will fit a specific theoretical distribution known as the *normal* or *gaussian* distribution. Because the mean of a gaussian distribution plus or minus 2 standard deviations encloses 95% of its contents, it became a tempting way to define the normal several years ago, and came into general use. It is unfortunate that it did, for three logical consequences of its use have led to enormous confusion and the creation of a new field of medicine: the diagnosis of non-disease. First, diagnostic test results simply do not fit the gaussian distribution (actually, we should be grateful that they do not; the gaussian distribution extends to infinity in both directions, necessitating occasional patients with impossibly high BNP results and others on the minus side of zero!). Second, if the highest and lowest 2.5% of diagnostic test results are called abnormal, then all the diseases they represent have exactly the same frequency, a conclusion that is also clinically nonsensical.

The third harmful consequence of the use of the gaussian definition of normal is shared by its more recent replacement, the *percentile*. Recognising the failure of diagnostic test results to fit a theoretical distribution such as the gaussian, some laboratorians have suggested that we ignore the shape of the distribution and simply refer (for example) to the lower (or upper) 95% of BNP or other test results as normal. Although this percentile definition does avoid the problems of infinite and negative test values, it still suggests

that the underlying prevalence of all diseases is similar – about 5% – which is silly, and still contributes to the "upper-limit syndrome" of non-disease because its use means that the only "normal" patients are the ones who are not yet sufficiently worked up. This inevitable consequence arises as follows: if the normal range for a given diagnostic test is defined as including the lower 95% of its results, then the probability that a given patient will be called "normal" when subjected to this test is 95%, or 0.95. If this same patient undergoes two independent diagnostic tests (independent in the sense that they are probing totally different organs or functions), the likelihood of this patient being called normal is now $(0.95) \times (0.95) = 0.90$. So, the likelihood of any patient being called normal is 0.95 raised to the power of the number of independent diagnostic tests performed on them. Thus, a patient who undergoes 20 tests has only 0.95 to the 20th power, or about one chance in three, of being called normal; a patient undergoing 100 such tests has only about six chances in 1000 of being called normal at the end of the work up.*

Other definitions of normal, in avoiding the foregoing pitfalls, present other problems. The *risk factor* definition is based on studies of precursors or statistical predictors of subsequent clinical events; by this definition, the normal range for BNP or serum cholesterol or blood pressure consists of those levels that carry no additional risk of morbidity or mortality. Unfortunately, however, many of these risk factors exhibit steady increases in risk throughout their range of values; indeed, some hold that the "normal" total serum cholesterol (defined by cardiovascular risk) might lie well below 3.9 mmol/L (150 mg%), whereas our local laboratories employ an upper limit of normal of 5.2 mmol/L (200 mg%), and other institutions employ still other definitions.

Another shortcoming of this risk factor definition becomes apparent when we examine the health consequences of acting upon a test result that lies beyond the normal range: will altering BNP or any other risk factor really change risk? For example, although obesity is a risk factor for hypertension, controversy continues over whether weight reduction improves mild hypertension. One of us led a randomised trial in which we peeled 4.1 kg (on average) from obese, mildly hypertensive women with a behaviourally oriented weight reduction programme the (control women lost less than 1 kg).[4] Despite both their and our efforts (the cost of the experimental group's behaviourally oriented weight reduction programme came to US$60 per kilo), there was no accompanying decline in blood pressure.

A related approach defines the normal as that which is *culturally desirable*, providing an opportunity for what HL Mencken called "the corruption of

*This consequence of such definitions helps explain the results of a randomised trial of hospital admission multitest screening that found no patient benefits, but increased healthcare costs, when such screening was carried out.[20]

medicine by morality" through the "confusion of the theory of the healthy with the theory of the virtuous".[5] Although this definition does not fit our BNP example, one sees such definitions in their mostly benign form at the fringes of the current lifestyle movement (for example, "It is better to be slim than fat,"† and "Exercise and fitness are better than sedentary living and lack of fitness"), and in its malignant form in the healthcare system of the Third Reich. Such a definition has the potential for considerable harm, and may also serve to subvert the role of medicine in society.

Two final definitions are highly relevant and useful to the clinician because they focus directly on the clinical acts of diagnosis and therapy. The *diagnostic* definition identifies a range of BNP (or other diagnostic test) results beyond which LVD (or another specific target disorder) is (with known probability) present. It is this definition that we focus on in this book. The "known probability" with which a target disorder is present is known formally as the positive predictive value, and depends on where we set the limits for the normal range of diagnostic test results. This definition has real clinical value and is a distinct improvement over the definitions described above. It does, however, require that clinicians keep track of diagnostic ranges and cut-offs.

The final definition of normal sets its limits at the level of BNP beyond which specific treatments for LVD (such as ACE inhibitors) have been shown conclusively to do more good than harm. This *therapeutic* definition is attractive because of its link with action. The therapeutic definition of the normal range of blood pressure, for example, avoids the hazards of labelling patients as diseased unless they are going to be treated. Thus, in the early 1960s the only levels of blood pressure conclusively shown to benefit from antihypertensive drugs were diastolic pressures in excess of 130 mmHg (phase V). Then, in 1967, the first of a series of randomised trials demonstrated the clear advantages of initiating drugs at 115 mmHg, and the upper limit of normal blood pressure, under the therapeutic definition, fell to that level. In 1970 it was lowered further to 105 mmHg with a second convincing trial, and current guidelines about which patients have abnormal blood pressures that require treatment add an element of the risk factor definition and recommend treatment based on the combination of blood pressure with age, sex, cholesterol level, blood sugar, and smoking habit. These days one can even obtain evidence for blood pressure treatment levels based on the presence of a second disease: for example, in type 2 diabetes the "tight control" of blood pressure reduces the risk of major complications in a cost effective way. Obviously, the use of this therapeutic definition requires that clinicians (and guideline developers) keep abreast of advances in therapeutics, and that is as it should be.

†But the tragic consequences of anorexia nervosa teach us that even this definition can do harm.

In summary, then, before you start any diagnostic study you need to define what you mean by normal, and be confident that you have done so in a sensible and clinically useful fashion.

The question is everything

As in other forms of clinical research, there are several different ways in which one could carry out a study into the potential or real diagnostic usefulness of a physical sign or laboratory test, and each of them is appropriate to one sort of question and inappropriate for others. Among the questions one might pose about the relation between a putative diagnostic test (say, BNP) and a target disorder (say, LVD), four are most relevant:

- **Phase I questions**: Do patients with the target disorder have different test results from normal individuals? (Do patients with LVD have higher BNP than normal individuals?)
- **Phase II questions**: Are patients with certain test results more likely to have the target disorder than patients with other test results? (Are patients with higher BNP more likely to have LVD than patients with lower BNP?)
- **Phase III questions**: Among patients in whom it is clinically sensible to suspect the target disorder, does the level of the test result distinguish those with and without the target disorder? (Among patients in whom it is clinically sensible to suspect LVD, does the level of BNP distinguish those with and without LVD?)
- **Phase IV questions**: Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not? (Of greatest interest in evaluating early diagnosis through screening tests, this might be phrased: Do patients screened with BNP (in the hope of achieving the early diagnosis of LVD) have better health outcomes (mortality, function, quality of life) than those who do not undergo screening?).

At first glance the first three questions may appear indistinguishable or even identical. They are not, because the strategies and tactics employed in answering them are crucially different, and so are the conclusions that can be drawn from their answers. The first two differ in the "direction" in which their results are analysed and interpreted, and the third differs from the first two as well in the fashion in which study patients are assembled. The fourth question gets at what we and our patients would most like to know: are they better off for having undergone it? The conclusions that can (and, more importantly, cannot) be drawn from the answers to these questions are crucially different, and there are plenty of examples of the price paid by patients and providers when the answers to Phase I or II questions are

interpreted as if they were answering a Phase III (or even a Phase IV) question.

These questions also nicely describe an orderly and efficient progression of research into the potential usefulness of a clinical sign, symptom, or laboratory result, and we will use the BNP story to show this sequence.

Phase I questions: Do patients with the target disorder have different test results from normal individuals?

Question 1 often can be answered with a minimum of effort, time, and expense, and its architecture is displayed in Table 2.1.

For example, a group of investigators at a British university hospital measured BNP precursor in convenience samples of "normal controls" and in patients who had various combinations of hypertension, ventricular hypertrophy, and LVD.[6] They found statistically significant differences in median BNP precursors between patients with and normal individuals without LVD, and no overlap in their range of BNP precursor results. It was not surprising, therefore, that they concluded that BNP was "a useful diagnostic aid for LVD".

Note, however, that the direction of interpretation here is from known diagnosis back to diagnostic test. Answers to Phase I questions cannot be applied directly to patients because they are presented as overall (usually average) test results. They are not analysed in terms of the diagnostic test's sensitivity, specificity, or likelihood ratios. Moreover, Phase I studies are typically conducted among patients known to have the disease and people known not to have the disease (rather than among patients who are suspected of having, but not known to have, the disease). As a result, this phase of diagnostic test evaluation cannot be translated into diagnostic action.

Why, then, ask Phase I questions at all? There are two reasons. First, such studies add to our biologic insights about the mechanisms of disease, and may serve later research into therapy as well as diagnosis. Second, such studies are quick and relatively cheap, and a negative answer to their question removes the need to ask the tougher, more time-consuming, and costlier questions of Phases II–IV. Thus, if a convenience (or "grab")

Table 2.1 Answering a Phase I question: Do patients with LVD have higher BNP than normal individuals?

	Patients known to have the target disorder (LVD)	Normal controls
Average diagnostic test (BNP precursor) result (and its range)	493.5 (range from 248.9 to 909)	129.4 (range from 53.6 to 159.7)

sample of patients with LVD already known to the investigators displays the same average levels and distribution of BNP as apparently healthy laboratory technicians or captive medical students, it is time to abandon it as a diagnostic test and devote scarce resources to some other lead.

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

Following a positive answer to a Phase I question, 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 the Phase II questions often can be asked in the same dataset that generated the Phase I answer, the architecture of asking and answering them differs. For example, a second group of investigators at a Belgian university hospital measured BNP in "normal subjects" and 3 groups of patients with coronary artery disease and varying degrees of LVD.[7] Among the analyses they performed (including the creation of ROC curves; see Chapter 7) was a simple plot of individual BNP results, generating the results shown in Table 2.2 by picking the cut-off that best distinguished their patients with severe LVD from their normal controls.

As you can see, the results in Table 2.2 are extremely encouraging. Whether it is used to "rule out" LVD on the basis of its high sensitivity (SnNout)[8] or to "rule in" LVD with its high specificity (SpPin),[9] BNP looks useful, so it is no wonder that the authors concluded: "BNP concentrations are good indicators of the severity and prognosis of

Table 2.2 Answering a Phase II question: Are patients with higher BNP more likely to have LVD than patients with lower BNP?

	Patients known to have the target disorder (LVD)	Normal controls
High BNP	39	2
Normal BNP	1	25
Test characteristics and their 95% confidence intervals	Lower	Upper
Sensitivity = 98%	87%	100%
Specificity = 92%	77%	98%
Positive predictive value = 95%	84%	99%
Negative predictive value = 96%	81%	100%
Likelihood ratio for an abnormal test result = 13	3.5	50
Likelihood ratio for a normal test result = 0.03	0.0003	0.19

congestive heart failure". But is Table 2.2 overly encouraging? It compares test results between groups of patients who already have established diagnoses (rather than those who are merely suspected of the target disorder), and contrasts extreme groups of normals and those with severe disease. Thus, it tells us whether the test shows diagnostic promise under ideal conditions. A useful way to think about this difference between Phase II

Table 2.3 Explanatory and pragmatic studies of diagnostic tests and treatments.

	Promising diagnostic test		Promising treatment	
Feature	Explanatory (Phase II study)	Pragmatic (Phase III study)	Explanatory	Pragmatic
Question	Can this test discriminate under ideal circumstances?	Does this test discriminate in routine practice?	Efficacy: Can this treatment work under ideal circumstances?	Effectiveness: Does this treatment work in routine practice?
Selection of patients	Preselected groups of normal individuals and of those who clearly have the target disorder	Consecutive patients in whom it is clinically sensible to suspect the target disorder	Highly compliant, high-risk, high-response patients	All comers, regardless of compliance, risk or responsiveness
Application of manoeuvre	Carried out by expert clinician or operator on best equipment	Carried out by usual clinician or operator on usual equipment	Administered by experts with great attention to compliance	Administered by usual clinicians under usual circumstances
Definition of outcomes	Same reference standard for those with and without the target disorder	Often different standards for patients with and without the target disorder; may invoke good treatment-free prognosis as proof of absence of target disorder	May focus on pathophysiology, surrogate outcomes, or cause-specific mortality	"Hard" clinical events or death (often all-cause mortality)
Exclusion of patients or events	Often exclude patients with lost results and indeterminate diagnoses	Include all patients, regardless of lost results or indeterminate diagnoses	May exclude events before or after treatment is applied	Includes all events after randomisation
Results confirmed in a second, independent ("test") sample of patients	Usually not	Ideally yes		
Incorporation into systematic review	Usually not	Ideally yes	Sometimes	Ideal

and Phase III studies is by analogy with randomised clinical trials, which range from addressing explanatory (efficacy) issues of therapy (can the new treatment work under ideal circumstances?) to management (pragmatic, effectiveness) issues (does the new treatment work under usual circumstances?). We have summarised this analogy in Table 2.3.

As shown in Table 2.3, the Phase II study summarised in Table 2.2 is explanatory in nature: preselected groups of normal individuals (ducks) and those who clearly have the target disorder (yaks) undergo testing under the most rigorous circumstances possible, with the presence or absence of the target disorder being determined by the same reference standard. No attempt is made to validate these initial ("training set") results (especially the cut-off used to set the upper limit of normal BNP) in a second, independent "test" set of ducks and yaks. On the other hand, and as with the Phase I study, this relatively easy Phase II investigation tells us whether the promising diagnostic test is worth further, costlier evaluation; as we have said elsewhere,[10] if the test cannot tell the difference between a duck and a yak it is worthless in diagnosing either one. As long as the writers and readers of a Phase II explanatory study report make no pragmatic claims about its usefulness in routine clinical practice, no harm is done. Furthermore, criticisms of Phase II explanatory studies for their failure to satisfy the methodological standards employed in Phase III pragmatic studies do not make sense.

Phase III questions: Among patients in whom it is clinically sensible to suspect the target disorder, does the level of the test result distinguish those with and without the target disorder?

Given its promise in Phase I and II studies, it is understandable that BNP would be tested in the much costlier and more time-consuming Phase III study, in order to determine whether it was really useful among patients in whom it is clinically sensible to suspect LVD. As we were writing this chapter, an Oxfordshire group of clinical investigators reported that they did just that by inviting area general practitioners "to refer patients with suspected heart failure to our clinic".[11] Once there, these 126 patients underwent independent, blind BNP measurements and echocardiography. Their results are summarised in Table 2.4.

About one third of the patients referred by their general practitioners had LVD on echocardiography. These investigators documented that BNP measurements did not look nearly as promising when tested in a Phase III study in the pragmatic real-world setting of routine clinical practice, and concluded that "introducing routine measurement [of BNP] would be unlikely to improve the diagnosis of symptomatic [LVD] in the

Table 2.4 Answering a Phase III question: Among patients in whom it is clinically sensible to suspect LVD, does the level of BNP distinguish patients with and without LVD?

	Patients with LVD on echocardiography	Patients with normal echoes
High BNP (>17.9 pg/ml)	35	57
Normal BNP (<18 pg/ml)	5	29
Prevalence or pretest probability of LVD	40/126 = 32%	
Test characteristics and their 95% confidence intervals	Lower	Upper
Sensitivity = 88%	74%	94%
Specificity = 34%	25%	44%
Positive predictive value = 38%	29%	48%
Negative predictive value = 85%	70%	94%
Likelihood ratio for an abnormal test result = 1.3	1.1	1.6
Likelihood ratio for a normal test result = 0.4	0.2	0.9

Table 2.5 Answering a Phase III question with likelihood ratios.

	Patients with LVD on echocardiography	Patients with normal echoes	Likelihood ratio and 95% CI
High BNP (>76 pg/ml)	26 (0.650)	11 (0.128)	5.1 (2.8–9.2)
Mid BNP (10–75 pg/ml)	11 (0.275)	60 (0.698)	0.4 (0.2–0.7)
Low BNP (<10 pg/ml)	3 (0.075)	15 (0.174)	0.4 (0.1–1)
Total	40 (1.000)	86 (1.000)	

community". However, their report of the study also documented the effect of two other cut-points for BNP. This led both to a counterclaim on the usefulness of BNP in the subsequent email letters to the editor, and to an opportunity for us to describe an alternative way of presenting information about the accuracy of a diagnostic test: the multilevel likelihood ratio (LR). The original report makes it possible for us to construct Table 2.5.

By using multilevel likelihood ratios to take advantage of the full range of BNP results, we can be slightly more optimistic about the diagnostic usefulness of higher levels: the LR for BNP results >76 pg/ml was 5.1. These levels were found in 29% of the patients in this study, and their presence raised the pretest probability of LVD in the average patient from 32% to a post-test probability of 70%. This can be determined directly from Table 2.5 for this "average" patient with a pretest probability of 32% and a high BNP: reading horizontally across the top row, the result is 26/(26+11) = 70%.

However, if the patient has a different pretest likelihood, say 50%, then either the table must be reconstructed for this higher figure, or the pretest probability needs to be converted to a pretest odds (50% is a pretest odds of $(1-0.5)/0.5 = 1$), and then multiplied by the likelihood ratio for the test result (5.1 in this case), giving a post-test odds of 5.1, which then can be converted back into a post-test probability of $5.1/(1+5.1) = 84\%$. These calculations are rendered unnecessary by using a nomogram, as in Figure 2.1.

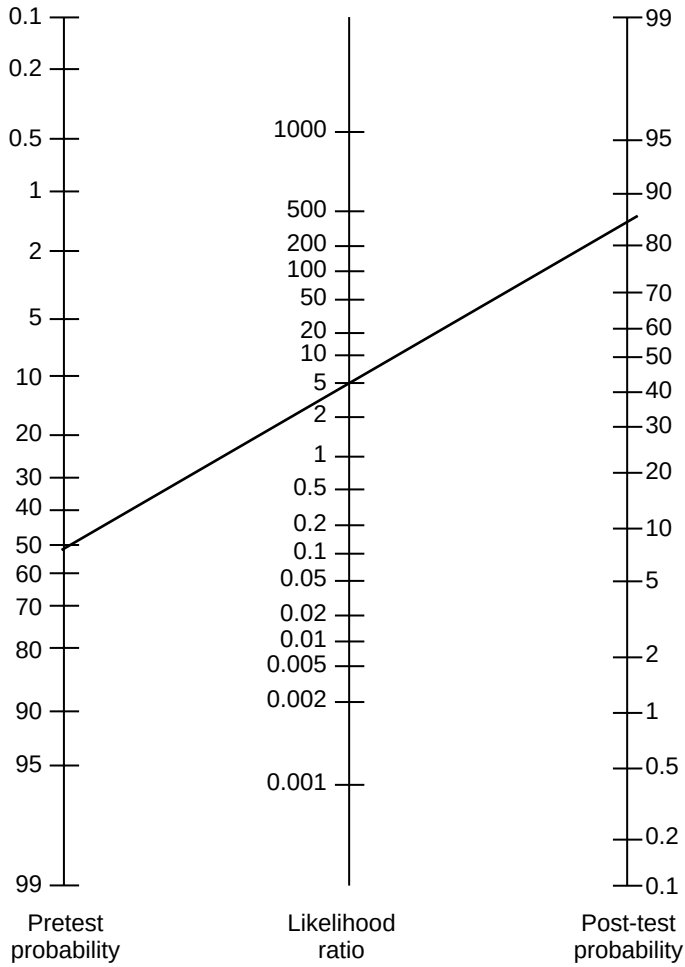

Figure 2.1 Nomogram for converting pretest likelihoods (left column) to post-test likelihoods (right column) by drawing a straight line from the pretest likelihood through the likelihood ratio for the test result.

Given the quite wide confidence intervals around these LRs, further type III studies may be fruitful (and readers can check to see whether this was done after this chapter was published).

Threats to the validity of Phase III studies

There are several threats to the validity of Phase III studies that distort their estimates of the accuracy of the diagnostic test, and the first batch are violations of the old critical appraisal guide: "Has there been an independent, blind comparison with a gold standard of diagnosis?"[12] By *independence* we mean that *all* study patients have undergone *both* the diagnostic test *and* the reference ("gold") standard evaluation and, more specifically, that the reference standard is applied *regardless of the diagnostic test result*. By *blind* we mean that the reference standard is applied and interpreted in total ignorance of the diagnostic test result, and vice versa. By anticipating these threats at the initial question forming phase of a study, they can be avoided or minimised.

Although we prefer to conceptualise diagnostic test evaluations in terms of 2×2 tables such as the upper panel of Table 2.6 (and this is the way that most Phase II studies are performed), in reality Phase III studies generate the 3×3 tables shown in the lower panel of Table 2.6. Reports get lost, their results are sometimes incapable of interpretation, and sometimes we are unwilling to apply the reference standard to all the study patients.

The magnitude of the cells ***v***–***z*** and the method of handling patients who fall into these cells will affect the validity of the study. In the perfect study these cells are kept empty, or so small that they cannot exert any important

Table 2.6 The ideal Phase III study meets the real world.

	Reference standard	
The ideal study	Target disorder present	Target disorder absent
Diagnostic test result		
Positive	***a***	***b***
Negative	***c***	***d***

	Reference standard		
The real study	Target disorder present	Lost, not performed, or indeterminate	Target disorder absent
Diagnostic test result			
Positive	***a***	***v***	***b***
Lost, not performed, or indeterminate	***w***	***x***	***y***
Negative	***c***	***z***	***d***

effect on the study conclusions. However, there are 6 situations in which they become large enough to bias the measures of test accuracy. First, when the reference standard is expensive, painful, or risky, investigators will not wish to apply it to patients with negative diagnostic test results. As a consequence, such patients risk winding up in cell ***z***. Furthermore, there is an understandable temptation to shift them to cell ***d*** in the analysis. Because no diagnostic test is perfect, some of them surely belong in cell ***c***. Shifting all of them to cell ***d*** falsely inflates both sensitivity and specificity. If this potential problem is recognised before the study begins, investigators can design their reference standard to prevent such patients from falling into cell ***z***. This is accomplished by moving to a more pragmatic study and adding another, prognostic dimension to the reference standard, namely the clinical course of patients with negative test results who receive no intervention for the target disorder. If patients who otherwise would end up in cell ***z*** develop the target disorder during this treatment-free follow up, they belong in cell ***c***. If they remain free of disease, they join cell ***d***. The result is an unbiased and pragmatic estimate of sensitivity and specificity.

Second, the reference standard may be lost; and third, it may generate an uninterpretable or indeterminate result. As before, arbitrarily analysing such patients as if they really did or did not have the target disorder will distort measures of diagnostic test accuracy. Once again, if these potential biases are identified in the planning stages they can be minimised, a pragmatic solution such as that proposed above for cell ***z*** considered, and clinically sensible rules established for shifting them to the definitive columns in a manner that confers the greatest benefit (in terms of treatment) and the least harm (in terms of labelling) to later patients.

Fourth, fifth, and sixth, the diagnostic test result may be lost, never performed, or indeterminate, so that the patient winds up in cells ***w***, ***x***, or ***y***. Here the only unforgivable action is to exclude such patients from the analysis of accuracy. As before, anticipation of these problems before the study begins should minimise tests that are lost or never performed to the point where they would not affect the study conclusion regardless of how they were classified. If indeterminate results are likely to be frequent, a decision can be made before the study begins as to whether they will be classified as positive or negative. Alternatively, if multilevel likelihood ratios are to be used, these patients can form their own stratum.

In addition to the 6 threats to validity related to cells ***v–z***, there are two more. The seventh threat to validity noted in the above critical appraisal guide arises when a patient's reference standard is applied or interpreted by someone who already knows that patient's diagnostic test result (and vice versa). This is a risk whenever there is any degree of interpretation (even in reading off a scale) involved in generating the result of the diagnostic test or reference standard. We know that these situations lead to biased inflations of sensitivity and specificity.

The eighth and final threat to the validity of accuracy estimates generated in Phase III studies arises whenever the selection of the "upper limit of normal" or cut-point for the diagnostic test is under the control of the investigator. When they can place the cut-point wherever they want, it is natural for them to select the point where it maximises sensitivity (for use as a SnNout), specificity (for use as a SpPin), or the total number of patients correctly classified in that particular "training" set. If the study were repeated in a second, independent "test" set of patients, employing that same cut-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.

The foregoing threats apply whether the diagnostic test comprises a single measurement of a single phenomenon or a multivariate combination of several phenomena. For example, Philip Wells and his colleagues 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.[13] Although their study generated similar results in three different centres (two in Canada and one in Italy), even they recommended further prospective testing before widespread use.

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 made by assuming 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 individuals with early disease, whereas tertiary care settings deal with patients with advanced or florid disease. No wonder, then, that sensitivity and specificity often vary between these settings. Moreover, 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. There is very little empirical evidence addressing this issue, and we acknowledge our debt to Dr James Wagner of the University of Texas at Dallas for tracking down and systematically reviewing diagnostic data from over 2000 patients with clinically suspected appendicitis seen in primary care and on inpatient surgical wards (personal communication, 2000). The diagnostic tests comprised the clinical signs that are sought when clinicians suspect appendicitis, and the reference standard is a combination of pathology

Table 2.7 The accuracy of right lower quadrant tenderness in the diagnosis of appendicitis.

	Primary care settings Appendicitis		Tertiary care settings Appendicitis	
	Yes (%)	No (%)	Yes (%)	No (%)
Right lower quadrant tenderness				
Present	84	11	81	84
Absent	16	89	19	16
Total	100	100	100	100
Frequency of appendicitis	14%		63%	
Frequency of positive sign	21%		82%	
Sensitivity	84%		81%	
Specificity	89%		16%	
LR+	7.6		1	
LR−	0.2		1	

reports on appendices when operations were performed, and a benign clinical course when they were not. The results for the diagnostic test of right lower quadrant tenderness are shown in Table 2.7.

A comparison of the results in primary and tertiary care shows, as we might expect, an increase in the proportions of patients with appendicitis (from 14% to 63%). 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 frequency of this sign from 21% of patients in primary care to 82% of patients in tertiary care. Although this sort of increase in a positive diagnostic test result 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 a dramatic one from 89% down to 16%. As a result, a diagnostic sign of real value in primary care (LR+ of 8, LR− of 0.2) is useless in tertiary care (LR+ and LR− 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 sort of setting to another setting where

‡Although not germane to this book on research methods, there are two major clinical ramifications of this phenomenon. First, because clinical signs and other diagnostic tests often lose their value along the referral pathway, tertiary care clinicians might be forgiven for proceeding immediately to applying invasive reference standards. Second, tertiary care teachers should be careful what they teach primary care trainees about the uselessness of clinical signs.

Table 2.8 The accuracy of abdominal rigidity in the diagnosis of appendicitis.

	Primary care settings Appendicitis		Tertiary care settings Appendicitis	
	Yes (%)	No (%)	Yes (%)	No (%)
Rigid abdomen				
Present	40	26	23	6
Absent	60	74	77	94
Total	100	100	100	100
Frequency of appendicitis	14%		47%	
Frequency of positive sign	28%		14%	
Sensitivity	40%		24%	
Specificity	74%		94%	
LR+	1.5		5	
LR−	0.8		0.8	

the mix of test results may differ. Overcoming this limitation is another bonus that attends the replication of a promising Phase III study in a second "test" setting attended by patients of the sort that the test is claimed to benefit.

Does specificity always fall between primary care and tertiary care settings? Might this be employed to generate a "standardised correction factor" for extrapolating test accuracy between settings? Have a look at the clinical sign of abdominal rigidity in Table 2.8.

In this case, a clinical sign that is useless in primary care (LR+ barely above 1 and LR− close to 1) is highly useful in tertiary care (LR+ of 5), and in this case specificity has risen (from 74% to 95%), not fallen, along the referral pathway. The solution to this paradox is revealed in the frequency of the sign in these two settings; it has fallen (from 28% to 14%), not risen, along the pathway from primary to tertiary care. We think that the explanation is that primary care clinicians, who do not want to miss any patient's appendicitis, are "over-reading" abdominal rigidity compared to their colleagues in tertiary care. At this stage in our knowledge of this phenomenon we do not think the "standard correction factors" noted in the previous paragraph are advisable, and this paradox once again points to the need to replicate promising Phase III study results in "test" settings attended by patients (and clinicians!) of the sort that the test is claimed to benefit. In this regard we welcome the creation of the CARE consortium of over 800 clinicians from over 70 countries[14] for their performance of web-based, large, simple, fast studies of the clinical examination.[15] It is hoped that this group, which can be contacted at www.carestudy.com, can make a large contribution to determining the wide applicability of the

diagnostic test information obtained from the medical history and physical examination.

For clinicians who wish to apply the bayesian properties of diagnostic tests, accurate estimates of the pretest probability of target disorders in their locale and setting are required. These can come from five sources: personal experience, population prevalence statistics, practice databases, the publication that described the test, or one of a growing number of primary studies of pretest probability in different settings.[16]

Phase IV questions: Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who do not?

The ultimate value of a diagnostic test is measured in the health outcomes produced by the further diagnostic and therapeutic interventions it precipitates. Sometimes this benefit is self-evident, as in the correct diagnosis of patients with life threatening target disorders who thereby receive life saving treatments. At other times these outcomes can be hinted at in Phase III studies if the reference standard for the absence of the target disorder is a benign clinical course despite the withholding of treatment. More often, however, Phase IV questions are posed about diagnostic tests that achieve the early detection of asymptomatic disease, and can only be answered by the follow up of patients randomised to undergo the diagnostic test of interest or some other (or no) test.

Methods for conducting randomised trials are discussed elsewhere,[17] and we will confine this discussion to an example of the most powerful sort, a systematic review of several randomised trials of faecal occult blood testing.[18] In these trials, over 400 000 patients were randomised to undergo annual or biennial screening or no screening, and then carefully followed for up to 13 years in order to determine their mortality from colorectal cancer. The results are summarised in Table 2.9.

Table 2.9 A systematic review of randomised trials of screening for colorectal cancer.

Outcome	Unscreened group	Screened group	Relative risk reduction	Absolute risk reduction	Number needed to screen to prevent one more colorectal cancer death
Colorectal cancer mortality	0.58%	0.50%	16%	0.08%	1237

In this example, patients were randomised to undergo or not undergo the diagnostic test. Because most of them remained cancer free, the sample size requirement was huge and the study architecture is relatively inefficient. It would have been far more efficient (but unacceptable) to randomise the disclosure of positive test results, and this latter strategy was employed in a randomised trial of a developmental screening test in childhood.[19] In this study, the experimental children whose positive test results were revealed and who subsequently received the best available counselling and interventions fared no better in their subsequent academic, cognitive or developmental performance than control children whose positive test results were concealed. However, parents of the "labelled" experimental children were more likely to worry about their school performance, and their teachers tended to report more behavioural problems among them. This warning that diagnostic tests can harm as well as help those who undergo them is a suitable stopping point for this chapter.

References

1 Hobbs R. Can heart failure be diagnosed in primary care? *BMJ* 2000;**321**:188–9.
2 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: a Basic Science for Clinical Medicine*, 2nd edn. Boston: Little, Brown and Company, 1991; pp. 58–61.
3 Murphy EA. The normal, and perils of the sylleptic argument. *Perspect Biol Med* 1972;**15**:566.
4 Haynes RB, Harper AC, Costley SR, *et al*. Failure of weight reduction to reduce mildly elevated blood pressure: a randomized trial. *J Hypertension* 1984;**2**:535.
5 Mencken HL. *A Mencken chrestomathy*. Westminster: Knopf, 1949;12.
6 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.
7 Selvais PL, Donickier JE, Robert A, *et al*. Cardiac natriuretic peptides for diagnosis and risk stratification in heart failure. *Eur J Clin Invest* 1998;**28**:636–42.
8 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 83.
9 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 77.
10 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 57.
11 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.
12 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. op cit; p. 52.
13 Wells PS, Hirsh J, Anderson DR, *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.
14 McAlister FA, Straus SE, Sackett DL, on behalf of the CARE-COAD group. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. *Lancet* 1999;**354**:1721–4.
15 Straus SE, McAlister FA, Sackett DL, Deeks JJ. The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airway disease. *JAMA* 2000; **283**:1853–7.
16 Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: how to Practise and Teach EBM*, 2nd edn. Edinburgh: Churchill Livingstone, 2000;82–4.
17 Shapiro SH, Louis TA. *Clinical Trials: Issues and Approaches*, 2nd edn. New York: Marcel Dekker, 2000; (in press).

18 Towler BP, Irwig L, Glasziou P, Weller D, Kewenter J. *Screening for colorectal cancer using the faecal occult blood test, Hemoccult.* Cochrane Review, latest version 16 Jan 1998. In: The Cochrane Library, Oxford: Update Software.
19 Cadman D, Chambers LW, Walter SD, Ferguson R, Johnston N, McNamee J. Evaluation of public health preschool child development screening: The process and outcomes of a community program. *Am J Public Health* 1987;**77**:45–51.
20 Dunbridge TC, Edwards F, Edwards RG, Atkinson M. *An evaluation of multiphasic screening on admission to hospital.* Précis of a report to the National Health and Medical Research Council. *Med J Aust* 1976;**1**:703–5.