# 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.

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

## Guides for reading articles about diagnostic tests

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

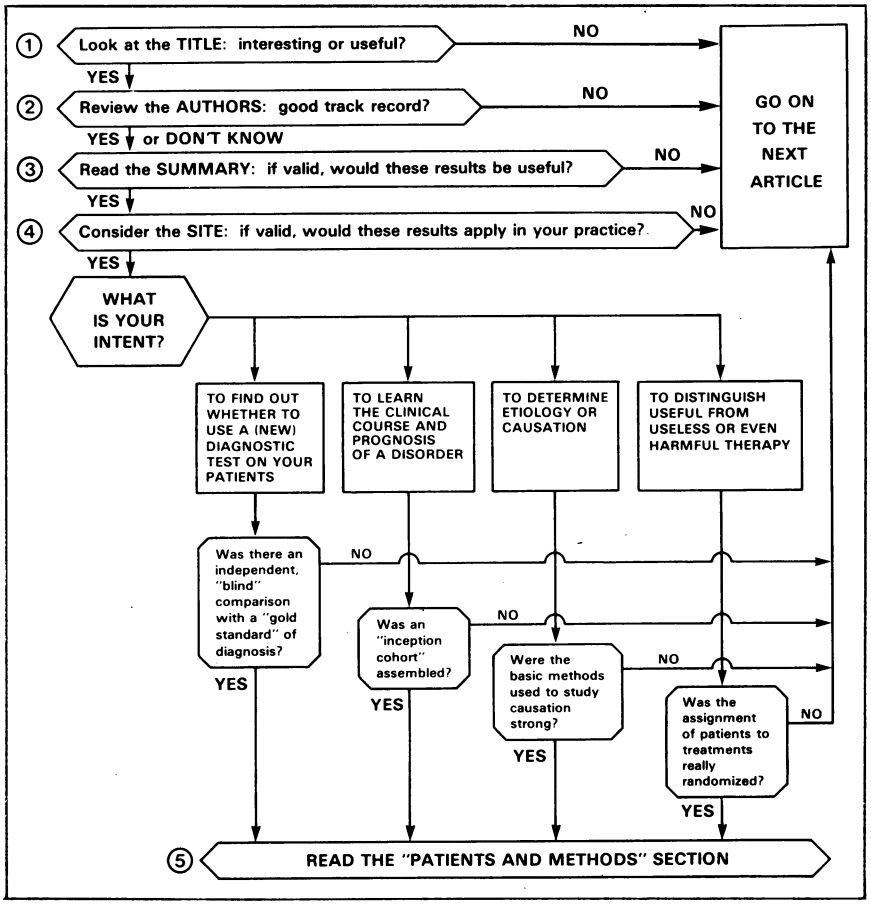


**FIG. 1—The first steps in how to read articles in a clinical journal.**

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 longstanding 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.

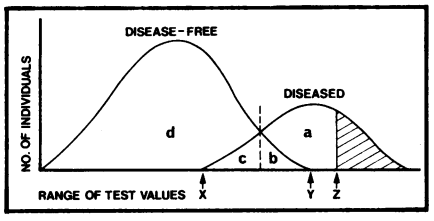


**FIG. 2—Diagnostic and therapeutic definitions of "normal".**

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*

# Clinical Epidemiology Rounds

## Interpretation of diagnostic data: 1. How to do it with pictures

DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS, MCMASTER UNIVERSITY, HAMILTON, ONT.

This series of Clinical Epidemiology Rounds will focus on strategies and tactics for interpreting diagnostic data. Although most explanations of how to interpret such data confine themselves to reports from paraclinical services, the same principles apply to items of the clinical history and findings in the physical examination (with which the vast majority of diagnoses are confirmed). Moreover, elements of proper history-taking and physical examination regularly generate more powerful tests of diagnostic hypotheses than do the results of laboratory testing. Thus, although we shall repeatedly use a laboratory test, measurement of the serum creatine kinase (CK) level, to illustrate the different ways of interpreting diagnostic data, we shall also include several examples of findings in the history and the physical examination to remind readers of the wider applicability of the principles of interpreting diagnostic data.

Much has been written about the interpretation of diagnostic tests, and it is very easy to drown in the associated graphs, tables, curves, formulae and technical jargon. We have tried to remove the unessential, pedantic bits and synthesize the rest into bite-sized chunks, each of which is self-contained, is digestible at a single sitting and has some features that can be put to good clinical use. Part 1, for example, is simple but powerful, and is done entirely with pictures, not computations and new technical terms. The other chunks become progressively complex and require learning some new terminology and performing computations, ultimately with a hand calculator.

In an earlier set of Rounds, one of the sections was on reading clinical journals to find out more about a diagnostic test.[1] You may find it useful as parallel reading, but you should go through at least part 1 of this series before turning to it.

### Doing it with pictures

*Case presentation*

At the Royal Infirmary in Edinburgh all patients under 70 years of age who are suspected of having had a myocardial infarction within the previous 48 hours are admitted to the coronary care unit (CCU). Like most such units, this one became crowded shortly after it opened in 1966, and those in charge recognized the need to differentiate, as quickly as possible, between the patients who had actually had a myocardial infarction and therefore ought to stay in the unit and those who had not and could therefore be transferred.[2]

The CCU staff thought that checking for rises in the serum CK level might help them diagnose a myocardial infarction sooner than studying enzymes whose levels rise later, such as aspartate aminotransferase. Accordingly, they measured the serum CK level at the time of admission to the CCU and the next two mornings for 360 consecutive patients who lived long enough to have blood samples taken.

A clinician who was "blind" to the serum CK measurements reviewed the electrocardiograms (ECGs), clinical records and autopsy reports for all 360 patients. Patients with pathologic Q waves, ST-segment elevation and subsequent T-wave inversions or positive findings on autopsy were considered to have "very probable" infarcts; those with less diagnostic ECG changes (usually just ST-segment and T-wave abnormalities) were considered to have "possible" infarcts. These two groups totalled 230 patients; the remaining 130 patients were judged not to have myocardial infarcts.

*Comments*

The highest serum CK levels for the 230 patients who had infarcts and the 130 patients who did not are given in Fig. 1.* At the upper end of the scale are maximum levels of 480 U/l or higher, recorded for 35 patients who had an infarct but none of those who did not. At the lower end are maximum levels of less than 40 U/l, recorded for only 2 patients who had an infarct but for 88 of those who did not.

*Since laboratory methods vary, this distribution of serum CK levels may not apply at your institution.

Reprint requests to: Dr. David L. Sackett, Rm. 3V43E, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton, Ont. L8N 3Z5

This is the first of six articles that focus on the strategies and tactics for interpreting diagnostic data, both the clinical data from the history and physical examination and the paraclinical data from the clinical laboratories, the radiology department and the surgical pathology service. The remaining articles will appear in the next five issues of the Journal.

### Conversion of numbers to pictures

Although we could proceed directly to a discussion of diagnostic levels of serum CK, let's convert these sets of numbers to pictures that will be easier to follow. First, converting these numbers to percentages will make them easier to compare. The 35 patients with an infarct whose maximum serum CK levels were 480 U/l or higher constituted 15% of all the patients with an infarct, whereas the 88 patients without an infarct whose maximum levels were below 40 U/l constituted 67% of all the patients without an infarct.

Because most people find graphs and other pictures easier to deal with than tables of numbers, we have gone one step further by illustrating the percentages in a bar chart (Fig. 2). Now the percentages are easier to comprehend. The serum CK levels for the patients with an infarct peak once between 80 and 160 U/l (forming the "possible" infarct group) and again beyond 480 U/l (forming the "very probable" infarct group). On the right side of the scale the serum CK levels for the patients without an infarct are clustered below 40 U/l.

Unfortunately, and like most clinical data, there is a fair degree of overlap: maximum serum CK levels of less than 40 U/l all the way up to 320 U/l were found in both patients with and patients without infarcts. None the less, can we take advantage of the fact that *only* those with infarcts had maximum levels of 320 U/l or more? Yes. We could set a dividing or cut-off line between the normal and abnormal serum CK levels at 320 U/l (Fig. 3). By setting the cut-off line at 320 U/l (labelled y) we can be confident that *every* patient (in the absence of skeletal muscle damage from, for example, resuscitation or intramuscular injection) who has a serum CK level of 320 U/l or higher has an infarct. Thus, when the level is 320 U/l or higher we can "rule in" an infarct.

Have we gained much by selecting a cut-off point above which only patients with an infarct will appear?

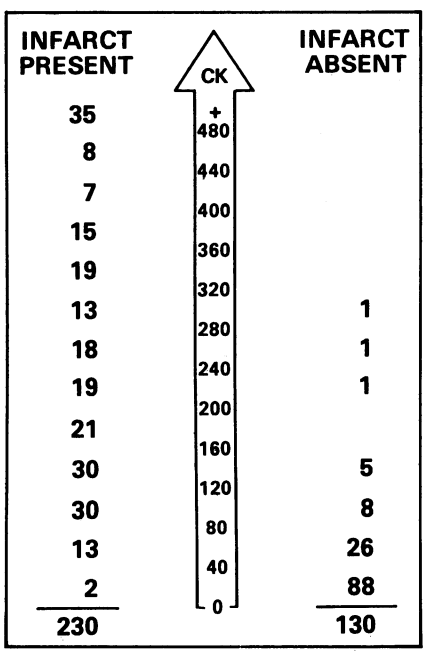


FIG. 1—Maximum serum creatine kinase (CK) levels (U/l) for patients with and without myocardial infarcts.

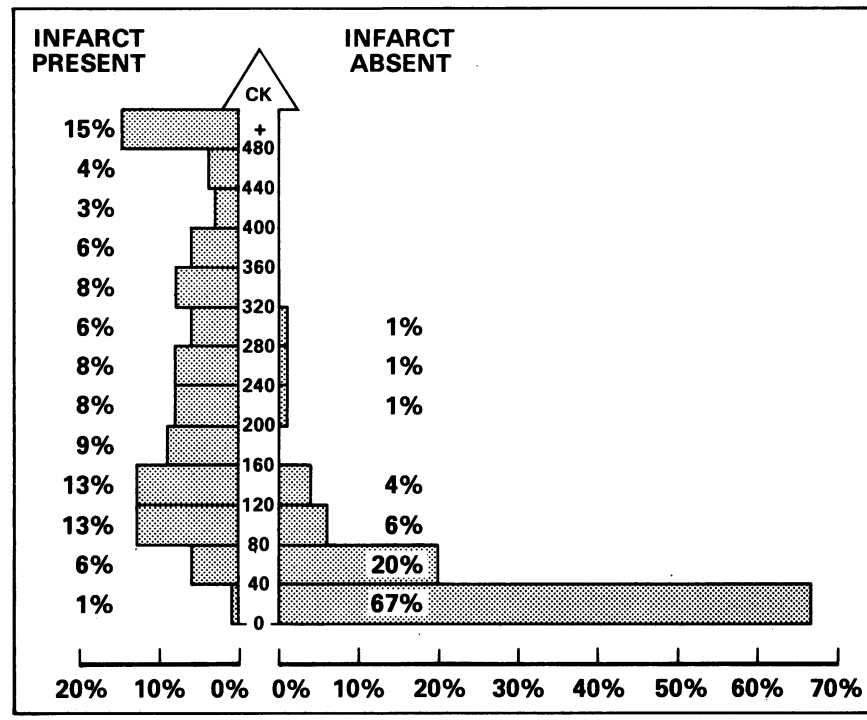


FIG. 2—Bar chart of maximum serum CK levels for patients with and without myocardial infarcts.

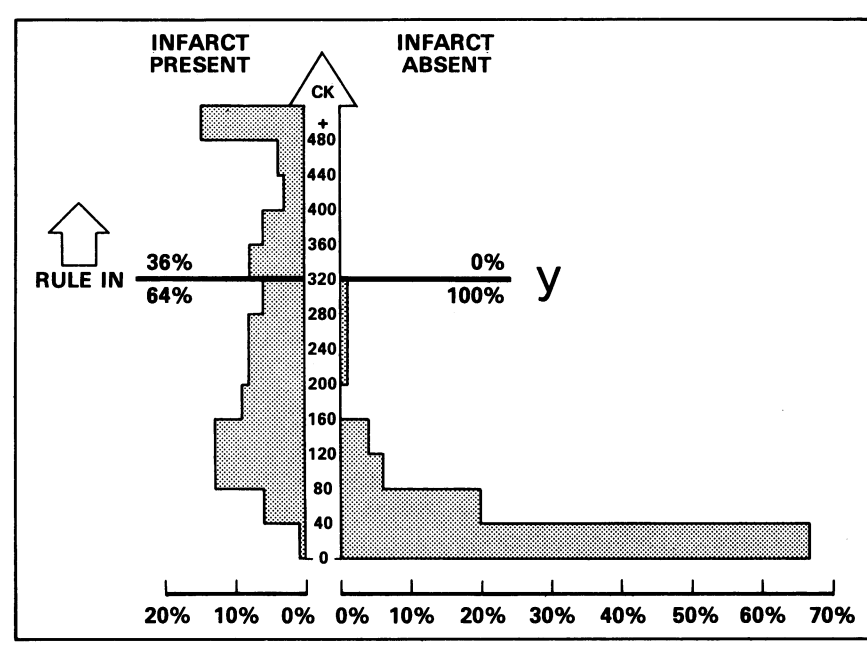


FIG. 3—Absolute (100%) cut-off point (y) of maximum serum CK level for diagnosis of myocardial infarct.

Yes and no. On the one hand, we can confidently diagnose a myocardial infarction in at least some patients and treat them accordingly. On the other hand, the cut-off point captures only about one third (36%) of the 230 patients with infarcts and cannot distinguish the remaining 64% from all those without infarcts. Thus, of the 360 patients admitted to the CCU (230 with and 130 without infarcts) we have achieved a firm diagnosis in (35 + 8 + 7 + 15 + 19)/360 (Fig. 1), or 23%.

Another problem with this "absolute" cut-off point becomes apparent when we think about the *next* 360 patients admitted to the unit and the 360 admitted after them. Will they generate an identical range of serum CK levels? Almost, but not quite. Subsequent groups of patients will, on average, show the same distribution as our group. However, the *range* of levels can only become *larger*, never smaller, as more patients are examined. The more patients we encounter, the more likely we are to find patients who, although they have neither a myocardial infarction nor skeletal muscle damage, have higher and higher serum CK levels. Such patients will be rare, constituting no more than 1% or, at most, 2% of the total, but the more patients we see, the greater our chances will be of encountering these "outliers". As a result, if we stick with our "absolute" definition of the "rule-in" cut-off point y, we shall be forced to move the cut-off point to ever more extreme levels; inevitably, this cut-off point will help us make a diagnosis in an ever-shrinking proportion of the patients we encounter.

Is there a way out of this dilemma? Yes, if we relax our cut-off point just a little bit. Bearing in mind that extreme values will be rare, what would happen if we revised our cut-off point so that it excluded 99%, rather than 100%, of patients with or without an infarct (Fig. 4)? The "rule-in" cut-off point y, by allowing in the highest 1% of patients who do not have an infarct, lowers the cut-off point to 280 U/l and encompasses 42% rather than just 36% of all the patients with an infarct. Moreover, a second cut-off point (x) allows in the lowest 1% of patients with an infarct and can be used to rule out myocardial infarction when the maximum serum CK level remains below 40 U/l. Below this cut-off point are only 1% of the patients with infarcts but 67% of those without, so we have gained a great deal. Now we can rule in or rule out myocardial infarction in (35 + 8 + 7 + 15 + 19 + 13 + 88)/360 (Fig. 1), or 51% of the 360 patients at the cost of missing only two patients with an infarct (whose maximum serum CK level was less than 40 U/l) and of diagnosing myocardial infarction in one patient who did not have an infarct (but had a serum CK level of 280 U/l or higher). Thus, these simple rule-in and rule-out levels could handle more than half the patients who presented with clinically suspected myocardial infarction.

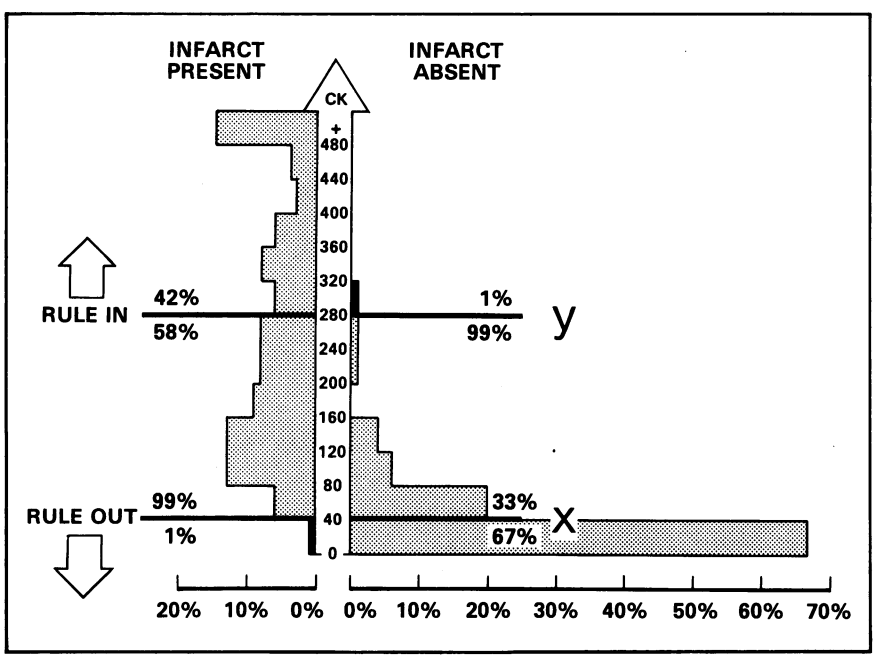


**FIG. 4—Revised (99%) cut-off points of maximum and minimum serum CK levels for diagnosis of myocardial infarct.**

There is nothing sacred about choosing a 99% cut-off point; it could be 98%, 96% or 99.5% if you wished. The major factor in deciding where to set the cut-off point should be your clinical judgement about what is best for the patients. For example, if the treatment for a disorder is innocuous, and if overdiagnosis does not produce shame or anguish in patients who are falsely labelled, you might want to relax the rule-in cut-off point y to exclude only 95% of patients who do not have the target disorder. On the other hand, if patients really suffer from being labelled with the target disorder (e.g., cancer, venereal disease or schizophrenia) the rule-in cut-off point y should be set very high so that it excludes 99% or 99.5% of those without the disorder. Similarly, if early diagnosis and therapy are essential for a satisfactory clinical outcome, as in many neonatal screening programs (e.g., for phenylketonuria and neonatal hypothyroidism) you would want a very low rule-out cut-off point x so that it captures all or almost all of the patients with the target disorder (this is why so many babies who have positive results of screening for these disorders are found on further testing to be normal; we do not want to miss any cases).

Let us briefly consider how these rule-in and rule-out cut-off points fit into commonly used diagnostic strategies.[3] For example, we might be using *pattern recognition* (instantaneous recognition that the patient's face, skin, gait, sound or smell conforms to an identifiable picture or pattern of disease). In this case we use the rule-in cut-off point y, for we are seeking confirmation that our "snap" diagnosis was correct. On the other hand, we might be using the *hypotheticodeductive* approach (formulation, from the earliest clues about the patient, of a "short list" of potential diagnoses or actions, followed by performance of

the clinical and paraclinical maneuvers that will best reduce the length of the list). We will proceed most efficiently by paying attention to the rule-out cut-off point x, for this will help us eliminate hypotheses from our list and concentrate on the most likely diagnoses.

## Summary

- Pick symptoms, signs or tests with which the overlap between patients who do and do not have the target disorder will be as small as possible. If you do not know which symptoms, signs or tests these are, ask clinicians, radiologists or laboratory workers who are expert in the relevant subspecialty.
- See if you can find pictures like Figs. 3 and 4; if not, you should be able to construct them from raw data such as those in Fig. 1 (just follow the steps we carried out in generating the later figures). The patient groups from which the pictures are generated should not just be persons with "classic cases" on the one hand and robust, young medical students on the other; most of us can tell the difference between these groups without any diagnostic tests. The patient groups should be those in whom the target disorder is a legitimate entry in the list of hypotheses. For more on this key issue, check our earlier Rounds on how to read clinical journals.[1]
- Decide whether you want to set strict (99%) or loose (95%) cut-off points for your rule-in and rule-out levels.
- Add the cut-off points to the pictures and identify the diagnostic test results that correspond to the rule-in (y) and rule-out (x) cut-off points.
- For fun and continuing education, add to the clinical "track-record" log we described in an earlier paper, on avoiding clinical disagreement and learning from one's mistakes.[4] Enter the rule-in and rule-out cut-off points as you generate them, and then keep track of how they work for you. Do your rule-in cut-off points agree with your diagnoses when they are achieved by pattern recognition? For what proportion of patients in whom you suspect a target disorder can these cut-off points help in making a decisive diagnosis?

Straightforward though it may be, you cannot help but have noticed that "doing it with pictures" leaves a large number of patients in the middle, with too much of a symptom, sign or laboratory result to rule out the target disorder but not enough of it to rule it in. Our next paper will show you how to diagnose a disorder in all the patients by means of a simple table.

## References


1. Department of clinical epidemiology and biostatistics, McMaster University Health Sciences Centre: How to read clinical journals: II. To learn about a diagnostic test. *Can Med Assoc J* 1981; 124: 703–710
2. SMITH AF: Diagnostic value of serum-creatine-kinase in a coronary care unit. *Lancet* 1967; 2: 178–182
3. SACKETT DL: Clinical diagnosis and the clinical laboratory. *Clin Invest Med* 1978; 1: 37–43
4. Department of 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


*VASODILATORS*
*continued from page 428*


42. PACKER M, FRISHMAN WH: Verapamil therapy for stable and unstable angina pectoris: calcium channel antagonists in perspective. *Am J Cardiol* 1982; 50: 881–885
43. SUBRAMANIAN B, BOWLES MJ, DAVIES AB, RAFTERY EB: Combined therapy with verapamil and propranolol in chronic stable angina. *Am J Cardiol* 1982; 49: 125–132
44. CHATTERJEE K, ROULEAU JL, PARMLEY WW: Haemodynamic and myocardial metabolic effects of captopril in chronic heart failure. *Br Heart J* 1982; 47: 233–238
45. ROULEAU JL, CHATTERJEE K, BENGE W, PARMLEY WW, HIRAMATSU B: Alterations in left ventricular function and coronary hemodynamics with captopril, hydralazine and prazosin in chronic ischemic heart failure: a comparative study. *Circulation* 1982; 65: 671–678
46. NAKASHIMA Y, FOUAD FM, TARAZI RC: Long-term captopril therapy in congestive heart failure: serial hemodynamic and echocardiographic changes. *Am Heart J* 1982; 104: 827–833
47. SLUTSKY R: Hemodynamic effects of inhaled terbutaline in congestive heart failure patients without lung disease: beneficial cardiotonic and vasodilator beta-agonist properties evaluated by ventricular catheterization and radionuclide angiography. *Am Heart J* 1981; 101: 556–560
48. CODY RJ, FRANKLIN KW, KLUGER J, LARAGH JH: Sympathetic responsiveness and plasma norepinephrine during therapy of chronic congestive heart failure with captopril. *Am J Med* 1982; 72: 791–797
49. THADANI U, MANYARI D, PARKER JO, FUNG HL: Tolerance to the circulatory effects of oral isosorbide dinitrate. Rate of development and cross-tolerance to glyceryl trinitrate. *Circulation* 1980; 61: 526–535


**Working hypotheses**

*We have multitudes of facts, but we require, as they accumulate, organisations of them into higher knowledge; we require generalisations and working hypotheses.*

*—Hughlings Jackson (1835–1911)*

# Clinical Epidemiology Rounds

## Interpretation of diagnostic data: 2. How to do it with a simple table (part A)

DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS, MCMASTER UNIVERSITY, HAMILTON, ONT.

In part 1 of our series on interpreting diagnostic data we showed how you can "do it with pictures". In parts 2 and 3 you can learn how to do it with a simple two-by-two, or fourfold, table. To ease the transition, we will use the same sample of patients as in part 1.

### Doing it with a simple table

*Case presentation*

In 360 patients consecutively admitted to a coronary care unit (CCU) blood was drawn for measurement of the serum creatine kinase (CK) level at the time of admission and the next two mornings.[1] A clinician who was "blind" to the measurements reviewed the patients' electrocardiograms (ECGs), clinical records and autopsy reports, and decided that 230 had had a myocardial infarct and 130 had not.

*Comment*

Fig. 1 shows the maximum serum CK levels* for the 230 patients who did have an infarct and the 130 patients who did not; in Fig. 2 the numbers are converted to percentages.

### Conversion of numbers to a simple table

In Fig. 3 we have swung the 230 patients with myocardial infarcts over to the same side as the 130 patients without myocardial infarcts. Note that the two lines cross

*Since laboratory methods differ, this distribution of serum levels may not apply at your institution.

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

**This is the second of six articles (the first appeared in the Sept. 1, 1983 issue of *CMAJ* on pages 429 to 432) that focus on the strategies and tactics for interpreting diagnostic data, both the clinical data from the history and physical examination and the paraclinical data from the clinical laboratories, the radiology department and the surgical pathology service. The remaining articles will appear in the next four issues of the Journal.**

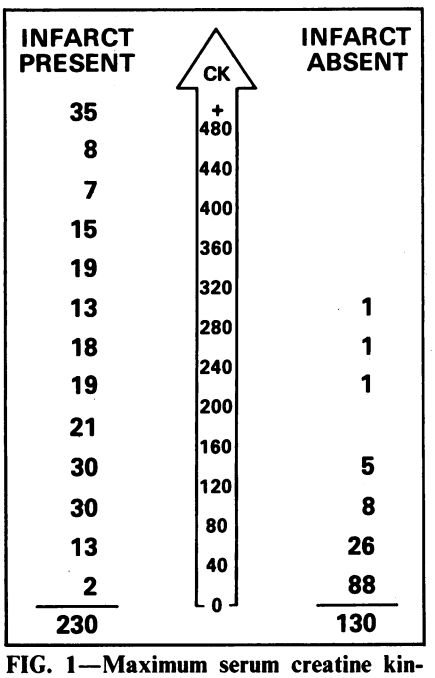


**FIG. 1—Maximum serum creatine kinase (CK) levels (U/l) in patients with and without myocardial infarcts.**

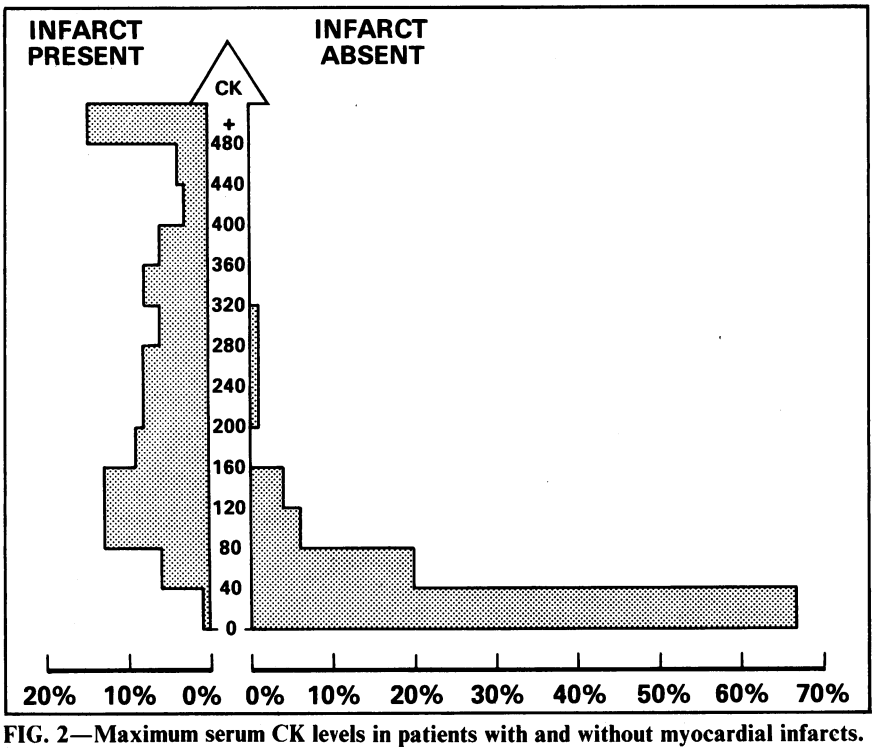


**FIG. 2—Maximum serum CK levels in patients with and without myocardial infarcts.**

at 80 U/l. By selecting this value as the cut-off point, we can then move the patients with infarcts back where they were (Fig. 4). By placing the cut-off point at 80 U/l, above which the CK test result is decreed positive and below which it is decreed negative, we have created four groups of patients:

a: those with infarcts whose test results were positive.

b: those without infarcts whose test results were positive.

c: those with infarcts whose test results were negative.

d: those without infarcts whose test results were negative.

The CK test correctly classified the patients in groups a and d. Group a's positive test results were telling the truth and are therefore called "true positives". Group d's negative test results were also telling the truth and are therefore called "true negatives". However, the test results were false for the patients in groups b and c. Group b's test results were positive, but the patients did not have infarcts; thus, these results are called "false positives". Group c's test results were negative, but the patients had infarcts; accordingly, these results are called "false negatives".

These relationships can be shown in a simple two-by-two, or fourfold, table (Table I). Table I corresponds exactly to Fig. 4: patients with infarcts are on the left, and those without infarcts are on the right. The upper half of the table shows the positive CK test results and the lower half the negative results. The four inner squares, called "cells", are labelled a, b, c and d to correspond with the four groups of patients.*

Once you are comfortable about the table you can put numbers in it. Add the 80-U/l cut-off point to Fig. 1 and then place the numbers in the appropriate cells in the table (Table II). Among the patients who did have infarcts, those whose serum CK levels were above the cut-off point (true positives) belong in cell a, and those whose levels were below the cut-off point (false negatives) belong in cell c. On the other side, among the patients who did not have infarcts, those whose levels were above the cut-off point (false positives) belong in cell b, and those whose levels were below the cut-off

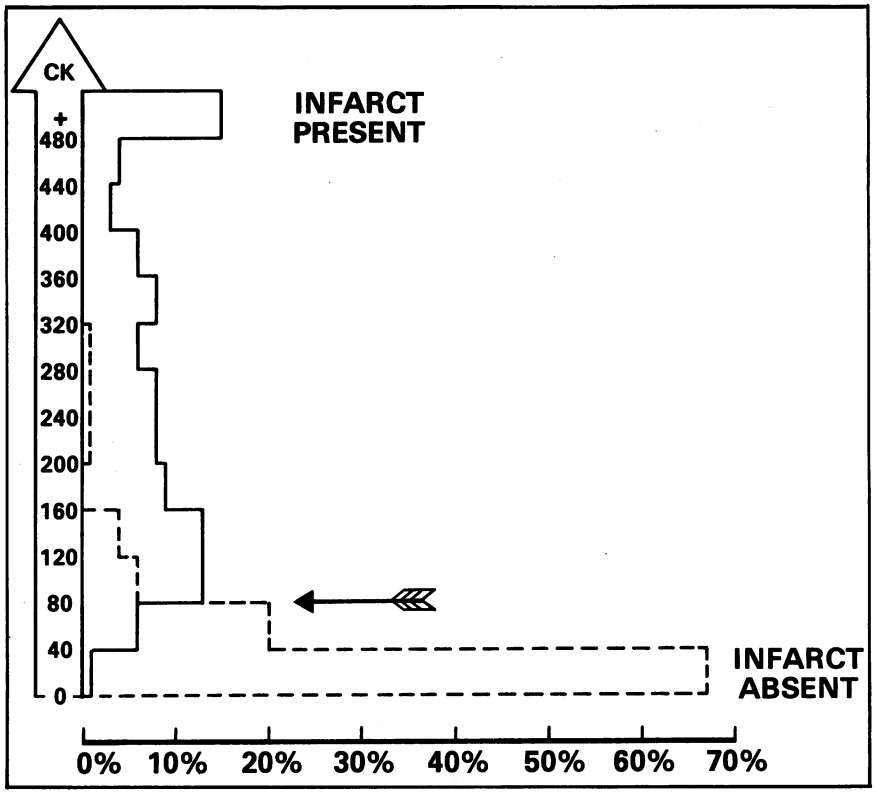


**FIG. 3—Intersection of serum CK levels for patients with (solid lines) and without (dotted lines) myocardial infarcts.**

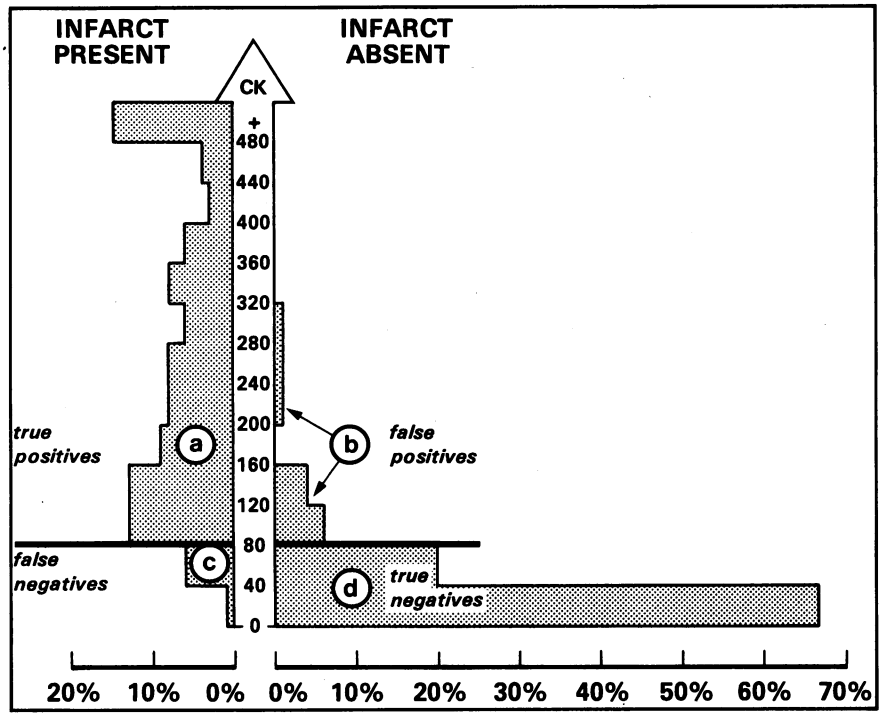


**FIG. 4—Use of serum CK level of 80 U/l as cut-off point in diagnosis of myocardial infarction.**

*While we will follow this layout throughout our series of rounds, you will come across articles in which the locations of affected and unaffected patients are switched or the whole table is rotated 90° and so forth. It is therefore important to check out how such tables are oriented and arranged before you try to interpret them. In fact, you may find it easier to redraw tables that have an unfamiliar format.

point (true negatives) belong in cell d. Thus, we have 215 true positives in cell a, 16 false positives in cell b, 15 false negatives in cell c and 114 true negatives in cell d.

**Table I—Conversion of Fig. 4 to a simple table**

| Serum creatine kinase (CK) test result | Myocardial infarct: Present | Myocardial infarct: Absent |
|---|---|---|
| Positive (level ≥ 80 U/l) | True positives | False positives |
| Negative (level < 80 U/l) | False negatives | True negatives |

a b
c d

**Table II—Filling in the simple table**

| Serum CK test result | Myocardial infarct: Present | Myocardial infarct: Absent |
|---|---|---|
| Positive | 35<br>8<br>7<br>15<br>19<br>13<br>18<br>19<br>21<br>30<br>30<br>215 | 1<br>1<br>1<br>5<br>8<br>16 |
| Negative | 13<br>2<br>15 | 26<br>88<br>114 |
| | a + c: 230 | b + d: 130 |

a b
c d

Having filled the table, we can now do some simple maths and generate some clinically useful indexes of the value of measuring serum CK levels in patients suspected of having a myocardial infarct. Table III shows:

- (a + c): the total number of patients with infarcts, regardless of the CK test results (230).
- (b + d): the total number of patients who do not have infarcts, regardless of the test results (130).
- (a + b): the total number of patients with positive test results, regardless of whether they have infarcts (231).
- (c + d): the total number of patients with negative test results, regardless of whether they have infarcts (129).

What have we gained by converting our attractive pictures into this squat table? It is now possible to elucidate two very important insights about signs, symptoms and laboratory tests. First, we can now describe, in a very concise, economical and easily remembered way, how "good" these data are at helping us decide whether patients have the target disorder. Second, we can now recognize a crucial fickleness of diagnostic data that, unless we understand it, causes a waste of time, money and diagnostic effort.

Let's start with the concise description of how good the CK test (or any diagnostic test) really is. To help us remember the various measures of how good a test is and how to communicate these measures to each other, a set of technical terms has been developed to designate each measure; we shall introduce the terms along the way.

The positive and negative results of the diagnostic test are in the upper and lower horizontal rows, respectively, of Table III. The vertical columns represent what the patients *really* had, according to all the other information we have about them, including autopsy findings. Because this information is "harder", or more certain, than the result of CK testing, we often refer to the vertical columns as representing the "gold standard". Thus, the gold standard of serial ECGs, clinical course and autopsy results tells us whether the patients *really* did or did not have infarcts.

How good is the CK test among patients with infarcts? Well, of the 230 patients who had infarcts, the 214 in cell a had positive test results; 214/230, or a/(a + c), is 0.93, or 93%. Thus, the CK test correctly identified 93% of the patients with infarcts; the shorthand term for this property is *sensitivity*.* Although in this series we use the term sensitivity to indicate "positivity in disease" (PiD), or the *proportion of patients with the target disorder who have a positive test result*, it has other scientific meanings.† You might also be more comfortable using other terms to denote a/(a + c) — for example, the *PiD rate* (with apologies to the gonococcus and the sciatic nerve). Another term is the *true-positive rate*, or *TP rate*, which is based on cell a. This term is often used in more advanced discussions of diagnosis (as in part 4 of our series), but it does not indicate clearly the denominator for the rate, which should be (a + c) (true positives plus false negatives). You can pick any of these three terms — sensitivity, PiD rate and TP rate — and relabel subsequent tables and discussions to suit yourself.

Now, what about the complementary ability to correctly identify the absence of the target disorder? Look again at Table III. Among the 130 patients who did not have infarcts, the 114 patients in cell d had negative test results; 114/130, or d/(b + d), is 0.88, or 88%. Thus, the CK

**Table III—Preparing to determine the clinical usefulness of the CK test**

| Serum CK test result | Myocardial infarct: Present | Myocardial infarct: Absent | |
|---|---|---|---|
| Positive | 215 | 16 | 231 (a + b) |
| Negative | 15 | 114 | 129 (c + d) |
| | a + c: 230 | b + d: 130 | |

a b
c d

*The terms "sensitivity" and "specificity" were introduced by Yerushalmy[2] in 1947 to describe the accuracy of chest x-ray films.

†Another meaning of sensitivity with which you are probably more familiar is the ability of an analytic method to detect minute amounts of a target substance.

test correctly identified 88% of the patients who did not have infarcts; the shorthand term for this property is *specificity*. Again, although specificity is the most commonly used term for this property, it too has another scientific meaning.* You might also want to use other terms for d/(b + d) — for example, the *NiH rate* (with apologies to the National Institutes of Health, Bethesda, Maryland), which is an acronym for "negativity in health". However, the H is not always correct since, although it refers to patients who do not have the target disorder, such patients are not necessarily healthy. Another term is the *true-negative rate*, or *TN rate*, which is based on cell d. It is the counterpart of the TP rate and has similar advantages and disadvantages. As with sensitivity, pick the term you find most sensible for d/(b + d) and use it, remembering that specificity = NiH rate = TN rate = d/(b + d).

*Another meaning of specificity is the ability of an analytic method to detect a single target substance and no others.

So, the CK test has a sensitivity of 93% and a specificity of 88%. Is this good or bad? How does the CK test compare with other diagnostic data? To find out, we have summarized the sensitivities and specificities of a variety of symptoms, signs and laboratory tests from other studies[1,3-9] in Table IV.

Table IV contains many lessons for us. In amniotic fluid acetylcholinesterase testing for the presence of neural tube defects[3] the cutoff point is set to miss as few cases of the target disorders (spina bifida and anencephaly) as possible. Inevitably, when you set the sensitivity higher, the specificity falls (Fig. 4). The four diagnostic tests for prostate cancer[4] underscore another lesson. Not only should we think of the physical signs of a target disorder as diagnostic tests for it, but these signs may, in fact, outperform laboratory tests for the same condition. Finally, the study of deep-vein thrombosis[5] reminds us that we may want to use combinations of tests in making diagnostic decisions. Pairs or groups of diagnostic tests can be combined in different ways, as we will show later.

But all this talk about sensitivity and specificity solves the wrong problem. When we use diagnostic tests clinically, we *do not know* who actually had and did not have the target disorder; if we did, we obviously wouldn't need the diagnostic test. Our clinical concern is not a vertical one of sensitivity and specificity, but a horizontal one of the meaning of positive and negative test results. Therefore, let's go back to Table III. The top row shows 231 patients with positive test results. Of these, the 215 in cell a did, indeed, have infarcts. Accordingly, a/(a + b) (or 93%) of the patients with positive CK test results had infarcts. The shorthand name for this propor-

**Table IV—Sensitivities and specificities of some diagnostic data and tests**

| Reference | Study group | Diagnostic data and tests | Target disorder (and gold standard*) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|---|
| 1 | Patients admitted to coronary care unit | Serum CK level (≥ 80 U/l) | Myocardial infarction (electrocardiograms, clinical course and autopsy) | 93 | 88 |
| 3 | Women with positive results of α-fetoprotein tests | Amniotic fluid acetylcholinesterase electrophoresis | Neural tube defects (direct examination) | 99.5 | 66 |
| 4 | Men with symptoms of urinary obstruction | Acid phosphatase assay | Prostate cancer (transrectal biopsy) | 56 | 94 |
| | | Prostatic secretion cytology | | 29 | 98 |
| | | Aspiration cytology | | 55 | 91 |
| | | Rectal exam for induration or nodules | | 69 | 89 |
| 5 | Patients with symptoms suggesting deep-vein thrombosis | One or both of impedance plethysmography and 125-iodine fibrinogen leg scanning | Deep-vein thrombosis (venography) | 90 | 95 |
| 6 | Patients referred for an upper gastrointestinal tract radiologic series | History of ulcer, over age 50, pain relieved by food or milk, or pain shortly after eating | Ulcer, hiatus hernia, abnormal motility or other important finding (roentgenography) | 95 | 30 |
| 7 | Patients with unstable angina referred for coronary angiography | History of crescendo angina (as opposed to angina of recent onset) | Stenosis of 50% or more of left mainstem coronary artery (coronary angiography) | 83 | 88 |
| 8 | Patients and volunteers | Absence of spontaneous pulsation of the retinal vein | Increased intracranial pressure (lumbar puncture, surgery or diagnostic imaging) | 100 | 88 |
| 9 | Patients suspected of having pancreatic disease | Ultrasonography | Pancreatic cancer or other pancreatic disease (surgery or autopsy) | 65 | 82 |
| | | Computerized tomographic scanning | | 90 | 82 |

*Diagnostic tests to determine presence or absence of disorder.

tion of patients with positive test results who have the target disorder is the *positive predictive value.* Other terms include *predictive value of a positive test, post-test likelihood of disease following a positive test* (which we'll encounter in part 5) and *posterior probability of disease following a positive test* (also in part 5).

The second row of Table III shows 129 patients with negative test results. Of these, the 114 in cell d did not have infarcts. Thus, d/(c + d) (or 88%) of the patients with negative CK test results did not have infarcts. The shorthand name for this proportion of patients with negative test results who do not have the target disorder is the *negative predictive value.* The other terms are analogous to the ones for positive predictive value: *predictive value of a negative test, post-test likelihood of no disease following a negative test* and *posterior probability of no disease following a negative test.** We have updated Table III to include all these terms (Table V), which makes it look much more useful clinically.†

*You might want to use the complement of the last two terms, the post-test likelihood or the posterior probability of disease following a negative test. In other words, c/(c + d), or 15/129 (12%); note that the 12% and the 88% total 100%.

Over 90% of the patients with CK test results of 80 U/l or more had infarcts, and almost 90% of the patients with results of less than 80 U/l did not have infarcts. Right? Not quite. This is where we gain the second very important insight about diagnostic data: their crucial fickleness.

The predictive values of diagnostic signs, symptoms and laboratory tests are not constant, since they must change with the proportion of patients who actually have the target disorder among those who undergo the diagnostic tests. Table V shows 360 patients consecutively admitted to a CCU who were suspected of having a myocardial infarction. Of these, 230 (a + c) — that is, (a + c)/(a + b + c + d), or 230/360 (64%) — did have myocardial infarcts. This proportion is also called *prevalence, pre-test likelihood of disease* or *prior probability of disease.* We'll use the term prevalence for now; the other terms appear in part 5.

Now, suppose that a group of clinicians at another hospital without a CCU were so impressed with the performance of the serum CK test in Edinburgh that they decided to use it routinely for all the patients (except those with skeletal muscle trauma) admitted to their hospital in whom myocardial infarction was even remotely suspected. Of course, we would expect the prevalence of myocardial infarction to be far lower among general admissions than among CCU admissions; indeed, the proportion might be 10% rather than 64%.

Would this really matter? Well, if the serum CK levels of patients with and without myocardial infarcts on general wards were the same as those of our patients in the CCU (that is, if the 80-U/l cut-off point produced a sensitivity of 93% and a specificity of 88%), the use of the test among the former group of patients would produce disappointing results (Table VI). The predictive values have changed dramatically, and now the majority of patients with positive test results did *not* have infarcts. The explanation is straightforward, though perhaps not immediately obvious, and can be grasped by comparing the entries in cell b of Tables V and VI. Although the test's specificity is the same (88%) in each table, the number of patients without infarcts rose from 130 in Table V to 2070 in Table VI. Because 12% of these patients wind up in cell b (since 88% of them go to cell d), the number in cell b rose from 16 in Table V to 248 in Table VI, thereby exceeding the 215 in cell a of both tables.‡

Although the decrease in the positive predictive value is the most

†Incidentally, that the sensitivity and positive predictive values and specificity and negative predictive values are identical (93% and 88% respectively) is happenstance and is due to the fact that cells b and c are almost equal. This is rarely the case, so you should not draw any generalizations from it.

‡The foregoing assumes there is *no change* in the sensitivity and specificity of the diagnostic test when the prevalence changes. Although convincing data are lacking, the sensitivity could decrease with changes in the prevalence if, for example, the infarcts among patients on general wards were less severe as well as less common. The specificity could decrease also if, for example, more patients without infarcts on general wards had accidentally been given intramuscular injections, which can produce false-positive CK test results. If the sensitivity fell, cell a would become smaller and cell c larger, and both predictive values would decrease further. Similarly, if the specificity fell, cell b would become larger and cell d smaller, and again both predictive values would decrease. Therefore, when the prevalence falls, the sensitivity and specificity may change, but the predictive value *must* change.

**Table V—Sensitivity, specificity and predictive values of the CK test in patients with and without myocardial infarcts admitted to a coronary care unit**

| Serum CK test result | Myocardial infarct: Present | Myocardial infarct: Absent | | |
|---|---|---|---|---|
| Positive | 215 (a) | 16 (b) | 231 | Positive predictive value (predictive value of a positive test or post-test likelihood or posterior probability of disease) $= \frac{a}{a+b} = \frac{215}{231} = 93\%$ |
| Negative | 15 (c) | 114 (d) | 129 | Negative predictive value (predictive value of a negative test or post-test likelihood or posterior probability of no disease) $= \frac{d}{c+d} = \frac{114}{129} = 88\%$ |
| | 230 | 130 | 360 | Prevalence (pretest likelihood or prior probability) of disease $= \frac{a+c}{a+b+c+d} = \frac{230}{360} = 64\%$ |
| | Sensitivity (PiD or TP rate) $= \frac{a}{a+c} = \frac{215}{230} = 93\%$ | Specificity (NiH or TN rate) $= \frac{d}{b+d} = \frac{114}{130} = 88\%$ | | |

dramatic effect of a decrease in prevalence, you may have noted a concomitant increase in the negative predictive value, from 88% in Table V to 99% in Table VI. The explanation for this change is analogous to that for the fall in the positive predictive value, and can be seen by comparing cell d in the two tables.

Is this fickleness a generalizable phenomenon? Unfortunately, it is. As the prevalence falls, so too must the positive predictive value, and the negative predictive value rises. Even an excellent symptom, sign or laboratory test with a sensitivity and a specificity of 95% will lose positive predictive value and gain negative predictive value as prevalence falls (Table VII). The table suggests that as prevalence falls you learn much more from the absence rather than the presence of a sign or symptom or from a negative rather than a positive test result. This is partly true and fits very nicely into the hypotheticodeductive approach to diagnosis because it helps us shorten our list of hypotheses by demonstrating that some of them are wildly improbable and therefore ought to be dropped. But if this is so, why did we say that it is only partly true that we learn more from negative than positive diagnostic data when the prevalence of the condition is low? Look at the bottom row of Table VII. It is the complement of the row above it and gives the probability that a patient *has* the disease despite a negative test result (thus, the column entries in the bottom two rows always add up to 100%). Now, compare the entries in the bottom row with those in the top row. In the middle of the table they are very far apart. For example, when the likelihood of disease is 50% before testing, the absence of an excellent sign or symptom or a negative test result drops it by 45%, to a post-test likelihood of disease of only 5%; hence, a formerly quite plausible diagnostic hypothesis should now be rejected. However, the drop from pretest to post-test likelihood because of a negative test result becomes progressively smaller as we move toward either extreme of the pretest likelihood. Thus, when a diagnostic hypothesis is extremely unlikely, say the pretest likelihood is 5%, the absence of the key sign or symptom or a negative test result produces a drop of only 4.7%, to a post-test likelihood of 0.3%, and you have learned very little.

What if the sign, symptom or laboratory study is not an excellent test? What if it is a more typical diagnostic test, whose sensitivity is 85% and whose specificity is 90%? You can calculate the answer for yourself and confirm (we hope) that the change (rise or drop) from pretest to post-test likelihood of disease is muted (hence you learn less) as the sensitivity or specificity or both decrease.

You will, on average, learn the most from a clinical sign, symptom or laboratory test when the pretest likelihood of disease is 40% to 60%. At this level, the presence of the clinical finding or a positive test result virtually clinches the diagnosis, and the negative test results effectively eliminate the target disorder from your list of hypotheses. In other words, a sign, symptom or laboratory test is of greatest benefit when you are in a 50–50 dilemma and cannot decide whether the patient has the target disorder. Thus, our tables and formulas converge with common sense.

How do you get a prevalence (pretest likelihood) of 40% to 60%? This achievement, long obscured as part of the "art of medicine", will be exposed for the science that it really is in part 3 of our series.

**Table VI—Sensitivity, specificity and predictive values of the CK test in patients with and without myocardial infarcts on general wards**

| Serum CK test result | Myocardial infarct: Present | Myocardial infarct: Absent | | |
|---|---|---|---|---|
| Positive | 215 | 248 | 463 | Positive predictive value $= \frac{a}{a+b} = \frac{215}{463} = 46\%$ |
| Negative | 15 | 1822 | 1837 | Negative predictive value $= \frac{d}{c+d} = \frac{1822}{1837} = 99\%$ |
| | 230 | 2070 | 2300 | Prevalence $= \frac{a+c}{a+b+c+d} = \frac{230}{2300} = 10\%$ |
| | Sensitivity $= \frac{a}{a+c} = \frac{215}{230} = 93\%$ | Specificity $= \frac{d}{b+d} = \frac{1822}{2070} = 88\%$ | | |

a b
c d

## References


1. SMITH AF: Diagnostic value of serum-creatine-kinase in a coronary care unit. *Lancet* 1967; 2: 178–182

2. YERUSHALMY J: Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Rep* 1947; 62: 1432–1449


*continued on page 587*

**Table VII—Effect of prevalence on the predictive values of an excellent sign, symptom or laboratory test (sensitivity and specificity 95% in all cases)**

| Variable | Values (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prevalence (pretest likelihood) | 99.0 | 95.0 | 90.0 | 50.0 | 10.0 | 5.0 | 1.0 | 0.5 | 0.1 |
| Positive predictive value | 99.9 | 99.7 | 99.4 | 95.0 | 68.0 | 50.0 | 16.0 | 9.0 | 2.0 |
| Negative predictive value | | | | | | | | | |
| No disease | 16.0 | 50.0 | 68.0 | 95.0 | 99.4 | 99.7 | 99.9 | 99.97 | 99.99 |
| Disease | 84.0 | 50.0 | 32.0 | 5.0 | 0.6 | 0.3 | 0.1 | 0.03 | 0.01 |

13. Oh WMC, Taylor RT, Olsen GJ: Aortic regurgitation in systemic lupus erythematosus requiring aortic valve replacement. *Br Heart J* 1974; 36: 413–416
14. Bidani AK, Roberts JI, Schwartz MM, Lewis EJ: Immunopathology of cardiac lesions in fatal systemic lupus erythematosus. *Am J Med* 1980; 69: 849–858
15. Shapiro RF, Gamble CN, Wiesner KB, Castles JJ, Wolf AW, Hurley EJ, Salel AF: Immunopathogenesis of Libman–Sacks endocarditis. Assessment by light and immunofluorescent microscopy in two patients. *Ann Rheum Dis* 1977; 36: 508–516
16. Murray FT, Fuleihan DS, Cornwall CS, Pinals RS: Acute mitral regurgitation from ruptured chordae tendineae in systemic lupus erythematosus. *J Rheumatol* 1975; 2: 454–459
17. Vaughton KC, Walker DR, Sturridge MF: Mitral valve replacement for mitral stenosis caused by Libman–Sacks endocarditis. *Br Heart J* 1979; 41: 730–733
18. Shulman HJ, Christian CL: Aortic insufficiency in systemic lupus erythematosus. *Arthritis Rheum* 1969; 12: 138–146

---

## *INTERPRETATION*

*continued from page 564*

3. Wald NJ, Cuckle HS: Amniotic fluid acetylcholinesterase electrophoresis as a secondary test in the diagnosis of anencephaly and open spina bifida in early pregnancy. Report of the Collaborative Acetylcholinesterase Study. *Lancet* 1981; 2: 321–324
4. Guinan P, Bush I, Ray V, Vieth R, Rao R, Bhatti R: The accuracy of the rectal examination in the diagnosis of prostate carcinoma. *N Engl J Med* 1980; 303: 499–503
5. Hull R, Hirsh J, Sackett DL, Taylor DW, Carter C, Turpie AGG, Zielinsky A, Powers P, Gent M: Replacement of venography in suspected venous thrombosis by impedance plethysmography and $^{125}$I-fibrinogen leg scanning. *Ann Intern Med* 1981; 94: 12–15
6. Marton KI, Sox HC Jr, Wasson J, Duisenberg CE: The clinical value of the upper gastrointestinal tract roentgenogram series. *Arch Intern Med* 1980; 140: 191–195
7. Plotnick GD, Greene HL, Carliner NH, Becker LC, Fisher ML: Clinical indicators of left main coronary artery disease in unstable angina. *Ann Intern Med* 1979; 91: 149–153
8. Levin BE: The clinical significance of spontaneous pulsations of the retinal vein. *Arch Neurol* 1978; 35: 37–40
9. Hessel SJ, Siegelman SS, McNeil BJ, Sanders RC, Adams OF, Alderson PO, Finberg HJ, Abrams HL: A prospective evaluation of computed tomography and ultrasound of the pancreas. *Radiology* 1982; 143: 129–133

# Clinical Epidemiology Rounds

# Interpretation of diagnostic data: 3. How to do it with a simple table (part B)

DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS, MCMASTER UNIVERSITY, HAMILTON, ONT.

Part 3 of our series on interpreting diagnostic data is a continuation of part 2, in which we showed you how to do it with a simple two-by-two, or fourfold, table. At the close of part 2 we discovered that a diagnostic sign, symptom or laboratory test is most helpful when the patient's pretest likelihood (prevalence or prior probability) of the target disorder is 40% to 60%.

But how do you get the pretest likelihood to 40% to 60%? Doing so is difficult to articulate and even tougher to teach or to study. Accordingly, it often is lumped with other clinical mysteries and referred to as the "art of medicine". Thus, when asked why they asked the key question of the patient, sought the pathognomonic physical sign or ordered the definitive laboratory test, seasoned diagnosticians often either shrug their shoulders, chuckle or mumble about Osler, "Irish hunches", Willie Sutton or "clinical judgement".

We are now beginning to demystify these and other elements of "the science of the art of medicine". We now recognize that these diagnosticians have elicited, without us (and sometimes them) noticing it, specific points in the patient's history and the results of the physical examination that, when present or positive, signify a prevalence of the target disorder of 40% to 60%. We will illustrate this with a common problem in primary care and internal medicine: making a diagnosis in the ambulatory patient with episodic chest pain.

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

## Doing it with a simple table

### *Case presentations*

Patient A is a 55-year-old mildly hypertensive man with a 4-week history of substernal pressure and pain that radiate to his neck and lower jaw and down the inner aspect of his left arm. The problem is precipitated by climbing stairs or walking uphill, and it disappears after 3 to 5 minutes of rest. He has a mild episode of pain while undressing for his examination, and as it is abating you think you hear an S4 gallop. You decide he has classic angina of effort and consider ordering an exercise electrocardiogram (ECG) to nail down the diagnosis.

Patient B is a 35-year-old man who is otherwise healthy and has no coronary risk factors. He has had "heartburn" for years and now reports a 6-week history of nonexertional, squeezing pain deep to his lower sternum and epigastrium, usually radiating through to his back. It usually occurs when he lies down after a heavy meal. The remainder of his history and the physical examination yield no abnormalities. You think that his pain is from esophageal spasm but judge that an exercise ECG will resolve the uncertainty.

Patient C is a 45-year-old, previously healthy man with no coronary risk factors save a pack-a-day cigarette habit. He reports a 3-week history of precordial and substernal pain that is usually fleeting and stabbing but occasionally feels as if a heavy weight were on his chest; it is inconsistently related to exertion. You find that one costochondral junction is slightly tender, but pressing on it does not reproduce the patient's pain. You conclude that he may have atypical angina and wonder if an exercise ECG would help confirm the diagnosis.

### *Comment*

The list of hypotheses for such patients usually includes coronary artery disease, and we often consider ordering an exercise ECG. A positive exercise test result will suggest severe coronary artery disease (for which you may want to proceed to coronary angiography and an aortocoronary bypass), and a negative exercise test result will reassure you and the patient and direct your diagnostic effort elsewhere.

Or will it? Table I summarizes the sensitivity and specificity of the exercise ECG in patients with and without coronary artery disease, as documented by Bartel and colleagues.[1] These patients had various chest pain syndromes and underwent

**This is the third of six articles (the first two appeared in the Sept. 1 and 15, 1983 issues of the Journal) that focus on the strategies and tactics for interpreting diagnostic data, both the clinical data from the history and physical examination and the paraclinical data from the clinical laboratories, the radiology department and the surgical pathology service. The remaining articles will appear in the next three issues of the Journal.**

both exercise electrocardiography and coronary angiography. If we use the latter as the "gold standard", the exercise ECG had a sensitivity of 60% and a specificity of 91% when 1 mm of ST-segment depression was selected as the cut-off point.

### Importance of pretest likelihood of disease

Now let's consider the three ambulatory patients with episodic chest pain for whom you are considering ordering an exercise ECG. What is your estimate of the pretest likelihood of significant coronary artery narrowing in each patient? (Jot it down in the margin or on a slip of paper.) Although we have very few details about these patients, we do know a fair bit about their pretest likelihood of disease. On the basis of only their age, sex and type of pain, we can be quite exact about the probability that each of them has clinically important coronary artery disease. For patient A the pretest likelihood is very high (about 90%), for patient B it is very low (about 5%), and for patient C it is intermediate (about 50%). These pretest rates were carefully documented by Diamond and Forrester,[2] who studied the results of coronary angiography or autopsy in thousands of patients with and without various chest pain syndromes. However, their exhaustive study only verifies what experienced clinicians do "naturally": most, on the basis of just the "soft" information supplied in our case presentations, obtain estimates of pretest likelihood that are very close to the "official" ones we have cited. How close were yours?

Given these estimates and the sensitivity and specificity of the exercise ECG, how useful will the exercise ECG be in our three patients? (Again, stop reading and jot down your estimates.) We think it would be diagnostically useless in patients A and B and therefore should not be done. In patient C, however, it will be extremely valuable. The reasons for our judgements, if not already obvious, should be clear from Table II. It shows that patients A's pretest likelihood of 90% is influenced very little by the result of exercise electrocardiography. We knew before the test that he almost certainly had significant coronary artery disease. Even if the result is negative, the likelihood that he has significant coronary artery disease is still 80%.* Therefore, he does not need an exercise ECG. Indeed, the important clinical decision for this patient is whether we should proceed directly to coronary angiography (depending on our judgement of the ultimate risks and benefits of finding out whether he is a suitable candidate for one or more bypass grafts).

Patient B does not need an exercise ECG either. We already estimated his pretest likelihood to be only 5%, and neither a positive nor a negative result of exercise electrocardiography will alter that likelihood in an important way. Even if the result is positive, the odds are still three to one against his having significant coronary artery disease. This is confirmed in Table II.

It may seem quite cavalier (especially to those of you still in subspecialty training) for us to spurn the exercise ECG for patient B. After all, he did have chest pain of a sort, and for every 1000 patients like him we will miss 30 who do have significant coronary artery disease and whose exercise ECG results will be positive (cell a of patient B's panel in Table II). Have we not done these 30 patients a disservice by forgoing further testing, especially since exercise electrocardiography is so safe? The answer lies in weighing the potential good we do the 30 patients who really have significant coronary artery disease and will have positive exercise ECG results against the potential harm we do the two other groups: first, the 20 people in cell c who will be given false-negative clean bills of health (they have coronary artery disease but negative exercise ECG results), and, second, the 86 patients in cell b who will be given false-positive, scary labels of severe coronary artery disease when they do not have it. As you probably know, being told you have a disease when you do not is frequently as disabling as actually having it[3] and is almost always more disabling than not knowing you have it when you do. In our judgement, the harm we do the 86 patients in cell b and the 20 in cell c outweighs the good we do the 30 in cell a. We will return to this case in part 6 of our series.

Patient C can really benefit (at least diagnostically) from undergoing exercise electrocardiography. We estimated his pretest likelihood to be 50%. If the exercise ECG result is positive, the likelihood rises

*Because the post-test likelihood of no disease in the presence of a negative result of a diagnostic test (negative predictive value) is d/(c + d), we subtract the result from 100% to get the post-test likelihood that the patient does have the disease despite a negative test result. We could get the same result by calculating c/(c + d).

**Table I—Sensitivity and specificity of the exercise electrocardiogram (ECG) in patients with and without coronary artery disease***

| Exercise ECG result | Coronary artery disease: Present | Coronary artery disease: Absent | |
|---|---|---|---|
| Positive (ST-segment depression ≥ 1 mm) | 137 (a) | 11 (b) | 148 (a + b) |
| Negative (ST-segment depression < 1 mm) | 90 (c) | 112 (d) | 202 (c + d) |
| | 227 (a + c) | 123 (b + d) | 350 |
| | Sensitivity = $\frac{a}{a+c} = \frac{137}{227}$ = 60% | Specificity = $\frac{d}{b+d} = \frac{112}{123}$ = 91% | |

*Adapted from reference 1.

37%, from 50% to 87%, and we have established his diagnosis. On the other hand, if the result is negative, the likelihood of significant coronary artery disease drops 19%, from 50% to 31%, and we had better consider looking elsewhere for an explanation for his pain.

Our finding of a post-test likelihood of 31% should raise two questions: first, at what post-test likelihood should we stop the diagnostic process (i.e., how low does it have to be to reject the diagnosis and how high to accept it?), and, second, do we really have to quantitate our diagnostic uncertainty this way?

To answer the first question we should consider the courses of action open to us when patient C's exercise ECG result is negative and the post-test likelihood is thus 31%. We have four courses of action, as follows:

- We could increase our own sophistication in interpreting the diagnostic data and see whether patient C's exercise ECG result can be interpreted as more than just "negative". For example, we have been assessing the value of the exercise ECG by noting whether the ST-segment depression was more or less than the 1-mm cut-off point. How-

**Table II—Usefulness of the exercise ECG in three patients**

**Patient A**

| Exercise ECG result | Coronary artery disease: Present | Coronary artery disease: Absent | Post-test likelihood | Change from pretest likelihood |
|---|---|---|---|---|
| Positive | 540 (a) | 9 (b) | 549 $= \frac{a}{a+b} = \frac{540}{549} = 98\%$ | +8% |
| Negative | 360 (c) | 91 (d) | 451 $= 100\% - \frac{d}{c+d} = 100\% - \frac{91}{451} = 80\%$ | −10% |
| | 900 $\frac{a}{a+c} = \frac{540}{900} = 60\%$ | 100 $\frac{d}{b+d} = \frac{91}{100} = 91\%$ | 1000 Pretest likelihood $= \frac{a+c}{a+b+c+d} = \frac{900}{1000} = 90\%$ | |

**Patient B**

| Exercise ECG result | Coronary artery disease: Present | Coronary artery disease: Absent | Post-test likelihood | Change from pretest likelihood |
|---|---|---|---|---|
| Positive | 30 (a) | 86 (b) | 116 $= \frac{a}{a+b} = \frac{30}{116} = 26\%$ | +21% |
| Negative | 20 (c) | 864 (d) | 884 $= 100\% - \frac{d}{c+d} = 100\% - \frac{864}{884} = 2\%$ | −3% |
| | 50 $\frac{a}{a+c} = \frac{30}{50} = 60\%$ | 950 $\frac{d}{b+d} = \frac{864}{950} = 91\%$ | 1000 Pretest likelihood $= \frac{a+c}{a+b+c+d} = \frac{50}{1000} = 5\%$ | |

**Patient C**

| Exercise ECG result | Coronary artery disease: Present | Coronary artery disease: Absent | Post-test likelihood | Change from pretest likelihood |
|---|---|---|---|---|
| Positive | 300 (a) | 45 (b) | 345 $= \frac{a}{a+b} = \frac{300}{345} = 87\%$ | +37% |
| Negative | 200 (c) | 455 (d) | 655 $= 100\% - \frac{d}{c+d} = 100\% - \frac{455}{655} = 31\%$ | −19% |
| | 500 $\frac{a}{a+c} = \frac{300}{500} = 60\%$ | 500 $\frac{d}{b+d} = \frac{455}{500} = 91\%$ | 1000 Pretest likelihood $= \frac{a+c}{a+b+c+d} = \frac{500}{1000} = 50\%$ | |

ever, it is possible — and more accurate — to grade the depression more finely than that. For instance, if there is less than 0.5 mm of ST-segment depression, the post-test likelihood of severe coronary artery stenosis is, in fact, only 17%. This eye-opener shows how much information we toss out when we reduce a range of diagnostic test results to a positive/negative dichotomy. We will return to this in part 4 of our series.

- We could choose to have patient C undergo coronary angiography. However, this is not a very attractive alternative in terms of discomfort, risk and cost, and the fact that two thirds of such patients would have negative results of angiography.
- We could apply other noninvasive tests (e.g., thallium-201 scanning and scintigraphy) in the hope that they would decisively shift the post-test likelihood.
- We could compare the post-test likelihood of 31% with the likelihoods we have assigned to the other diagnostic hypotheses on our short list. If pain in the chest wall now jumps to the fore with a likelihood of 65%, we may want to watch and wait.

The last option is the answer to the first question: you toss out (or accept) a diagnostic hypothesis when its likelihood (suitably tempered by considering the harm you will do if you are wrong) is substantially lower (or higher) than the likelihood of other diagnostic hypotheses on your short list. If the harm of missing a diagnosis is great you will need a very low likelihood (less than 10%) before you toss it out. (This is why patients with a probability of myocardial infarction of only 20% to 25% still go to the coronary care unit for serial enzyme determinations and daily electrocardiography.) Conversely, if the harm of overdiagnosis is great you will need a very high likelihood (greater than 95%) before you accept the diagnosis as established. This is why we insist on tissue diagnosis before we tell patients they have cancer and why we demand repeated blood pressure measurements before we label patients as hypertensive. When misdiagnosis carries lower penalties we relax and may reject a diagnostic hypothesis at 40% or accept it at 60%.

All of the foregoing underscores our second question: Do we really have to quantitate our diagnostic uncertainty this way? Only if we want to be better clinicians. Of course we deal in uncertainty in everything we do: we cannot be sure about the patient's history, the results of the physical examination, the results of laboratory tests, the diagnosis, the prognosis or whether the treatment will work (even if the patient does take the medicine we have prescribed). How shall we handle this uncertainty? Shall we simply ignore it or mystify it by calling it the "art" of medicine? Let's not. Let's use the science of the art of medicine and get the most out of it. Certainty is a delusion, and uncertainty can, within limits, be quantified to the benefit of the patient. We echo the advice of one of the heroes mentioned in our earlier set of rounds,[4,5] David Spodick:[6] "Physicians must be content to end not in certainties, but rather in statistical probabilities. The modern [physician] thus has a right to feel certain, within statistical constraints, but never cocksure. Absolute certainty remains for some theologians — and like-minded physicians."

## Multiple diagnostic tests

There is one other aspect of "doing it with a simple table" that we want to show you: the simultaneous use of two diagnostic tests. Suppose we have two different diagnostic tests for a target disorder. Test A has a sensitivity of 70% and a specificity of 95%, and test B has a sensitivity of 90% and a specificity of 75%. The pretest likelihood of the target disorder is 50%. Finally, suppose that the two tests are "independent" of one another — that is, the positivity rate of one test is the same for patients with positive and negative results of the second test. In other words, of the patients with the target disorder, 90% of both those with a *positive* result of test A and those with a *negative* result of test A will also have a positive result of test B.

Patients who undergo both test A and test B can have three sorts of results: positive results of both, positive results of one and negative results of the other, or negative results of both. If we are going to reduce this set of results to a simple table, however, we have to decide where to put the cut-off point for a combination of the results. Should we require that the results of both tests A and B be positive, or should we include as positive results those of the patients with just one positive test result? As shown in Table III, it makes a difference.

When the cut-off point requires that both test results be positive, the sensitivity is not very impressive (63%), but the specificity is splendid (99%). Hence, although you may miss several cases of the target disorder, the patients with positive results of both tests A and B (at least those with a pretest likelihood of 50%) virtually always have the disorder. This is good way to minimize the chances of mislabelling the "innocent".

**Table III—Two ways of combining the results of two independent diagnostic tests**

| Test results | Target disorder: Present | Target disorder: Absent | Test results: Both positive | Test results: One or both positive |
|---|---|---|---|---|
| Both positive | 630 | 12 | 630 (a) | 12 (b) |
| One positive | 340 | 276 | 370 (c) | 988 (d) |
| Neither positive | 30 | 712 | 970 (a+c) | 288 (b+d) |
| | | | 30 (a+c) | 712 (b+d) |
| | | Sensitivity | 63% | 97% |
| | | Specificity | 99% | 71% |

Alternatively, you could lower the cut-off point to include just one (or both) positive test result(s), as shown in the right-hand panel of Table III. The sensitivity jumps to 97%, but the specificity falls to 71%. This is a good way to be sure, when both test results are negative, that your patient does not have the target disorder. Almost every case will be caught, although there will also be a large number of false-positive results (cell b).

What if, instead, the two tests are dependent upon each other? Sup-

**Table IV—Doing it with a simple table**

**Panel A**

| Test result | Target disorder: Present | Target disorder: Absent | |
|---|---|---|---|
| Positive | a | b | a + b |
| Negative | c | d | c + d |
| | a + c | b + d | a + b + c + d |
| | | | 1000 |

**Panel B**

| Test result | Target disorder: Present | Target disorder: Absent | |
|---|---|---|---|
| Positive | a | b | a + b |
| Negative | c | d | c + d |
| | a + c | b + d | a + b + c + d |
| | 100 | 900 | 1000 |

Pretest likelihood = 10%

**Panel C**

| Test result | Target disorder: Present | Target disorder: Absent | |
|---|---|---|---|
| Positive | 83 | 81 | 164 |
| | a | b | a + b |
| | c | d | c + d |
| Negative | 17 | 819 | 836 |
| | a + c | b + d | a + b + c + d |
| | 100 | 900 | 1000 |
| | Sensitivity = 83% | Specificity = 91% | |

Pretest likelihood = 10%

**Panel D**

| Test result | Target disorder: Present | Target disorder: Absent | |
|---|---|---|---|
| Positive | 83 | 81 | 164 |
| | a | b | a + b |
| | c | d | c + d |
| Negative | 17 | 819 | 836 |
| | a + c | b + d | a + b + c + d |
| | 100 | 900 | 1000 |
| | Sensitivity = 83% | Specificity = 91% | |

Positive predictive value $= \frac{a}{a+b} = \frac{83}{164} = 51\%$

Negative predictive value $= \frac{d}{c+d} = \frac{819}{836} = 98\%$

Pretest likelihood = 10%

pose that when the result of test A is positive, that of test B is *more* likely to be positive than when the former is negative, as when both tests are detecting the same manifestation of the target disorder. This is called "convergence". This is the usual case when we use combinations of diagnostic tests for the same target disorder. When convergent diagnostic tests are used, the gain in diagnostic certainty from the second test is diminished, and the selection of the cut-off point will make less difference.

The second way in which two diagnostic tests can be dependent upon each other is when they are "divergent". Suppose that when the result of test A is positive, that of test B is *less* likely to be positive than when the former is negative, as when the target disorder is a family of diseases, each with its own distinct diagnostic features. Another situation in which the diagnostic tests may be divergent is when the disorder has distinct, irreversible stages such that progression from one stage to the next "switches off" some diagnostic tests as it "switches on" others. When combinations of diagnostic tests are divergent, the gain in diagnostic certainty from the second test is augmented, and the selection of the cut-off point will make a great difference (even more so than when the two tests are independent).

## Summary

The following guidelines are useful if you want to "do it with a simple table" (Table IV):

- First, identify the sensitivity and specificity of the sign, symptom or diagnostic test you plan to use. Many are already in the literature, and subspecialists should either know them for their field or be able to track them down for you. Depending on whether you are considering a sign, a symptom or a diagnostic laboratory test, you will want to track down a clinical subspecialist, a radiologist, a pathologist and so on.
- Start your table with a total of 1000 patients, as shown in location (a + b + c + d) of panel A.
- Using the information you have about the patient before you apply the diagnostic test, estimate the patient's pretest likelihood (prevalence or prior probability) of the target disorder — let's say 10%. Take this proportion of the total (100) and place it in location (a + c); the remaining 900 patients go in location (b + d) (panel B).
- Multiply (a + c) (100) by the sensitivity of the diagnostic test (let's say 83%) and place the result (83) in cell a and the difference (17) in cell c; similarly, multiply (b + d) (900) by the specificity of the diagnostic test (let's say 91%) and place the result (819) in cell d and the difference (81) in cell b (panel C). If (a + b) and (c + d) do not add up to 1000, you will know you have made a mistake.
- You can now calculate the positive predictive value, a/(a + b), and the negative predictive value, d/(c + d), as shown in panel D.

You have now reached a level of understanding a fair bit beyond the rule-in/rule-out strategy discussed in part 1 of our series. Furthermore, you can already do more than most clinicians, so you may want to stop here, at least for a while. On the other hand, you may want to go further and learn how to handle slightly more complex tables with multiple cut-off points. In the next article you will find more powerful ways to take advantage of the *degree* of positivity and negativity of diagnostic test results.

## References


1. BARTEL AG, BEHAR VS, PETER RH, ORGAIN ES, KONG Y: Graded exercise stress tests in angiographically documented coronary artery disease. *Circulation* 1974; 49: 348–356
2. 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
3. BERGMAN AB, STAMM SJ: The morbidity of cardiac nondisease in schoolchildren. *N Engl J Med* 1967; 276: 1008–1013
4. Department of clinical epidemiology and biostatistics, McMaster University, Hamilton, Ont.: Clinical disagreement: I. How often it occurs and why. *Can Med Assoc J* 1980; 123: 499–504
5. Idem: Clinical disagreement: II. How to avoid it and how to learn from one's mistakes. Ibid: 613–617
6. SPODICK DH: On experts and expertise: the effect of variability in observer performance. *Am J Cardiol* 1975; 36: 592–596


*VAGINAL MECHANICAL*
*continued from page 701*


19. WORTMAN JS: The diaphragm and other intravaginal barriers — a review. *Popul Rep [H]* 1976: 57–76
20. HATCHER RA, STEWART GK, STEWART F, GUEST F, SCHWARTZ DW, JONES SA: The diaphragm. In HATCHER RA, STEWART GK: *Contraceptive Technology, Nineteen Eighty–Eighty-One*, 10th ed, Irvington, New York, 1980: 76–83
21. GARA E: Nursing protocol to improve the effectiveness of the contraceptive diaphragm. *Am J Matern Child Nurs* 1981; 6: 41–45
22. HAWKINS DF, ELDER MG: Condoms, diaphragms and caps. In *Human Fertility Control — Theory and Practice*, Butterworths, Woburn, Mass., 1979: 139–142
23. WIDHALM MV: Vaginal lesion: etiology — a malfitting diaphragm? *J Nurse-Midwifery* 1979; 24: 39–40
24. MILLS JL, HARLEY EE, REED GF, BERENDES HW: Are spermicides teratogenic? *JAMA* 1982; 248: 2148–2151


# New Books of Interest

*This list is an acknowledgement of the books received that we intend to send out for review.*

**CRITICAL CARE OF THE NEWBORN.** W. Alan Hodson and William E. Truog. 201 pp. Illust. W.B. Saunders Company Canada Limited, Toronto, 1983. $19.45, spiral-bound. ISBN 0-7216-1071-4

**FAMILY THERAPY IN SCHIZOPHRENIA.** Guilford Family Therapy Series. Edited by William R. McFarlane. 355 pp. Illust. Guilford Press, New York, 1983. $25 (US). ISBN 0-89862-042-2

**GUIDE TO THE MANAGEMENT OF INFECTIOUS DISEASE.** Monographs in Family Medicine. Edited by Laurel G. Case. 238 pp. Illust. Academic Press Canada, Don Mills, Ont., 1983. $40. ISBN 0-8089-1506-1

**THE QUALITY OF MERCY:** The Lives of Sir James and Lady Cantlie. Jean Cantlie Stewart. 277 pp. Illust. George Allen & Unwin (Publishers) Ltd., London, 1983. Price not stated. ISBN 0-04-920066-6

**THE SICK CITADEL.** The American Academic Medical Center and the Public Interest. Irving J. Lewis and Cecil G. Sheps. 290 pp. Oelgeschlager, Gunn & Hain, Inc., Cambridge, Massachusetts, 1983. $25 (US). ISBN 0-89946-173-5

**STRUCTURE AND FUNCTION OF Fc RECEPTORS.** Receptors and Ligands in Intercellular Communication Series. Vol. 2. Edited by Arnold Froese and Frixos Paraskevas. 294 pp. Illust. Marcel Dekker, Inc., New York, 1983. $45 (US). ISBN 0-8247-1814-3



For prescribing information see page 765 →

# Clinical Epidemiology Rounds

## Interpretation of diagnostic data: 4. How to do it with a more complex table

DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS, MCMASTER UNIVERSITY, HAMILTON, ONT.

In parts 2 and 3 of our series on interpreting diagnostic data we showed you how to use a simple two-by-two, or fourfold, table. Now we will show you how to use a more complex table. To ease the transition, we will use the same sample of patients as in part 2.

### Doing it with a more complex table

*Case presentation*

In 360 patients consecutively admitted to a coronary care unit (CCU) blood was drawn for measurement of the serum creatine kinase (CK) level at the time of admission and the next two mornings.[1] A clinician who was "blind" to the CK measurements reviewed the patients' electrocardiograms (ECGs), clinical records and autopsy reports and decided that 230 had had a myocardial infarction and 130 had not.

*Comment*

Fig. 1 shows the maximum serum CK levels* for the 230 patients who had an infarct and the 130 patients who did not.

### Conversion of numbers to a more complex table

In parts 2 and 3 of our series we placed these numbers in simple two-by-two, or fourfold, tables. This time we shall extend the tables to several rather than just two vertical rows. As a result, the tables now retain more of the information presented in Fig. 1 (Table I).

Several of the cut-off points in Table I might look familiar. That at 40 U/l is the X, or rule-out, cut-off point we used in part 1, that at 80 U/l is the one we used in parts 2 and 3 and that at 280 U/l is the Y, or rule-in, cut-off point we used in part 1. The cut-off points need not be evenly spaced; the bottom three are 40 U/l apart, but there is a jump of 200 U/l to the top one.

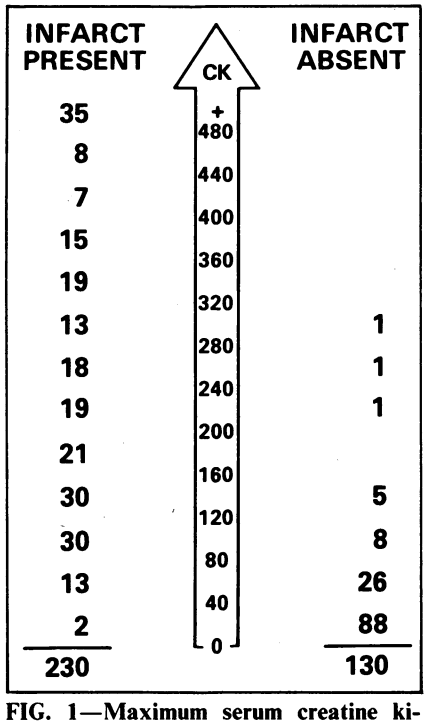


**FIG. 1—Maximum serum creatine kinase (CK) levels (U/l) in patients with and without myocardial infarcts.**

How do we generate sensitivity and specificity from this more complex table? We do it by creating a series of simple tables (Table II), each of which uses one of the cut-off points from the complex table. The left-hand panel of Table II simply replicates Table I and displays the numbers of patients with and without infarcts and their serum CK levels. The right-hand panels display all the simple tables that could be generated from the complex one. Thus, when the cut-off point is at 280 U/l, cell a contains the 97 patients with infarcts whose serum CK levels were 280 U/l or greater, cell c contains the 133 (118 + 13 + 2) patients with infarcts whose levels were less than 280 U/l, and so forth across the remaining panels.

Beneath the right-hand panels are

**Table I—Doing it with a more complex table**

| Serum creatine kinase (CK) level, U/l | Myocardial infarct | |
|---|---|---|
| | Present | Absent |
| ≥ 280 | 97 | 1 |
| 80–279 | 118 | 15 |
| 40–79 | 13 | 26 |
| 1–39 | 2 | 88 |

*Since laboratory methods differ, this distribution of serum levels may not apply at your institution.

Reprint requests to: Dr. P.X. Tugwell, Rm. 2C16, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton, Ont. L8N 3Z5

**This is the fourth of six articles (the first three appeared in the Sept. 1, Sept. 15 and Oct. 1, 1983 issues of the Journal) that focus on the strategies and tactics for interpreting diagnostic data, both the clinical data from the history and physical examination and the paraclinical data from the clinical laboratories, the radiology department and the surgical pathology service. The remaining articles will appear in the next two issues of the Journal.**

the sensitivities and specificities for each cut-off point. A comparison of the sensitivities and specificities reminds us once again that as one goes up, the other goes down. The extreme right-hand panel shows that we can always achieve a sensitivity of 100% if we are willing to drop the specificity to zero.

The increase in the post-test likelihood of disease may be very great when we switch from a simple table to a more complex table. Consider two patients who are admitted to a CCU in whom a myocardial infarct is suspected and who both have a pretest likelihood of 64% (as in Table II, where 230/360 = 64%). Suppose the first patient (A) has a serum CK level of 400 U/l and the second patient (B) a level of 30 U/l. The effect of the different cut-off points on the post-test likelihood of disease for each patient is shown in Table III.

The selection of a cut-off point of 280 U/l gave us more information on patient A (whose serum CK level was 400 U/l). Patient A's probability of infarction rose from a pretest value of 64% to a post-test value of 99%. We did not learn much about patient B (whose serum CK level was 30 U/l), but the change in likelihood was small for this patient, from 64% to 51%. However, we do learn more when the cut-off point is 40 U/l. Now the likelihood of infarction has fallen from 64% to 2%, and we can effectively rule out this diagnosis.

So, we can gain more with more complex tables than with single, simpler tables. This method also helps us understand combinations of tests (Table IV). Table IV replicates Table III from part 3 of our series. Once again, with the more complex table we see a greater change between the pretest and post-test likelihood of disease for patients with extreme test results (i.e., both positive or both negative). We will ex-

**Table III—Effect of different cut-off points on the post-test likelihood of disease***

| Serum CK level, U/l | Myocardial infarct: Present | Myocardial infarct: Absent | Post-test likelihood of disease |
|---|---|---|---|
| ≥ 280 | 97 (a) | 1 (b) | Patient A: $\frac{a}{a+b}=\frac{97}{98}=99\%$ |
| < 280 | 133 (c) | 129 (d) | Patient B: $\frac{c}{c+d}=\frac{133}{262}=51\%$ |

| Serum CK level, U/l | Myocardial infarct: Present | Myocardial infarct: Absent | Post-test likelihood of disease |
|---|---|---|---|
| ≥ 80 | 215 (a) | 16 (b) | Patient A: $\frac{a}{a+b}=\frac{215}{231}=93\%$ |
| < 80 | 15 (c) | 114 (d) | Patient B: $\frac{c}{c+d}=\frac{15}{129}=12\%$ |

| Serum CK level, U/l | Myocardial infarct: Present | Myocardial infarct: Absent | Post-test likelihood of disease |
|---|---|---|---|
| ≥ 40 | 228 (a) | 42 (b) | Patient A: $\frac{a}{a+b}=\frac{228}{270}=84\%$ |
| < 40 | 2 (c) | 88 (d) | Patient B: $\frac{c}{c+d}=\frac{2}{90}=2\%$ |

*The pretest likelihood of disease was 64%.

**Table II—Simple tables created from one complex table**

| Serum CK level, U/l | Myocardial infarct: Present | Myocardial infarct: Absent | Cut-off point; serum CK level (U/l): ≥ 280 | ≥ 80 | ≥ 40 | ≥ 1 |
|---|---|---|---|---|---|---|
| ≥ 280 | 97 | 1 | 97 (a) / 1 (b) | | | |
| 80–279 | 118 | 15 | | 215 (a) / 16 (b) | 228 (a) / 42 (b) | |
| 40–79 | 13 | 26 | 133 (c) / 129 (d) | | | 230 (a) / 130 (b) |
| 1–39 | 2 | 88 | | 15 (c) / 114 (d) | 2 (c) / 88 (d) | 0 (c) / 0 (d) |
| | 230 | 130 | | | | |
| | Sensitivity = $\frac{a}{a+c}$ | | 42% | 93% | 99% | 100% |
| | Specificity = $\frac{d}{b+d}$ | | 99% | 88% | 68% | 0% |

amine multiple diagnostic tests in greater detail in part 5 of our series.

As you have probably already noted, changes or differences in the pretest likelihood of disease are managed just as they were with the simple tables. Something that you may not have recognized, however, is that you now know a bit about ROC (receiver operating characteristic)* curves.

## ROC curves

Look again at Table II, where we have listed the sensitivity (true-positive [TP] rate) and specificity (true-negative [TN] rate) for each cut-off point.

The TN rate is, of course, d/(b + d) or true negatives/(false positives + true negatives). We could therefore generate a complementary rate, b/(b + d), and call it the false-positive (FP) rate. And, of course, for any table the FP rate plus the TN rate would sum to 100%. When we add the FP rates to Table II we get Table V. Since the FP rate = 100% — the TN rate, the TP and FP rates rise and fall together.

We can now draw an ROC curve, which is simply a graph of the pairs of TP rates and the FP rates that correspond to each possible cut-off point for the diagnostic test result (Fig. 2). For example, the point labelled "≥ 280" indicates the TP rate of 42% and the FP rate of 1% that apply when we use a cut-off point of 280 U/l or greater.

Fig. 2 provides a picture of the implications of using different cut-off points; such ROC curves have some interesting properties. For example, the upper left-hand corner of Fig. 2 denotes a perfect diagnostic test: a TP rate of 1.00 (all patients with the target disorder are detected) and an FP rate of 0 (no one without the target disorder is falsely labelled). It follows that the point on an ROC curve that is *closest* to the upper left-hand corner is the best cut-off point in terms of making the fewest mistakes (that is, it creates the smallest total number of false positives plus false negatives). You can confirm this by relating Fig. 2 to Table II. In the former, the closest point to the upper left-hand corner is for the cut-off point at 80 U/l or greater, and in Table II the sum of cells b and c (that is, the mistakes) for this cut-off point is 31; for any other cut-off point in Table II the sum of cells b and c is greater than 31. You also may want to look again at Fig. 3 in part 2 of our series, noting the point at which the

*This acronym comes from the early days of radar and other imaging strategies, when interpreters had to distinguish between "signals" caused by airplanes and "noise" from other sources.

**Table V—True-positive (TP), true-negative (TN) and false-positive (FP) rates at different cut-off points**

| Rate (%) | Cut-off point; serum CK level (U/l) | | | |
|---|---|---|---|---|
| | ≥ 280 | ≥ 80 | ≥ 40 | ≥ 1 |
| TP | 42 | 93 | 99 | 100 |
| TN | 99 | 88 | 68 | 0 |
| FP | 1 | 12 | 32 | 100 |

**Table IV—Two ways of combining the results of two independent diagnostic tests**

| Test results | Target disorder | | Test results | | | |
|---|---|---|---|---|---|---|
| | | | Both positive | | One or both positive | |
| | Present | Absent | | | | |
| Both positive | 630 | 12 | 630 (a) | 12 (b) | 970 (a) | 288 (b) |
| One positive | 340 | 276 | 370 (c) | 988 (d) | 30 (c) | 712 (d) |
| Neither positive | 30 | 712 | | | | |
| | | Sensitivity | 63% | | 97% | |
| | | Specificity | | 99% | | 71% |

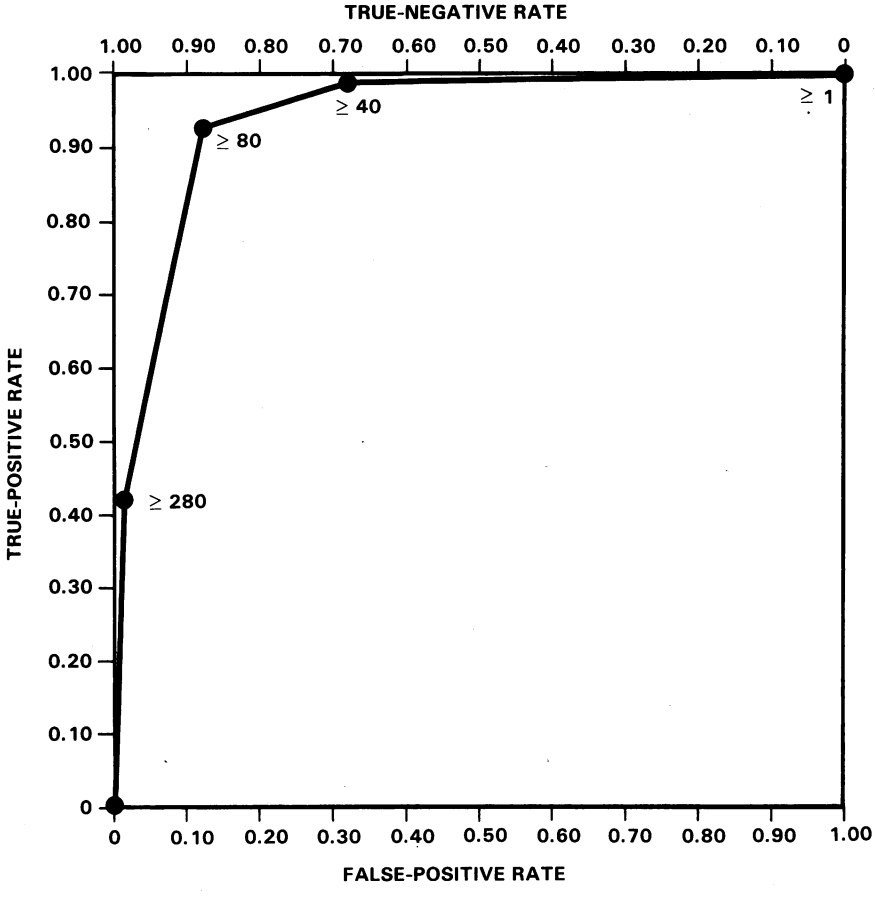


**FIG. 2—ROC (receiver operating characteristic) curve, showing serum CK levels (U/l) in patients with and without myocardial infarcts.**

serum CK levels in patients with and without infarcts cross one another. It is also 80 U/l.

Finally, the upper left-hand corner of Fig. 2 is also where the sum of the TP and TN rates attains its highest value. Of course, there is much more to picking the correct cut-off point than simply minimizing the sum of false positives and false negatives, as you will recall from our earlier discussions. You would seek the upper left-hand cut-off point only if your patient might suffer equally from false-positive or false-negative labelling. If false-positive labelling was going to be very harmful, you would select a lower cut-off point, which would minimize the FP rate. If, on the other hand, false-negative labelling was going to be highly dangerous, you would select a higher cut-off point, which would maximize the TP rate.

A more complex table and the resulting ROC curves can be used to compare the usefulness of two different signs, symptoms or diagnostic tests for the same target disorder. All you need to do is plot a separate ROC curve for each of them: the one that lies farthest to the "northwest" is the more accurate.*

## Summary


A more complex table is especially useful when a diagnostic test produces a wide range of results and your patient's levels are near one of the extremes. The following guidelines will be useful:

- Identify the several cut-off points that could be used.
- Fill in a complex table along the lines of Table I, showing the numbers of patients at each level who have and do not have the target disorder.
- Generate a simple table for *each* cut-off point, as in Table II, and determine the sensitivity (TP rate) and specificity (TN rate) at each of them.
- Select the cut-off point that makes the most sense for your patient's test result and proceed as in parts 2 and 3 of our series.
- Alternatively, construct an ROC curve by plotting the TP and FP rates that attend each cut-off point.

If you keep your tables and ROC curves close at hand, you will gradually accumulate a set of very useful guides. However, if you looked very hard at what was happening, you will probably have noticed that they are not very useful for patients whose test results fall in the middle zones, or for those with just one positive result of two tests; the post-test likelihood of disease in these patients lurches back and forth past 50%, depending on where the cut-off point is. We will show you how to tackle this problem in part 5 of our series. It involves some maths, but you will find that its very powerful clinical application can be achieved with a simple nomogram or with some simple calculations.


*Stated more formally, the sign, symptom or laboratory test result whose ROC curve encloses (below and to the right) the largest area is the most accurate. If you want to learn more about ROC curves you might start with an article by McNeil and colleagues.[2]

## References


1. SMITH AF: Diagnostic value of serum-creatine-kinase in a coronary-care unit. *Lancet* 1967; 2: 178–182
2. MCNEIL BJ, KEELER E, ADELSTEIN SJ: Primer on certain elements of medical decision making. *N Engl J Med* 1975; 293: 211–215


**Clue to correct diagnosis**

*When laboratory reports conflict with clinical judgment, don't discard the latter before repeating the laboratory tests.*

*—Paul Reznikoff (1896–)*

# Clinical Epidemiology Rounds

## Interpretation of diagnostic data: 5. How to do it with simple maths

DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS, MCMASTER UNIVERSITY, HAMILTON, ONT.

In parts 2 to 4 of our series on interpreting diagnostic data we showed you how to use simple or complex tables. This article presents quite a different way of interpreting diagnostic data — the use of simple maths. To ease the transition we will use the same sample of patients as in parts 2 and 4.

### Using simple maths

*Case presentation*

In 360 patients consecutively admitted to a coronary care unit (CCU) blood was drawn for measurements of the serum creatine kinase (CK) level at the time of admission and the next two mornings.[1] A clinician who was "blind" to the CK measurements reviewed the patients' electrocardiograms (ECGs), clinical records and autopsy reports, and decided that 230 of them had had a myocardial infarction and 130 had not.

Reprint requests to: Dr. David L. Sackett, Rm. 3V43E, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton, Ont. L8N 3Z5

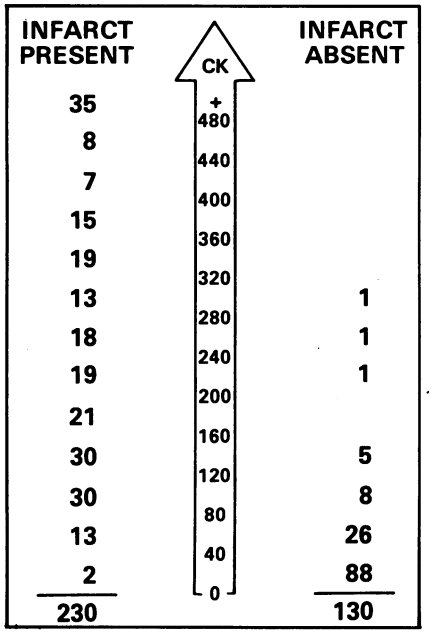


**FIG. 1—Maximum serum creatine kinase (CK) levels (U/l) in patients with and without myocardial infarcts.**

*Comment*

Fig. 1 shows the maximum serum CK levels* for the 230 patients who had an infarct and the 130 patients who did not. In Table I, which should look familiar to you by now, these figures are converted to a simple table.

Such tables are all very nice, but you must admit that they are cumbersome. Do you really want to carry graph paper and scratch pads wherever you go, and figure out a new table for every patient you see? Fortunately, once you have mastered the tables we used in parts 2 to 4, you are ready for a great leap forward. It involves more calculations initially, but once you become familiar with them you should be able to arm yourself with some simple tables (like the tables of "normal" values you probably carry now) and be able to come up with a patient's post-test probability of disease with just a simple nomogram or simple mental arithmetic.

*Since laboratory methods differ, this distribution of serum CK levels may not apply at your institution.

**Table I—Conversion of raw data to a simple table**

| Serum creatine kinase (CK) test result | Myocardial infarct: Present | Myocardial infarct: Absent | |
|---|---|---|---|
| Positive (level ≥ 80 U/l) | 215 (a) | 16 (b) | 231 (a + b) |
| Negative (level < 80 U/l) | 15 (c) | 114 (d) | 129 (c + d) |
| | 230 (a + c) | 130 (b + d) | 360 |

### Discussion

*The likelihood ratio*

We start by generating a new

This is the fifth of six articles (the first four appeared in the Sept. 1, Sept. 15, Oct. 1 and Oct. 15, 1983 issues of the Journal) that focus on the strategies and tactics for interpreting diagnostic data, both the clinical data from the history and physical examination and the paraclinical data from the clinical laboratories, the radiology department and the surgical pathology service. The last article in the series will appear in the next issue of the Journal.

index of how good a diagnostic test is. This index, which is called a "likelihood ratio",* contrasts the proportions of patients with and without the target disorder who have a given level of a diagnostic test result. By "given level" we mean the presence (or absence) of a sign or symptom, or any of the levels of a laboratory test result, such as those in Fig. 1.

Thus, *the likelihood ratio expresses the odds that a given diagnostic test result would be expected in a patient with (as opposed to one without) the target disorder.*

Let us calculate some likelihood ratios and learn their properties. Table II, constructed by adding the likelihood ratios to Table I, shows that the likelihood ratio for a positive test result (a serum CK level of 80 U/l or higher) is 7.75; in other words, this serum CK level is 7.75 times as likely to come from patients with infarcts as from those without. Let's take a closer look at the proportions that make up this likelihood ratio. The first likelihood is a/(a + c) = 215/230 = 0.93, our old friend sensitivity (the TP rate), and the second likelihood is b/(b + d) = 16/130 = 0.12, which is 1 − specificity (the FP rate). The likelihood ratio for a negative test result (a serum CK level of less than 80 U/l) is 0.08; in other words, this serum CK level is only about 1/10th (8/100ths to be exact) as likely to come from patients with infarcts as from those without. The first likelihood is c/(a + c) = 15/230 = 0.07, the complement of sensitivity (the false-negative [FN] rate), and the second one is d/(b + d) = 114/130 = 0.88, the specificity (the true-negative [TN] rate).

Likelihood ratios have three properties, which, when combined, form a very powerful diagnostic strategy. First, because the likelihoods that make up the likelihood ratio are calculated vertically, like sensitivity and specificity, the ratios need not change with changes in the prevalence (pretest probability) of the target disorder. In fact, they may be much more stable than sensitivity or specificity with changes in prevalence because of their second property — the option of their being calculated for *several* levels of the sign, symptom or laboratory test result, rather than for just the two levels we have worked with up to now.

Table III shows the second property in action. In Table III four levels of the CK test result are considered. However, rather than collapsing these four levels into individual two-by-two dichotomies, as we did in parts 2 and 3 of our series, we can preserve all four levels and assign a likelihood ratio to each one. Note how the range of likelihood ratios has dramatically widened, from 97-fold (0.08 to 7.75) in Table II to 4200-fold (0.01 to 42) in Table III. The clinical information from the diagnostic test result is therefore greatly increased.†

Let's look again at the 99% cut-off points of maximum and minimum serum CK levels (Fig. 2). The highest CK level (≥ 280 U/l) corresponds to the rule-in cut-off point y that we developed in the first part of our series, and the lowest level (1 to 39 U/l) corresponds to the rule-out cut-off point x. By using simple maths we can now distinguish between the subgroups of patients whose CK levels lie between the rule-in and rule-out cut-off points.

A comparison of Fig. 2 and Table III suggests a very pragmatic, two-step approach to evaluating the results of diagnostic tests. We can start by identifying the 99% cut-off points for a diagnostic test and rule in or rule out the diagnosis if a patient's test result lies beyond one of these points. If, however, the patient's test result lies between the cut-off points we can use the likelihood ratio to extract as much diagnostic information from the test as possible.

We have already said that likelihood ratios are more stable than sensitivity or specificity when the prevalence changes. If the mix of patients with either a mild or a severe form of the target disorder varies when the prevalence of the disorder varies, the sensitivity and specificity as well as the predictive values will change. However, be-

*A "likelihood" is analogous to a "probability", a "proportion" or a "rate". Thus, sensitivity (the true-positive [TP] rate) is the likelihood that patients who have the target disorder will have a positive test result. A "likelihood ratio" is simply the TP rate divided by the false-positive (FP) rate. Thus, a likelihood ratio could just as easily be called a "rate ratio", but the former term is more common.

†Of course, the information was there all the time. The likelihood ratio simply *preserves* the diagnostic information that is often lost with other methods of interpretation.

**Table II—How likelihood ratios are generated**

| Serum CK test result | Myocardial infarct: Present — No. | Present — Likelihood | Absent — No. | Absent — Likelihood | Likelihood ratio |
|---|---|---|---|---|---|
| Positive | 215 (a) | $\frac{a}{a+c} = \frac{215}{230} = 0.93$ | 16 (b) | $\frac{b}{b+d} = \frac{16}{130} = 0.12$ | $\frac{0.93}{0.12} = 7.75$ |
| Negative | 15 (c) | $\frac{c}{a+c} = \frac{15}{230} = 0.07$ | 114 (d) | $\frac{d}{b+d} = \frac{114}{130} = 0.88$ | $\frac{0.07}{0.88} = 0.08$ |
| | 230 (a + c) | | 130 (b + d) | | |

cause likelihood ratios can be generated for narrow "slices" of a diagnostic test result, they are less susceptible to such changes.

The third property of the likelihood ratio is the most delightful; it can be used in a very powerful way to shorten a list of diagnostic hypotheses because the pretest "odds" (the ratio of the probabilities for and against a diagnosis) of the target disorder × the likelihood ratio for the diagnostic test result = the post-test "odds" for the target disorder. If you start from your clinical estimate of the odds that your patient has a certain target disorder and then carry out a diagnostic test and apply the likelihood ratio that corresponds to your patient's test result, you can calculate a new, post-test odds of the target disorder.

Suppose you are doing a work-up on a man with chest pain, and you judge that the probability that he has had a myocardial infarction is about 60% (odds of 60:40 or 1.5:1). Suppose further that his initial serum CK level is 180 U/l. A quick look at Table III will confirm that the likelihood ratio for this CK level is 4.2 (i.e., 4.2:1); now you can apply the third property of the likelihood ratio: 1.5:1 (pretest odds) × 4.2:1 (likelihood ratio) = 6.3:1 (post-test odds). The post-test odds correspond to a probability of slightly more than 85%, so your tentative diagnosis is firming up nicely.

This example emphasizes both the diagnostic power of and a major drawback to the likelihood ratio strategy. Although it can help us get the most out of the diagnostic tests we use, the need to switch back and forth between probabilities and odds is off-putting at best and frequently scares clinicians. We suggest two solutions.

*Nomogram:* The first solution is to use a nomogram (adapted from that proposed by Fagan[2]), which obviates the need to switch back and forth between probabilities and odds (Fig. 3). The pretest and post-test odds are already converted to percentage probabilities, so we need not trip over the calculations. Let's go back to the 60% probability for our patient with chest pain. Anchor a ruler at the pretest probability of 60%, then rotate the ruler until it lines up with the likelihood ratio of 4.2. If you look along the ruler to the right you'll see that the post-test probability is about 86%. Hence, you have figured out the post-test likelihood with no maths or conversions between probabilities and

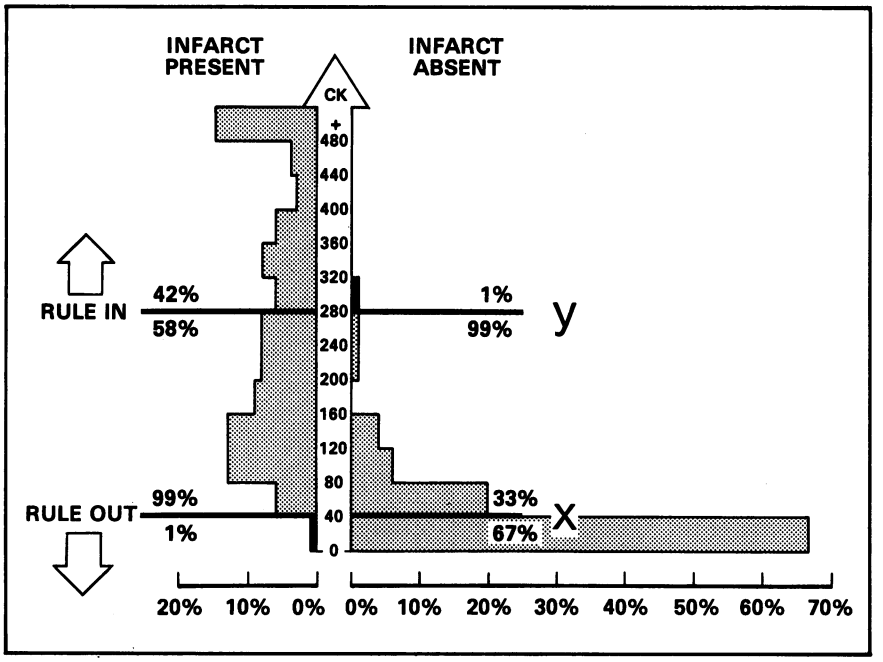


**FIG. 2—Cut-off points (99%) for maximum and minimum serum CK levels for diagnosis of myocardial infarct.**

**Table III—Likelihood ratios for several levels of a diagnostic test result**

| Serum CK level, U/l | Myocardial infarct: Present — No. | Present — Likelihood | Absent — No. | Absent — Likelihood | Likelihood ratio |
|---|---|---|---|---|---|
| ≥ 280 | 97 | $\frac{97}{230} = 0.42$ | 1 | $\frac{1}{130} = 0.01$ | $\frac{0.42}{0.01} = 42$ |
| 80–279 | 118 | $\frac{118}{230} = 0.51$ | 15 | $\frac{15}{130} = 0.12$ | $\frac{0.51}{0.12} = 4.2$ |
| 40–79 | 13 | $\frac{13}{230} = 0.06$ | 26 | $\frac{26}{130} = 0.20$ | $\frac{0.06}{0.20} = 0.30$ |
| 1–39 | 2 | $\frac{2}{230} = 0.01$ | 88 | $\frac{88}{130} = 0.68$ | $\frac{0.01}{0.68} = 0.01$ |
| | 230 | | 130 | | |

odds. If you find this solution attractive, make a photocopy of the nomogram to keep on hand.

*Simple conversion:* The second solution is to learn how to convert pretest probabilities to pretest odds. Simply divide the probability by its complement — that is, pretest probability/(1 − pretest probability) = pretest odds. Thus, for our patient with a pretest probability of myocardial infarction of 60% the pretest odds are 0.60/(1 − 0.60) = 0.60/0.40 = 1.50. In Table IV we have calculated the pretest odds for several clinically relevant pretest probabilities.

How do we get back to the post-test probability? As you will recall, the likelihood ratio for our patient's serum CK level of 180 U/l was 4.2, and his post-test odds for having an infarct were 6.3:1. The post-test odds are converted back to the post-test probability as follows: post-test odds/(post-test odds + 1) = post-test probability. Thus, with post-test odds of 6.3 the post-test probability is 6.3/(6.3 + 1) = 6.3/7.3 = 0.86, or 86%. In Table V we have calculated the post-test probabilities for a range of post-test odds.

As more clinicians recognize the power of the likelihood ratio strategy, they are asking subspecialists, radiologists, pathologists and laboratory specialists for precise likelihood ratios for various signs, symptoms and laboratory test results. In Table VI we have calculated likelihood ratios from data in several reports.[3-10]

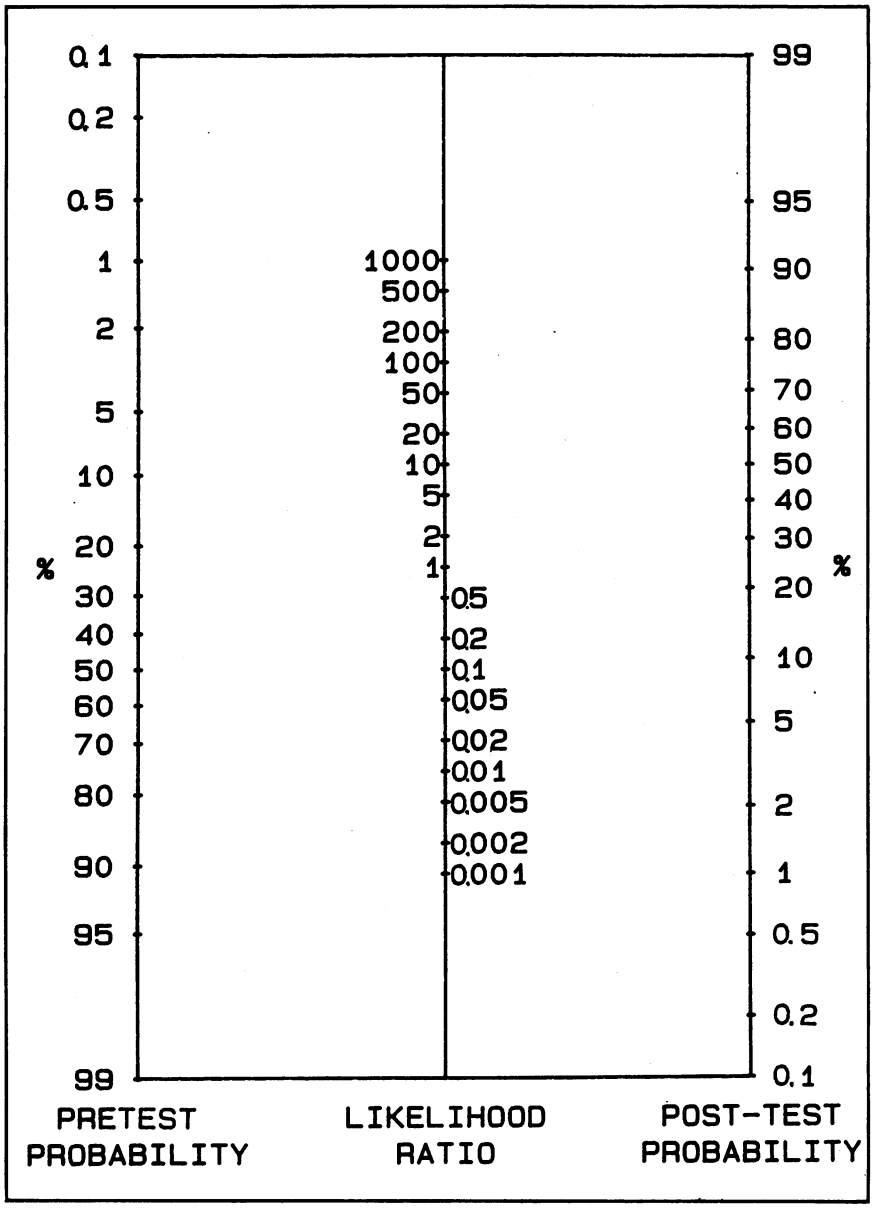


**FIG. 3—Nomogram for applying likelihood ratios. Adapted from Fagan.[2]**

*Tying it all together*

It may be hard for you to see how we got here from where we were in part 2 of our series. If so, let's go back and redo the example in part 2 with our new strategy (Table VII). The pretest probability of myocardial infarction was 64%, and the positive predictive value of a serum CK level of 80 U/l or greater was 93%. Would we get the same answer with

**Table IV—Converting pretest probabilities to odds**

| Pretest probability | Probability/(1 − probability) | Pretest odds |
|---|---|---|
| 0.1% (0.001) | 0.001/0.999 | 0.001 |
| 1% (0.01) | 0.01/0.99 | 0.01 |
| 2% (0.02) | 0.02/0.98 | 0.02 |
| 3% (0.03) | 0.03/0.97 | 0.03 |
| 4% (0.04) | 0.04/0.96 | 0.04 |
| 5% (0.05) | 0.05/0.95 | 0.05 |
| 10% (0.1) | 0.1/0.9 | 0.11 |
| 20% (0.2) | 0.2/0.8 | 0.25 |
| 30% (0.3) | 0.3/0.7 | 0.43 |
| 40% (0.4) | 0.4/0.6 | 0.67 |
| 50% (0.5) | 0.5/0.5 | 1.0 |
| 60% (0.6) | 0.6/0.4 | 1.5 |
| 70% (0.7) | 0.7/0.3 | 2.3 |
| 80% (0.8) | 0.8/0.2 | 4.0 |
| 90% (0.9) | 0.9/0.1 | 9.0 |
| 95% (0.95) | 0.95/0.05 | 19.0 |
| 99% (0.99) | 0.99/0.01 | 99.0 |

**Table V—Converting post-test odds to probabilities**

| Post-test odds | Odds/(odds + 1) | Post-test probability |
|---|---|---|
| 0.001 | 0.001/1.001 | 0.001 (0.1%) |
| 0.01 | 0.01/1.01 | 0.01 (1%) |
| 0.02 | 0.02/1.02 | 0.02 (2%) |
| 0.03 | 0.03/1.03 | 0.03 (3%) |
| 0.04 | 0.04/1.04 | 0.04 (4%) |
| 0.05 | 0.05/1.05 | 0.05 (5%) |
| 0.1 | 0.1/1.1 | 0.09 (9%) |
| 0.2 | 0.2/1.2 | 0.17 (17%) |
| 0.3 | 0.3/1.3 | 0.23 (23%) |
| 0.4 | 0.4/1.4 | 0.29 (29%) |
| 0.5 | 0.5/1.5 | 0.33 (33%) |
| 0.6 | 0.6/1.6 | 0.38 (38%) |
| 0.7 | 0.7/1.7 | 0.41 (41%) |
| 0.8 | 0.8/1.8 | 0.44 (44%) |
| 0.9 | 0.9/1.9 | 0.47 (47%) |
| 1.0 | 1/2 | 0.5 (50%) |
| 2.0 | 2/3 | 0.67 (67%) |
| 3.0 | 3/4 | 0.75 (75%) |
| 4.0 | 4/5 | 0.8 (80%) |
| 5.0 | 5/6 | 0.83 (83%) |
| 10.0 | 10/11 | 0.91 (91%) |
| 20.0 | 20/21 | 0.95 (95%) |
| 30.0 | 30/31 | 0.97 (97%) |
| 40.0 | 40/41 | 0.98 (98%) |
| 50.0 | 50/51 | 0.98 (98%) |
| 100.0 | 100/101 | 0.99 (99%) |

the likelihood ratio strategy? We can use either the nomogram or simple conversion between the pretest probabilities and odds.

With the nomogram, anchor your ruler at the pretest probability of 64%, then rotate it to a likelihood ratio of 7.75. If you look along the ruler to the post-test probability you will see that it lies just short of 95%, which is in good agreement with the positive predictive value of 93% in Table VII.

To use simple conversion you have to generate the pretest odds from the prevalence of 64% with the formula probability/(1 − probability) = 64%/36% = 1.78:1. Next, you have to generate the likelihood ratio for a serum CK level of 80 U/l or greater. Sensitivity is the same as the TP rate, which is 93%. The FP rate is 100% − specificity = 100% − 88% = 12%. Thus, the likelihood ratio is TP rate/FP rate = 93%/12% = 7.75. Now, multiply the pretest odds by the likelihood ratio for a CK level of 80 U/l or greater: 1.78 × 7.75 = 13.8:1. Finally, we can convert the post-test odds back to probability with the formula odds/(odds + 1) = 13.8/14.8 = 0.93, which is in perfect agreement with the 93% positive predictive value shown in Table VII.

Thus, all these methods produce the same results. In fact, you could obtain about the same answer in your head if you round off the

**Table VI—Likelihood ratios for various signs, symptoms and laboratory test results**

| Report | Diagnostic test | Target disorder (and gold standard*) | Test result | Likelihood ratio |
|---|---|---|---|---|
| Diamond et al[3] | Symptoms of typical angina | Coronary artery stenosis (angiography or autopsy) | Positive history | |
| | | | Men | 115 |
| | | | Women | 120 |
| | Symptoms of typical angina | Coronary artery stenosis (angiography or autopsy) | Positive history | |
| | | | Men | 14 |
| | | | Women | 15 |
| | Exercise electrocardiography | Coronary artery stenosis (angiography) | Nonsloping ST-segment depression (mm) | |
| | | | ≥ 2.5 | 39 |
| | | | 2.0–2.49 | 11 |
| | | | 1.5–1.99 | 4.2 |
| | | | 1.0–1.49 | 2.1 |
| | | | 0.05–0.99 | 0.92 |
| | | | < 0.05 | 0.23 |
| Singer[4] | Signs of deep-vein thrombosis (pain, warmth, colour change, induration or tenderness) or a difference in circumference greater than 3 cm | Proximal deep-vein thrombosis (venography) | Four signs or more, or increased circumference | 2.6 |
| | | | Fewer than four signs and no difference in circumference | 0.15 |
| Newman et al[5] | Radionuclide angiocardiography | Coronary artery stenosis (angiography) | Positive | 3.6 |
| | | | Negative | 0.05 |
| Hessel et al[6] | Ultrasonography | Pancreatic disease (biopsy, autopsy or clinical course) | Abnormal | |
| | | | Definitely | 5.6 |
| | | | Probably | 2.1 |
| | | | Possibly | 0.95 |
| | | | Normal | |
| | | | Probably | 0.43 |
| | | | Definitely | 0.32 |
| | Computerized tomography scan | Pancreatic disease (biopsy, autopsy or clinical course) | Abnormal | |
| | | | Definitely | 26 |
| | | | Probably | 4.8 |
| | | | Possibly | 0.35 |
| | | | Normal | |
| | | | Probably | 0.32 |
| | | | Definitely | 0.11 |
| Hawkins et al[7] | Test for HLA-B27 histocompatibility antigen | Ankylosing spondylitis | Positive | 15 |
| | | | Negative | 0.11 |
| National Cancer Institute of Canada/ American Cancer Society[8] | Measurement of serum carcinoembryonic antigen (CEA) level | Colorectal cancer (biopsy or operation) | CEA level (ng/ml) | |
| | | | ≥ 20 | 3.5 |
| | | | 10–19.9 | 2.3 |
| | | | 5–9.9 | 1.4 |
| | | | 1–4.9 | 0.94 |
| | | | < 1 | 0.46 |
| Boyd et al[9] | Sputum smears | Tuberculosis (culture) | Positive | 31 |
| | | | Negative | 0.79 |
| Mundt et al[10] | Antibody-coated bacteria assay | Upper urinary tract infection (bilateral ureteral catheterization or bladder wash-out) | Positive | 3.6 |
| | | | Negative | 0.22 |

*Diagnostic tests to determine presence or absence of target disorder.

figures. The pretest odds of 64%/36% is about 60%/40%, which converts to a pretest odds of 1.5:1 (Table IV). The likelihood ratio of 93%/12% is about 90%/10%, or 9:1. The product of the pretest odds and the likelihood ratio then becomes 1.5 × 9 = 13.5, or about 14:1. Again, from Table V you can see that this result lies between 91% (for 10:1) and 95% (for 20:1). Thus, you can do quite well by rounding off the pretest odds and the likelihood ratio and calculating it all in your head, at least after a bit of practice.

This example also shows that you can use simple maths even when the diagnostic test results have a single cut-off point. If you know the sensitivity and specificity you also know the TP rate and 100% − the FP rate.

However, you are always in a better position if likelihood ratios are available for several different levels of the diagnostic test result, because the degree of abnormality in a test result can only be estimated when there are many levels, each with its own likelihood ratio. For example, in Table VI 1.0 to 1.49 mm of ST-segment depression on the exercise ECG has a likelihood ratio for severe coronary artery stenosis of only 2:1. However, the ratio jumps to 11:1 for a depression of 2.0 to 2.49 mm and to 39:1 for a depression of greater than 2.5 mm.

Now that you are getting comfortable at doing these simple maths, we will let you in on a secret: you have been applying a modification of "Bayes' theorem", an approach to diagnosis that is usually presented with all sorts of arcane symbols and complex formulas. However, we haven't shown any of these symbols or formulas because you just don't need them.

The separate consideration of pretest probabilities on the one hand and likelihood ratios on the other underscores an issue we raised in our series on clinical disagreement.[11,12] One strategy for minimizing the risk of clinical disagreement and error is to "blind" the assessment of raw diagnostic data. That is, those who interpret x-ray films, ECGs, biopsy results and so on should, at least at the time of their initial interpretations, simply generate the likelihood ratios for the target disorders that might produce such an image or tracing, without having other information about the patient. The diagnostician who ordered the test can then apply likelihood ratios to the appropriate pretest and post-test probabilities.[13]

The "blind" approach may resolve the diagnostic problem from the start and, at any rate, will avoid several sorts of bias. On the other hand, a knowledge of the diagnostic hypothesis may be crucial if specific radiologic views, histologic stains and other variations in diagnostic testing are required for certain target disorders. In this instance negotiation between the diagnostician and the laboratory specialist is essential. None the less, keeping the pretest probabilities and likelihood ratios separate whenever possible will both reduce the risk of clinical disagreement and error and sharpen our diagnostic power.

### *Multiple diagnostic tests*

The likelihood ratios strategy also allows you to carry out sequences of diagnostic tests. That is, with this strategy the post-test probability for one test becomes the pretest probability for a second, independent diagnostic test.

For example, let's say a 45-year-old woman calls your receptionist saying that she's had intermittent chest pain for 1 month. The receptionist gives her an appointment to see you. Like a good hypotheticodeductive diagnostician, as soon as you hear the patient's chief complaint you form a mental "short list" of possible explanations: (a) there's something in the chest wall, possibly related to emotional stress; (b) there's something in the esophagus or upper gastrointestinal tract; (c) she has coronary artery disease; or (d) none of the above.

**Table VII—Results of serum CK test in patients with and without myocardial infarcts**

| Serum CK test result | Myocardial infarct: Present | Myocardial infarct: Absent | | |
|---|---|---|---|---|
| Positive | 215 (a) | 16 (b) | 231 (a + b) | Positive predictive value (predictive value of a positive test or post-test likelihood or posterior probability of disease) $= \frac{a}{a+b} = \frac{215}{231} = 93\%$ |
| Negative | 15 (c) | 114 (d) | 129 (c + d) | Negative predictive value (predictive value of a negative test or post-test likelihood or posterior probability of no disease) $= \frac{d}{c+d} = \frac{114}{129} = 88\%$ |
| | 230 (a + c) Sensitivity $= \frac{a}{a+c} = \frac{215}{230} = 93\%$ | 130 (b + d) Specificity $= \frac{d}{b+d} = \frac{114}{130} = 88\%$ | 360 | Prevalence (pretest likelihood or prior probability of disease) $= \frac{a+c}{a+b+c+d} = \frac{230}{360} = 64\%$ |

When the patient walks through your door her pretest probability of coronary artery disease is only about 0.01, so her pretest odds are 0.01/(1 − 0.01) = 0.01/0.99 = 0.01:1. As you talk with her, however, you learn that her pain is substernal, more "heavy" than painful, radiates down her left inner arm, is brought on by physical effort and is relieved by 4 to 6 minutes of rest. In fact, she has typical angina; the likelihood ratio for the presence of this symptom is enormous — that is, over 100 (Table VI). So the first diagnostic test, careful history-taking, produces a large change in the likelihood of coronary artery disease, the calculation being 0.01 (pretest odds) × 100 (likelihood ratio) = 1:1 (post-test odds). The post-test odds can be converted to a probability of 50% by calculating odds/(odds + 1) = 1/2 = 50%, or by consulting Table V. Alternatively, you could have done all the calculations with the nomogram.

Having boosted the patient's post-test probability of coronary artery disease to 40% to 60%, a range at which most diagnostic tests are most helpful, you suggest that an exercise ECG be done. The ECG shows 2.2 mm of ST-segment depression. This degree of depression has a likelihood ratio of 11.1:1 (Table VI).

From both diagnostic tests you can now determine the post-test odds of coronary artery disease with your standard formula, as follows: from the history-taking the post-test odds are 1:1 (0.01:1 [pretest odds] × 100:1 [likelihood ratio]) and from the exercise ECG the post-test odds are 10:1 (1:1 [pretest odds] × 10:1 [likelihood ratio]). With the post-test odds of 10:1 for the second test, the post-test probability is now 10/11, or 91%; therefore, the presence of coronary artery disease is confirmed.*

You would get the same result with the nomogram. However, when you have the post-test probability from the first test be careful to reanchor your ruler at the pretest probability for the second test.

This example shows us something else. Look at the relative size of the likelihood ratios for the brief, immediate and relatively inexpensive history-taking and the longer, delayed and relatively expensive exercise ECG. There is no contest. The likelihood ratios for key points of the history and physical findings for both coronary artery disease and most other target disorders are mammoth and usually dwarf those derived from most excursions through high technology.

Also, use of the two independent diagnostic tests for the same target disorder now becomes much less cumbersome. The likelihood ratios for the different test results can be rewritten as shown in the top panel of Table VIII. Suppose that a patient who is judged to have a pretest probability of 10% for the target disorder has positive results of both tests. The calculations are in the bottom panel of the table; you can confirm the results with the nomogram. Note that the post-test odds for test A (1.554) are the same as the pretest odds for test B. Note also how very "robust" the strategy is: even if we round off the likelihood ratio from test A from 14 to 15 and

*Note also that if the patient had had less than 0.5 mm of ST-segment depression the likelihood ratio would be 0.23, and 1 × 0.23 = 0.23:1 = 19%. Thus, almost any result of the exercise ECG would have been enormously useful clinically.

**Table VIII—Likelihood ratios for two independent tests for the same target disorder**

| Test | Sensitivity | 1 − specificity | Likelihood ratio |
|---|---|---|---|
| A | 0.70 | 1 − 0.95 = 0.05 | $\frac{0.70}{0.05}$ = 14 |
| B | 0.90 | 1 − 0.75 = 0.25 | $\frac{0.90}{0.25}$ = 3.6 |

$$\frac{\text{Pretest probability}}{1 - \text{pretest probability}} = \frac{0.10}{0.90} = 0.111 = \text{pretest odds}$$

| Type of calculation | Odds before test A | × | Likelihood ratio with test A | = | Odds after test A | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Odds before test B | × | Likelihood ratio with test B | = | Odds after tests A and B |
| "Proper" | 0.111 | × | 14 | = | 1.554 | × | 3.6 | = | 5.594 |
| "Rough" | 0.1 | × | 15 | = | 1.5 | × | 4 | = | 6 |

$$\frac{\text{Post-test odds}}{\text{Post-test odds} + 1}: \frac{5.594}{6.594}\ (\text{"proper"}) = 0.85;\ \text{or}\ \frac{6}{7}\ (\text{"rough"}) = 0.86$$

that from test B from 3.6 to 4, we get the same answer for the final post-test odds. We could have reversed the order of the two tests. For example, the rounded-off calculation for test B is $0.1 \times 4 = 0.4$, and that for test A is then $0.4 \times 15 = 6$, and $6/7 = 0.86$.

The sequence of tests can be as long as we want. If the tests are independent the final post-test probability will be true. If the tests are of the very common *convergent* sort (i.e., test A's result is more likely to be positive when test B's result is positive than when the latter is negative), the final post-test probability will overestimate the patient's true likelihood of disease. Conversely, if the tests are of the rare *divergent* variety (i.e., test A's result is more likely to be negative when test B's result is positive than when the latter is negative), the final post-test probability will underestimate the patient's true likelihood of disease.

### Summary

The use of simple maths with the likelihood ratio strategy fits in nicely with our clinical views. By making the most out of the entire range of diagnostic test results (i.e., several levels, each with its own likelihood ratio, rather than a single cut-off point and a single ratio) and by permitting us to keep track of the likelihood that a patient has the target disorder at each point along the diagnostic sequence, this strategy allows us to place patients at an extremely high or an extremely low likelihood of disease. Thus, the numbers of patients with ultimately false-positive results (who suffer the slings of labelling and the arrows of needless therapy) and of those with ultimately false-negative results (who therefore miss their chance for diagnosis and, possibly, efficacious therapy) will be dramatically reduced.

The following guidelines will be useful in interpreting signs, symptoms and laboratory tests with the likelihood ratio strategy:

- Seek out, and demand from the clinical or laboratory experts who ought to know, the likelihood ratios for key symptoms and signs, and several levels (rather than just the positive and negative results) of diagnostic test results.
- Identify, when feasible, the logical sequence of diagnostic tests.
- Estimate the pretest probability of disease for the patient, and, using either the nomogram or the conversion formulas, apply the likelihood ratio that corresponds to the first diagnostic test result.
- While remembering that the resulting post-test probability or odds from the first test becomes the pretest probability or odds for the next diagnostic test, repeat the process for all the pertinent symptoms, signs and laboratory studies that pertain to the target disorder. However, these combinations may not be independent, and convergent diagnostic tests, if treated as independent, will combine to overestimate the final post-test probability of disease.

You are now far more sophisticated in interpreting diagnostic tests than most of your teachers. In the last part of our series we will show you some rather complex strategies that combine diagnosis and therapy, quantify our as yet nonquantified ideas about use, and require the use of at least a hand calculator.

### References


1. SMITH AF: Diagnostic value of serum-creatine-kinase in a coronary care unit. *Lancet* 1967; 2: 178–182
2. FAGAN TJ: Nomogram for Bayes's theorem (C). *N Engl J Med* 1975; 293: 257
3. 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
4. SINGER J: Value of clinical signs in diagnosis of deep venous thrombosis (C). *Lancet* 1980; 1: 1186
5. NEWMAN GE, RERYCH SK, UPTON MT, SABISTON DC JR, JONES RH: Comparison of electrocardiographic and left ventricular functional changes during exercise. *Circulation* 1980; 62: 1204–1211
6. HESSEL SJ, SIEGELMAN SS, MCNEILL BJ, SANDERS RC, ADAMS DF, ALDERSON PO, FINBERG HG, ABRAMS HL: A prospective evaluation of computed tomography and ultrasound of the pancreas. *Radiology* 1982; 143: 129–133
7. HAWKINS BR, DAWKINS RL, CHRISTIANSEN FT, ZILKO PJ: Use of the B27 test in the diagnosis of ankylosing spondylitis: a statistical evaluation. *Arthritis Rheum* 1981; 24: 743–746
8. A collaborative study of a test for carcinoembryonic antigen (CEA) in the sera of patients with carcinoma of the colon and rectum. A Joint National Cancer Institute of Canada/American Cancer Society Investigation. *Can Med Assoc J* 1972; 107: 25–33
9. BOYD JC, MARR JJ: Decreasing reliability of acid-fast smear techniques for detection of tuberculosis. *Ann Intern Med* 1975; 82: 489–492
10. MUNDT KA, POLK BF: Identification of site of urinary tract infections by antibody-coated bacteria assay. *Lancet* 1979; 2: 1172–1175
11. Department of clinical epidemiology and biostatistics, McMaster University, Hamilton, Ont.: Clinical disagreement: I. How often it occurs and why. *Can Med Assoc J* 1980; 123: 499–504
12. Idem: Clinical disagreement: II. How to avoid it and how to learn from one's mistakes. Ibid: 613–617
13. SCHWARTZ WB, WOLFE HJ, PAUKER SG: Pathology and probabilities: a new approach to interpreting and reporting biopsies. *N Engl J Med* 1981; 305: 917–923


## BOOKS

*continued from page 937*

**ESSENTIALS OF HUMAN METABOLISM.** The Relationship of Biochemistry to Human Physiology and Disease. 2nd ed. W.C. McMurray. 331 pp. Illust. J.B. Lippincott Co.; Harper & Row, Publishers, Inc., Philadelphia, 1983. $19.50 (US), paperbound. ISBN 0-06-141643-6

**ETHICS OF WITHDRAWAL OF LIFE-SUPPORT SYSTEMS.** Case Studies on Decision Making in Intensive Care. Contributions in Philosophy. Douglas N. Walton. 259 pp. Greenwood Press, Westport, Connecticut, 1983. $29.95 (US). ISBN 0-313-23752-2

**FUNCTIONAL GASTROINTESTINAL DISORDERS.** A Behavioral Medicine Approach. Paul R. Latimer. 177 pp. Illust. Springer Publishing Company, Inc., New York, 1983. $23.95 (US). ISBN 0-8261-4310-5

**INFANT NUTRITION.** A Study of Feeding Practices and Growth from Birth to 18 Months. David L. Yeung. A project of the H.J. Heinz Company of Canada Ltd. supported by the National Research Council of Canada. 184 pp. Illust. Canadian Public Health Association, Ottawa, 1983. $10, paperbound. ISBN 0-919245-18-8

**LABORATORY-ACQUIRED INFECTIONS.** History, Incidence, Causes and Prevention. C.H. Collins. 277 pp. Illust. Butterworths & Co. (Publishers) Ltd., Woburn, Massachusetts, 1983. $49.95 (US). ISBN 0-408-10650-6

**LITHIUM TREATMENT OF MANIC-DEPRESSIVE ILLNESS.** A Practical Guide. 2nd, revised ed. Mogens Schou. 49 pp. Illust. S. Karger AG, Basel, Switzerland, 1983. $9 (US), paperbound. ISBN 3-8055-3678-X

**MANAGEMENT OF LABOR.** Edited by Wayne R. Cohen and Emanuel A. Friedman. 362 pp. Illust. University Park Press, Baltimore, Maryland, 1983. $34.50 (US). ISBN 0-8391-1816-3

*continued on page 975*