14/7/09

CLINICAL EPIDEMIOLOGY ROUNDS

How to read clinical journals: II. To learn about a diagnostic test

DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS,
MCMASTER UNIVERSITY HEALTH SCIENCES CENTRE

The first round in this series (*Can Med Assoc J* 1981; 124: 555–558) presented 10 reasons to read clinical journals and introduced a flowchart of guides for reading them (Fig. 1) that suggests four universal guides for any article (consider the title, the authors, the summary and the site) and points out that further guides for reading (and discarding) articles depend on why they are being read.

This round will present guides for reading articles that describe diagnostic tests, both old and new. First, however, we must give some nominal definitions.

The serum level of thyroxine (T_4) can be measured in at least four circumstances, and it is important for us to tell them apart. First, a group of passers-by in a shopping plaza or the members of a senior citizens' club may be invited to have a free T_4 test; this testing of apparently healthy volunteers from the general population for the purpose of separating them into groups with high and low probabilities for thyroid disease is called *screening*. Second, patients who come to a clinician's office for any illness may have a T_4 test routinely added to whatever laboratory studies are undertaken to diagnose their chief complaints; this testing of patients for disorders that are unrelated to the reason they came to the clinician is called *case finding*. Third, a T_4 test may be specifically ordered to explain the exact cause for a patient's presenting illness; this, of course, is *diagnosis*. Finally, a T_4 test may be ordered for a patient who is taking a replacement hormone or who has previously received therapeutic radioiodine in order to *test for achievement of a treatment goal*. This round will deal mostly with diagnosis, and later rounds will take up the other three uses of paraclinical data such as a T_4 determination.

Guides for reading articles about diagnostic tests

When encountering an article that looks like it might be describing a

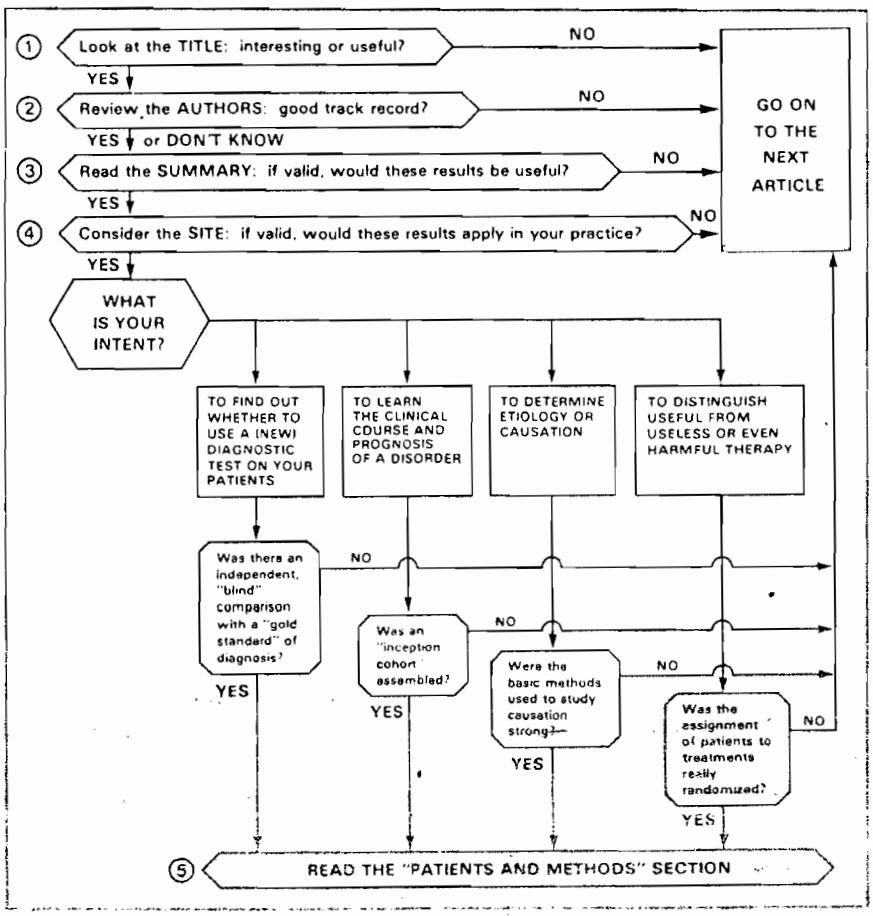

FIG. 1—The first steps in how to read articles in a clinical journal.

Reprint requests to: Dr. R.B. Haynes, McMaster University Health Sciences Centre, Rm. 3V43D, 1200 Main St. W, Hamilton, Ont. L8N 3Z5

useful diagnostic test (that is, the title is interesting, the authors have a good track record, the summary shows it would be very helpful if it really works as claimed, and the site is similar to your own), what should you seek in the Methods portion of the paper?

The eight elements of a proper clinical evaluation of a diagnostic test appear in Table I.[1-4] They constitute guides for the clinical reader and will be considered in order.

1. *Was there an independent, "blind" comparison with a "gold standard" of diagnosis?*

Patients shown (by application of an accepted "gold standard" of diagnosis, such as a biopsy) to have the disease of interest, plus a second group of patients shown (by application of the same gold standard) not to have this disease should have undergone the diagnostic test, and the test should have been interpreted by clinicians who didn't know (that is, they were "blind" to) whether a given patient really had the disease. Afterward, these diagnostic test results should be compared with the gold standard.

The most straightforward method of displaying the comparison of a diagnostic test and a gold standard is with a "two-by-two" or "fourfold" table (Table II). The key words in such comparisons are *sensitivity*, *specificity* and *predictive value*. If you don't see at least the first two words in the abstract, beware. If you don't find or cannot construct a fourfold table from a sneak preview of the Results section, it's probably not worth your time to read any further; toss the article out and go to the next one.

If the article survives this quick screening test, a great deal of useful information can be derived from comparing the diagnostic test results and the gold standard. Here are the basic concepts:

First, the gold standard refers to a definitive diagnosis attained by biopsy, surgery, autopsy, long-term follow-up or another acknowledged standard. If you can't accept the gold standard (within reason, that is — nothing's perfect!) then you should abandon the article.* If you do accept the gold standard, then consider the diagnostic test: Does it have something to offer that the gold standard does not? For example, is it less risky, less uncomfortable or less embarrassing for the patient, less costly or applicable earlier in the course of the illness? Again, if the proposed diagnostic test offers no theoretical advantage over the gold standard, why read further?

Having satisfied yourself that it's worth proceeding, you are now ready to study the comparison between the diagnostic test results and the gold standard. There are several useful elements of this comparison and we will cover them one by one, introducing their associated technical jargon along the way.

The first two elements of this comparison consider how well the diagnostic test correctly identifies patients with and without the disease of interest. Consider the vertical columns of Table II. The gold standard has identified (a + c) patients as having the disease of interest, and the "a" patients had positive diagnostic test results. Thus,

*Of course, the gold standard mustn't include the diagnostic test result as one of its components, for the resulting "incorporation bias" would invalidate the whole comparison.[3]

Table I—Elements of the proper clinical evaluation of a diagnostic test

1. Was there an independent, "blind" comparison with a "gold standard" of diagnosis?
2. Did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?
3. Was the setting for the study, as well as the filter through which study patients passed, adequately described?
4. Was the reproducibility of the test result (precision) and its interpretation (observer variation) determined?
5. Was the term "normal" defined sensibly?
6. If the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined?
7. Were the tactics for carrying out the test described in sufficient detail to permit their exact replication?
8. Was the "utility" of the test determined?

Table II—Fourfold table demonstrating "blind" comparison with "gold standard"

		Gold standard: Patient *has* the disease	Gold standard: Patient does *not have* the disease	
Test result (conclusion drawn from the results of the test)	Positive: Patient appears *to have* the disease	True positive (a)	False positive (b)	a + b
	Negative: Patient appears *not to have* the disease	False negative (c)	True negative (d)	c + d
		a + c	b + d	a + b + c + d

Stable properties:
- a/(a + c) = sensitivity
- d/(b + d) = specificity

Frequency-dependent properties:
- a/(a + b) = positive predictive value*
- d/(c + d) = negative predictive value
- (a + d)/(a + b + c + d) = accuracy
- (a + c)/(a + b + c + d) = prevalence

*Positive predictive value can be calculated other ways too. One of them uses Bayes' theorem:

$$\frac{(\text{prevalence})(\text{sensitivity})}{(\text{prevalence})(\text{sensitivity}) + (1 - \text{prevalence})(1 - \text{specificity})}$$

an index of the diagnostic test's ability to detect the disease when it is present is a/(a + c), usually expressed as a percentage and, for purposes of quick communication, referred to as sensitivity. Similarly, the ability of the diagnostic test to correctly identify the absence of the disease is shown in the next vertical column as d/(b + d); this index goes by the name specificity. Sensitivity and specificity can be considered the *stable properties* of the test because they do not change when different proportions of diseased and well patients are tested; this is an important issue, and we'll come back to it.

But stop a moment to consider the usual clinical situation. When we attempt to diagnose a patient's illness we do not have the results of a gold standard to go by. (If we did we would not bother to order the less definitive diagnostic test because we would already have more information than it can provide.) We are operating horizontally in Table II, not vertically. Thus, in judging the value of a diagnostic test, what we wish to know is not its sensitivity and specificity but what it means when it is positive or negative. That is, we want to know how well its results will predict the results of applying the gold standard. If this prediction is good enough we will add it to our bag of diagnostic tricks.

Accordingly, we are primarily interested in the horizontal properties of the diagnostic test. Among (a + b) patients, those with a positive diagnostic test result, in what proportion, a/(a + b), have we correctly predicted, or "ruled in", the correct diagnosis? This proportion a/(a + b), again usually expressed as a percentage, goes by the name *positive predictive value.*

Similarly, we want to know how well a negative test result correctly predicts the absence of, or "rules out", the disease in question. This proportion, d/(c + d), is named the *negative predictive value.*

Another property of interest is the overall rate of agreement between the diagnostic test and the gold standard. Table II reveals that this could be expressed by the fraction (a + d)/(a + b + c + d); this rate is usually called *accuracy.**

If a diagnostic test's predictive value constitutes the focus of our clinical interest, why waste time considering its sensitivity and specificity? The reason is a fundamental one that has major implications, not just for the rational use of diagnostic tests, but also for the basic education of clinicians. Put simply, a diagnostic test's positive and negative predictive values fluctuate widely, depending on the proportion of truly diseased individuals among patients to whom the test is applied — in Table II this is the proportion (a + c)/(a + b + c + d), a property called *prevalence.*

Although a diagnostic test's sensitivity and specificity remain constant (or "stable") with changes in the proportions of diseased and well people who are tested, its predictive values and accuracy can change markedly (and thus are "unstable") when the prevalence of illness changes. This is not a theoretical concern: in the real world the prevalence of a given condition varies considerably between patients tested in primary and tertiary care centres, as we saw in the hypertension example that concluded the previous round. Furthermore, during their initial development most diagnostic tests are evaluated among equal numbers of individuals with and without the disease of interest (i.e., a contrived prevalence of 50%). This is almost always a higher prevalence than is seen in clinical practice, even in tertiary care centres.

Because an understanding of the effect of prevalence on the stable and unstable properties of diagnostic tests is central to their rational use, and because those of us who are generating these rounds are convinced that active problem-solving beats passive absorption, we invite you to work through the following example.[1]

Several investigators carefully studied a group of men referred with chest pain. Following graded treadmill stress testing (the diagnostic test) and selective coronary arteriography (the gold standard), they obtained the results shown in Table III.[5] The ability of the post-exercise electrocardiogram (ECG) to predict the results of selective coronary arteriography was revealed in its positive predictive value of 89% (the percentage of men with positive ECGs whose arteriograms showed stenosis of 75% or more) and its negative predictive value of 63% (the percentage of men with negative ECGs whose arteriograms showed less than 75% stenosis). Accordingly, the authors concluded: "In men a positive multistage stress test is useful in predicting the presence of significant coronary artery disease although a negative stress test cannot be relied upon to rule out the presence of significant disease."[5]

As you can see from the gold

*Galen and Gambino,[4] who have written a very thorough and easily understood book on this topic, call this property "efficiency". We won't.

Table III—Postexercise electrocardiogram as a predictor of coronary artery stenosis when the disease is present in half the men tested[5]

		≥ 75% stenosis		
		Present	Absent	
Postexercise electrocardiogram	Positive	55 (a)	7 (b)	62
	Negative	49 (c)	84 (d)	133
		104	91	195

Positive predictive value = a/(a + b) = 55/62 = 89%
Negative predictive value = d/(c + d) = 84/133 = 63%
Sensitivity = a/(a + c) = 55/104 = 53%
Specificity = d/(b + d) = 84/91 = 92%
Prevalence = (a + c)/(a + b + c + d) = 104/195 = 53%

standard arteriographic results (a + c)/(a + b + c + d) or 104/195 or 53% of the patients had marked coronary artery stenosis — a highly selected group of patients indeed. What would happen if enthusiasts adopted the multistage stress test for wider use in an effort to detect significant coronary disease in men who want to take up jogging or other sports, regardless of whether they had any chest pain?* Would a positive stress test still be useful?

The results of applying this test to a less carefully selected group of men are entirely predictable (Table IV). If the true prevalence of marked coronary artery stenosis, as assessed by the gold standard of arteriography, was only 1/6 (104/624 or 17%) rather than better than 1/2 (104/195 or 53%), the test's positive predictive value would fall from 89% to 57% and its negative predictive value would rise from 63% to 91% — the reverse of the original situation.†

Now, we said that this result could be forecast from Table III, and it is this forecasting feature that permits a reader to translate the results of a diagnostic test evaluation to his or her own setting. All that are needed are a rough estimate of the prevalence of the disease in one's own practice (from personal experience) or practices like it (from other articles) and some simple arithmetic. For example, as we've charitably estimated for Table IV, approximately one sixth of all men (both symptomatic and asymptomatic) sent for coronary arteriography from a primary care setting might ultimately be found to have coronary artery stenosis. Thus, if we started with the original number of patients with coronary artery disease (104), five times this number (520) would be free of the disease. Because sensitivity remains constant, 55 (53%) of the 104 diseased men would have positive exercise ECGs. Similarly, because specificity remains at 92%, 478 of the 520 nondiseased men would have negative tests. The rest of the table can then be completed by adding or subtracting to fill in the appropriate boxes, and the predictive values and accuracy can then be calculated. In this or any other example, then, the positive predictive value falls and the negative predictive value rises when a diagnostic test developed for patients with a high prevalence of the target disorder is subsequently applied to patients with a lower prevalence of the disorder.

Our analysis derives its relevance from the very real differences in prevalence of various disorders in primary and tertiary care settings. But individual clinicians seldom work at more than one level of specialization and so it might be assumed that a given clinician need not be concerned about the effect of shifts in disease prevalence on his or her interpretation of diagnostic tests. This assumption is quite incorrect, however. We have already mentioned the difference in prevalence among men and women in the same clinical setting. Patients usually have a variety of easily discernible features that permit a fairly precise estimate of the diagnosis before any diagnostic tests are performed. For example, a 30-year-old man with a history of nonanginal chest pain has a low likelihood of coronary artery stenosis (Diamond and Forrester[6] put this likelihood at 5%), whereas a 62-year-old man with typical angina has a very high likelihood of coronary stenosis (94%[6]). When these "pretest likelihoods" or "prevalences" are fed into our diagnostic-test model for exercise electrocardiography, the information provided by this test varies greatly. For the younger man it can be calculated that the likelihood of coronary artery stenosis is 26% if the exercise test is positive (positive predictive value) and 3% if the test is negative (this is the complement of the negative predictive value or d/[c + d]). The exercise test is of little value here: a negative test merely informs us of the obvious (ischemic heart disease is unlikely in this man) and a positive test does not imply a sufficiently high probability of the disease to justify invasive testing under most circumstances.

The exercise test is also not very helpful for the 62-year-old man with typical angina. If the exercise test is positive the likelihood of disease rises only from 94% to 99%. If the test is negative the likelihood falls only to 89%, hardly reassuring enough to forgo further testing.

The important use of the exercise test (or any other test) lies in its application in cases of uncer-

Table IV—Postexercise electrocardiogram as a predictor of coronary artery stenosis when the disease is present in one sixth of the men tested[1]

		≥ 75% stenosis		
		Present	Absent	
Postexercise electrocardiogram	Positive	55 (a)	42 (b)	97
	Negative	49 (c)	478 (d)	527
		104	520	624

Positive predictive value = a/(a + b) = 55/97 = 57%
Negative predictive value = d/(c + d) = 478/527 = 91%
Sensitivity = a/(a + c) = 55/104 = 53% (as in Table III)
Specificity = d/(b + d) = 478/520 = 92% (as in Table III)
Prevalence = (a + c)/(a + b + c + d) = 104/624 = 17%

*The authors of the work cited in this example made no such recommendation.[5]

†This hypothetical case closely approximates what actually happened among women in the study cited here.[5] Roughly one sixth had 75% stenosis or more and the stress test had a sensitivity of 50%, a specificity of 78% (values close to those observed among men), and positive and negative predictive values of 33% and 88% respectively. The authors concluded: "'In women, a positive exercise test is of little value in predicting the presence of significant coronary artery disease, whereas a negative test is quite useful in ruling out the presence of significant disease."

tainty. Let us consider another example, that of a 45-year-old man with atypical angina. Clinical studies demonstrate that such a patient has a 46% likelihood of coronary artery stenosis.[6] Should he go on to angiography or not? If an exercise test is done and is positive, the likelihood of ischemic heart disease can be calculated to be 85%, and he should therefore have an angiogram if clinically warranted. If an exercise test is negative, however, the likelihood of significant coronary stenosis drops to 30% and the need for further investigation diminishes.

Thus, the exercise test is of value, but only for selected patients for whom the likelihood of coronary artery disease is neither high nor low. To act on the results of the exercise test in the last two circumstances makes little sense because it provides little information beyond that already apparent from the clinical presentation.

Having discussed the fourfold comparison with a gold standard, what about the element of "blindness"? This simply means that those who are carrying out or interpreting the results of the diagnostic test should not know whether the patient being tested really does or does not have the disease of interest; that is, they should be "blind" to each patient's true disease status. Similarly, those who are applying the gold standard should not know the diagnostic test result for any patient. It is only when the diagnostic test and gold standard are applied in a blind fashion that we can be assured that conscious or unconscious bias (in this case the "diagnostic suspicion" bias) has been avoided.[7] As you may recall, this bias was discussed in an earlier round on clinical disagreement.[8]

2. *Did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?*

Florid disease (such as long-standing rheumatoid arthritis) usually presents a much smaller diagnostic challenge than the same disease in an early or mild form; the real clinical value of a new diagnostic test often lies in its predictive value among equivocal cases. Moreover, the apparent diagnostic value of some tests actually resides in their ability to detect the manifestations of therapy (such as radiopaque deposits in the buttocks of ancient syphilitics) rather than disease, and the reader must be satisfied that the two are not being confused.

Finally, just as a duck is not often confused with a yak even in the absence of chromosomal analyses, the ability of a diagnostic test to distinguish between disorders not commonly confused in the first place is scant endorsement for its widespread application. Again, the key value of a diagnostic test often lies in its ability to distinguish between otherwise commonly confused disorders, especially when their prognoses or therapies differ sharply. It is this discriminating property that makes the T_4 determination so helpful in sorting out tense, anxious, tremulous and perspiring patients into those with abnormal thyroid function and those with other disorders.

3. *Was the setting for the study, as well as the filter through which study patients passed, adequately described?*

In the previous round we saw how the proportion of hypertensive patients with surgically curable lesions varied almost 10-fold depending on whether the same diagnostic tests were applied in a general practice or in a tertiary care centre. Because a test's predictive value changes with the prevalence of the target disease, the article ought to tell you enough about the study site and patient selection filter to permit you to calculate the diagnostic test's likely predictive value in your own practice.

The selection of control subjects who do not have the disease of interest should be described as well. Although lab technicians and janitors may be appropriate control subjects early in the development of a new diagnostic test (especially with the declining use of medical students as laboratory animals), the definitive comparison with a gold standard demands equal care in the selection of patients with and without the target disease. The reader deserves some assurance that differences in diagnostic test results are due to a mechanism of disease and not simply to differences in such features as age, sex, diet and mobility of case and control subjects.

4. *Was the reproducibility of the test result (precision) and its interpretation (observer variation) determined?*

Validity of a diagnostic test demands both the absence of systematic deviation from the truth (that is, the absence of bias) and the presence of precision (the same test applied to the same unchanged patient must produce the same result). The description of a diagnostic test ought to tell readers how reproducible they can expect the test results to be. This is especially true when expertise is required in performing the test (for example, ultrasonography currently has enormous variation in the quality of its results when performed by different operators) or in interpreting it (as you may recall from an earlier round, observer variation is a major problem for tests involving x-rays, electrocardiography and the like).[9]

5. *Was the term "normal" defined sensibly?*

If the article uses the word "normal" its authors should tell you what they mean by it. Moreover, you should satisfy yourself that their definition is clinically sensible. Several different definitions of normal are used in clinical medicine; we contend that some of them probably lead to more harm than good. We have listed six definitions of normal in Table V and acknowledge our debt to Tony Murphy for pointing out most of them.[2,10]

Perhaps the most common definition of normal assumes that the diagnostic test results (or some arithmetic manipulation of them) for everyone, for a group of presumably normal people or for a carefully characterized "reference" population will fit a specific theoretical distribution known as the normal or gaussian distribution. One of the nice properties of the gaussian

distribution is that its mean ± two standard deviations (SDs) encloses 95% of its contents, leaving 2.5% at each of its upper and lower ends. Thus, the "mean ± 2 SDs" became a tempting way to define normal and came into general use.

It's too bad 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 nondisease.[11] First, diagnostic test results simply do not fit the gaussian distribution. (Actually, we should be grateful that they don't; the gaussian distribution extends to infinity in both directions, necessitating occasional patients with impossibly high hemoglobin concentrations 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 diseases have 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*. Recognizing the failure of diagnostic test results to fit a theoretical distribution such as the gaussian, many laboratories have suggested that we ignore the shape of the distribution and simply refer, for example, to the lower 95% of test results as normal. Although the percentile definition does avoid the problem of negative test values, it still leads to the conclusion that all diseases are of equal prevalence and still contributes to the "upper limit syndrome" of nondisease because its use means that the only "normal" patients are the ones who are not yet sufficiently worked up.[2]

This inevitable consequence arises as follows: if the normal range includes the lower 95% of diagnostic test results, then the likelihood that a given patient will be called normal when subjected to this test is 0.95 (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 that the patient will be called normal is now 0.95 × 0.95 = 0.90. Indeed, the likelihood of a patient's being called normal is 0.95 raised to the power of the number of independent diagnostic tests performed. Thus, a patient who undergoes 20 tests has only 0.95^{20} or about 1 chance in 3 of being called normal; a patient undergoing 100 such tests has only about 6 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* approach is based upon studies of precursors or statistical predictors of subsequent clinical events; by this definition, the normal range for serum cholesterol concentration or blood pressure consists of 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, it has been pointed out that the normal serum cholesterol concentration, by this definition, might lie below 150 mg/dl (3.9 mmol/l).[13] Another shortcoming of this risk factor definition becomes apparent when we examine the consequences of acting upon a test result that lies beyond the normal range: Will altering a risk factor really change the risk? Recent experience with the treatment of "abnormal" serum cholesterol levels with clofibrate (in which mortality went up, not down, with treatment) underscores the danger of this assumption.[14]

A related approach defines normal as that which is *culturally desirable*, providing an opportunity for what Mencken[15] called "the corruption of medicine by morality" through the "confusion of the theory of the healthy with the theory of the virtuous". One sees such definitions in their benign form at the fringes of the current lifestyle movement (e.g., "It is better to be slim than fat" and "Exercise and fitness are better than sedentary living and lack of fitness"[16]), and in their malignant form in the health care 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. Mencken[15] offered a similarly pungent point of view on the latter: "The true aim of medicine is not to make men virtuous; it is to safeguard and rescue them from the consequences of their vices."

Two final definitions are of much greater utility to the clinician because they focus directly on the clinical acts of diagnosis and therapy.[1] The *diagnostic* definition identifies a range of diagnostic test results beyond which a specific disease is, with known probability, present. It is this definition that is

*This consequence of such definitions helps explain the results of a randomized trial of multitest screening at the time of admission to hospital that found no patient benefits but increased health care costs.[12]

Table V—Properties and consequences of different definitions of "normal"

Property	Term	Consequences of its clinical application
The distribution of diagnostic test results has a certain shape	Gaussian	Ought to occasionally obtain minus values for hemoglobin level etc. All diseases have the same prevalence. Patients are normal only until they are assessed.
Lies within a preset percentile of previous diagnostic test results	Percentile	All diseases have the same prevalence. Patients are normal only until they are assessed.
Carries no additional risk of morbidity or mortality	Risk factor	Assumes that altering a risk factor alters risk.
Socially or politically aspired to	Culturally desirable	Confusion over the role of medicine in society.
Range of test results beyond which a specific disease is, with known probability, present or absent	Diagnostic	Need to know predictive values for your practice.
Range of test results beyond which therapy does more good than harm	Therapeutic	Need to keep up with new knowledge about therapy.

used in the first guide to reading about a diagnostic test: comparison with a gold standard. The "known probability" with which a disease is present is our old friend the positive predictive value.

This definition is illustrated in Fig. 2, where we see the usual overlap in diagnostic test results between patients shown, by application of a gold standard, to be disease-free or diseased (the a, b, c and d in Fig. 2 correspond to cells a, b, c and d of Tables II to IV). The known probability (or predictive value) with which a disease is present or absent depends on where we set the limits for the normal range of diagnostic test results. If we simply wanted to maximize the number of times the diagnostic test result was correct, we'd set the limits for normal at the dotted line where the curves cross, but that might not be very helpful clinically. If we lower these normal limits to point X, cell c approaches zero, sensitivity and negative predictive values approach 100% and we can use the normal diagnostic test result to rule out the disease (because nobody with the disease has test results below X). Similarly, if we raise the limits of normal for the diagnostic test result to point Y, cell b approaches zero, specificity and positive predictive values approach 100% and we can use the abnormal diagnostic test result to rule in the disease (because no nondiseased patients have test results above Y). Thus, this definition has clinical utility and is a distinct improvement over the definitions described earlier. However, it does require that clinicians keep track of both the predictive values of individual diagnostic tests and the test levels at points X and Y that apply in their own practices.

The final definition of normal sets its limits at the point beyond which specific treatments have been shown to do more good than harm, and is indicated in Fig. 2 as point Z. This *therapeutic* definition is attractive because of its link to action. The therapeutic definition of the normal range of blood pressure, for example, avoids the hazards of labelling patients as diseased[17] unless they are going to be treated. The use of this definition requires that clinicians keep abreast of advances in therapeutics and become adept at sorting out therapeutic claims; a later article in this series of Clinical Epidemiology Rounds is devoted to this topic.

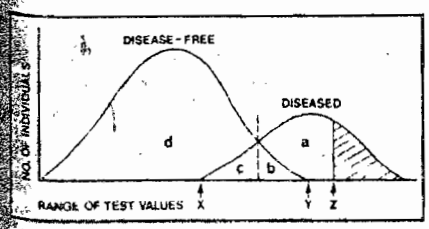

FIG. 2—Diagnostic and therapeutic definitions of "normal".

When reading a report of a new diagnostic test, then, you should satisfy yourself that the authors have defined what they mean by normal and that they have done so in a sensible and clinically useful fashion.

6. *If the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined?*

In many conditions an individual diagnostic test examines but one of several manifestations of the underlying disorder. For example, in diagnosing deep vein thrombosis impedance plethysmography examines venous emptying, whereas leg scanning with iodine-125-labelled fibrinogen examines the turnover of coagulation factors at the site of thrombosis.[18] Furthermore, plethysmography is much more sensitive for proximal than distal venous thrombosis, whereas the reverse is true for leg scanning. As a result, these tests are best applied in sequence: if the plethysmogram is positive, the diagnosis is made and treatment begins at once; if it is negative, leg scanning begins and the diagnostic and treatment decisions await its results.

This being so, it is clinically nonsensical to base a judgement of the value of leg scanning on a simple comparison of its results alone against the gold standard of venography. Rather, its agreement with venography among suitably symptomatic patients with a negative impedance plethysmogram is one appropriate assessment of its validity and clinical usefulness. Another valid assessment would be the agreement of results of the combination of leg scanning and impedance plethysmography with venography.

In summary, any single component of a cluster of diagnostic tests should be evaluated in the context of its clinical use.

7. *Were the tactics for carrying out the test described in sufficient detail to permit their exact replication?*

If the authors have concluded that you should use their diagnostic test, they have to tell you how to use it; this description should cover patient issues as well as the mechanics of performing the test and interpreting its results. Are there special requirements for fluids, diet or physical activity? What drugs should be avoided? How painful is the procedure and what is done to relieve any pain? What precautions should be taken during and after the test? How should the specimen be transported and stored for later analysis? These tactics and precautions must be described if you and your patients are to benefit from this diagnostic test.

8. *Was the "utility" of the test determined?*

The ultimate criterion for a diagnostic test or any other clinical maneuver is whether the patient is better off for it. If you agree with this point of view you should scrutinize the article to see whether the authors went beyond the foregoing issues of accuracy, precision and the like to explore the long-term consequences of their use of the diagnostic test.

In addition to telling you what happened to patients correctly classified by the diagnostic test, the authors should describe the fate of the patients who had false-positive results (those with positive test results who really did not have the disease) and those with false-negative results (those with negative test results who really did have the disease). Moreover, when the execution of a test requires a delay in the initiation of definitive therapy (while the procedure is being rescheduled, the test tubes are incubating or the slides are waiting

to be read) the consequences of this delay should be described.

For example, we are part of a team that has studied the value of noninvasive tests in the diagnosis of patients with clinically suspected deep leg vein thrombosis, and have tested the policy of withholding anticoagulants from patients with a negative impedance plethysmogram (a quick test) until or unless the ^{125}I-fibrinogen leg scan becomes positive.[18] The scan takes several hours to several days to become positive when venous thrombi are small or confined to the calf; it is therefore important to determine and report whether any patients suffer clinical embolic events during this interval (fortunately, they do not). Moreover, comparisons of these investigations against the gold standard of venography have included documentation of the consequences of treating patients with false-positive results and withholding treatment from those with false-negative results. The resemblance of this approach to the "therapeutic" definition of normalcy is worth noting.*

Use of these guides to reading

By applying the foregoing guides you should be able to decide if a diagnostic test will be useful in your practice, if it won't or if it still hasn't been properly evaluated. Depending on the context in which you are reading about the test, one or another of the eight guides will be the most important one and you can go right to it. If it has been met in a credible way, you can go on to the others; if the most important guide hasn't been met you can discard the article right there and go on to something else. Thus, once again, you can improve the efficiency with which you use your scarce reading time. When trying to pick the best test from an array of competing diagnostic tests you could carry out on a given patient, these guides will help you compare them with each other. On the basis of this comparison you can pick the one that will best meet your clinical requirements.

The next round will consider articles that describe the clinical course and prognosis of disease.

*In this regard, we think it's a shame that the term "diagnostic efficacy" has crept into the literature, especially since it is used as a synonym for accuracy rather than utility.

References

1. SACKETT DL: Clinical diagnosis and the clinical laboratory. *Clin Invest Med* 1978; 1: 37–43
2. MURPHY EA: *The Logic of Medicine*, Johns Hopkins, Baltimore, 1976: 117–160
3. RANSOHOFF DF, FEINSTEIN AR: Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978; 299: 926–930
4. GALEN RS, GAMBINO SR: *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*, Wiley, New York, 1975: 30–40
5. SKETCH MH, MOHIUDDIN SM, LYNCH JD, ZENCKA AE, RUNCO V: Significant sex differences in the correlation of electrocardiographic exercise testing and coronary arteriograms. *Am J Cardiol* 1975; 36: 169–173
6. DIAMOND GA, FORRESTER JS: Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *N Engl J Med* 1979; 300: 1350–1358
7. SACKETT DL: Bias in analytic research. *J Chronic Dis* 1979; 32: 51–63
8. Department and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.: Clinical disagreement: II. How to avoid it and how to learn from one's mistakes. *Can Med Assoc J* 1980; 123: 613–617
9. Idem: Clinical disagreement: I. How often it occurs and why. Ibid: 499–504
10. MURPHY EA: The normal, and the perils of the sylleptic argument. *Perspect Biol Med* 1972; 15: 566–582
11. MEADOR CK: The art and science of nondisease. *N Engl J Med* 1965; 272: 92–95
12. DURBRIDGE TC, EDWARDS F, EDWARDS RG, ATKINSON M: An evaluation of multiphasic screening on admission to hospital. Precis of a report to the National Health and Medical Research Council. *Med J Aust* 1976; 1: 703–705
13. KANNEL WB, DAWBER TR, GLENNON WE, THORNE MC: Preliminary report: the determinants and clinical significance of serum cholesterol. *Mass J Med Technol* 1962; 4: 11–29
14. OLIVER MF, HEADY JA, MORRIS JN, COOPER J: A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. Report from the Committee of Principal Investigators. *Br Heart J* 1978; 40: 1069–1118
15. MENCKEN HL: A *Mencken Chrestomathy*, Knopf, Westminster, 1949
16. LALONDE M: *A New Perspective on the Health of Canadians. A Working Document*, Department of National Health and Welfare, Ottawa, Apr 1974: 58
17. HAYNES RB, SACKETT DL, TAYLOR DW, GIBSON ES, JOHNSON AL: Increased absenteeism from work following the detection and labelling of hypertensives. *N Engl J Med* 1978; 299: 741–744
18. HULL R, HIRSH J, SACKETT DL, POWERS P, TURPIE AGG, WALKER I: Combined use of leg scanning and impedance plethysmography in suspected venous thrombosis. An alternative to venography. *N Engl J Med* 1977; 296: 1497–1500

BOOKS

continued from page 697

THE HEALTHY HYPOCHONDRIAC. Recognizing, Understanding and Living with Anxieties about our Health. Richard Ehrlich. 211 pp. W.B. Saunders Company Canada, Ltd., Toronto, 1980. $14.75 (Can.), clothbound; $8.50 (Can.), paperbound. ISBN 0-7216-334-X, clothbound; ISBN 0-7216-333-1, paperbound

THE HOSPITAL CARE OF CHILDREN. A Review of Contemporary Issues. Geoffrey C. Robinson and Heather F. Clarke. 270 pp. Illust. Oxford University Press, Toronto, 1980. $25.25. ISBN 0-19-502673-X

THE HYPOTHALAMO-PITUITARY CONTROL OF THE OVARY. Volume 2. J.S.M. Hutchinson. 215 pp. Eden Press, Westmount, PQ, 1980. $28. ISBN 0-88831-091-9

AN INTRODUCTION TO HUMAN BIOCHEMISTRY. C.A. Pasternak. 271 pp. Illust. Oxford University Press, Toronto, 1979. $20.50, paperbound. ISBN 0-19-261127-5

LANGUAGE AND COMMUNICATION IN THE ELDERLY. Clinical, Therapeutic and Experimental Issues. Edited by Loraine K. Obler and Martin L. Albert. 220 pp. Illust. Lexington Books, Lexington, Massachusetts; D.C. Heath Canada Ltd., Toronto, 1980. $25.95. ISBN 0-669-03868-7

MANAGING HEALTH SYSTEMS IN DEVELOPING AREAS. Experiences from Afghanistan. Edited by Ronald W. O'Conner. 314 pp. Illust. Lexington Books, Lexington, Massachusetts; D.C. Heath Canada Ltd., Toronto, 1980. $30.50. ISBN 0-669-03646-3

MANUAL OF CLINICAL PROBLEMS IN ONCOLOGY WITH ANNOTATED KEY REFERENCES. Carol S. Portlock and Donald R. Goffinet. 323 pp. Little, Brown and Company (Inc.), Boston, 1980. Price not stated, spiralbound. ISBN 0-316-71424-0

continued on page 751