DRAWING CONCLUSIONS FROM TEST RESULTS

MICHAEL W. KITA, MD
Vice President & Associate Medical Director
UNUM Life Insurance Company
Portland, ME

Some of the terminology surrounding the interpretation of laboratory tests is a little esoteric and subject to confusion. But for underwriters, medical directors and others who use such data, the basic concepts and probability principles are really fairly simple.

Let's take a hypothetical case. The applicant is a 50 year-old executive who applies to your company, Few-Hoops Life Insurance Company, for a $2,000,000 life policy. On his application, he notes a recent "false positive" treadmill. An Attending Physician's Statement also notes that treadmill result and reports that "he has perhaps a 20 to 90% chance of coronary artery disease (CAD)." The case is a rush. What do you do?

Well, you *could* consider issuing standard since the risk might be small — after all, it could be a "false positive" — but the uncertainty element makes you nervous. You might rate for coronary artery disease; but maybe it isn't, and maybe you'll lose the case. You could rate Table D (as a "don't know" hedge), but that's not very scientific. You could *reject the applicant* since this might even be unstable angina. You could *reject the test* results, but that would be risky, too. You could *repeat the test* — the size of the case might justify it — but that would create delay and could be perceived by the applicant as *one hoop too many* to jump through. Or, you could consider the statistical logic behind the AP's statement.

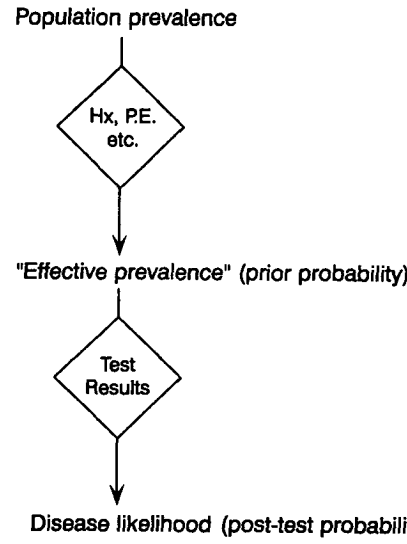
When the attending physician said that the chance was 20 to 90% for coronary disease, why the broad span? Was he being clever or ridiculous? Actually, he was being Bayesian. Bayes' theorem is a way of expressing the likelihood that a disease is present, given a positive test result for it. It is an expression of "conditional probability."

In All Probability

Probability should be more familiar than baffling to any of us, since it is merely the expression of the likelihood of something happening, expressed on a scale of 0 to 1 (0% to 100%). Probability is simply a means of "quantitatively expressing uncertainty or risk." We live in a world in which we deal with uncertainty and probability all the time — 30% chance of rain, 10% chance the airline will be late, 50% chance that your cholesterol level will lead to a heart attack in your lifetime. In the medical and insurance world, we have coined some expressions peculiar to ourselves to connote our estimates of probability, terms like "consistent with," "suggestive of," "essentially normal," "borderline standard," and "send it to Reinsurance."

"Conditional probability" is simply the likelihood of something being the case *given that* something *else* is already the case. In other words, **conditional probability** statements would include the following: a) the likelihood of having a positive test result, given the presence of a disease, and b) the likelihood of having a disease, given a positive test result. Note that "a" and "b" are not the same thing: "a" describes the **sensitivity** of a test, and "b" the **predictive value of a positive test** result. These concepts will be explained and developed further in a little bit.

Figure 1
Evaluation Process



Look at Figure 1. What the evaluative process of deciding the significance of a test result actually is, is a process of *revising probabilities.* In the absence of other information, you can begin with the general population prevalence for a disease or condition, assuming that the person could reasonably belong to that population. For heart disease, you could start with published findings of Framingham studies, American Heart Association data, or other such measurements. This will give the probability of disease for an applicant walking in off the street. Then, in light of the relevant *history* and *physical exam* (conducted by a physician as part of his diagnostic process, or as part of an application for insurance), the initial "population prevalence" is modified to come up with an "effective prevalence" based on the totality of information at hand. This effective prevalence becomes the "prior probability" of disease, i.e., the probability of disease prior to conducting the next diagnostic test.

After the test is done, its result is factored into an estimation of disease likelihood and a "post-test probability" of disease (or posterior probability) is calculated.

Expected Variation

In order to understand test results, it is important to be aware of the several sources of *expected* variation — including *biologic* variation, *sample* variation and *analytic* variation.

Biologic variation refers to variability of test results due to age, sex, diurnal variation, pregnancy, fasting state and other such *biologic* factors. *Sample* variation refers to the variability caused by specimen-handling (spun or unspun, "time in the tube," exposure to heat or cold, etc.) and processing. *Analytic* variation refers to variability due to test methodology (e.g., type of assay) and the inherent accuracy and precision of the technique. (Accuracy refers to the degree to which the lab result is identical to the actual or absolute value (e.g., when compared to a known "standard") and precision refers to the reproductibility of a test result on repetitive runs.)

For a given test, the person attempting to interpret the test result needs to factor in *all* such sources of variation known to him in order to interpret a test result as *abnormal*, and then to judge whether the abnormality is of clinical or underwriting *significance*, and finally to be able to conclude what it is likely *due to*.

What is Normal?

Before one can work with *ab*normals, one needs to be clear about what is meant by normal. The term *"normal"* gets tossed around rather loosely, and Murphy has suggested that "normal" as used in everyday speech can have one (or more) of seven different meanings (Figure 2).

Figure 2
Meanings of the Word "Normal"

Preferable Term	Paraphrase
Gaussian normal	Bell-shaped distribution
Mean/median/mode	Average/representative
Habitual	Commonly seen
Desirable/optimal	Fittest for survival
Innocuous	Carrying no penalty
Conventional/approved	Consensus/fashionable
Ideal	Aspired to

(after Murphy)

You will note on the report page for most blood test results that the term "normal range" is not used much anymore because it may wrongly imply a bell-shaped distribution for the lab results. (Many lab results, in fact, have a skewed, bimodal, or other-than-bell-shaped distribution.) Instead they are called "reference ranges" or "usual clinical ranges." This implies that these are consensus norms intended to reflect an "average" or typical reference population with which the sample can be compared.

For example, a reference laboratory may report a usual clinical range for cholesterol as 160-240 mg/dl, whereas "desirable" or "optimal" for a healthy or low-risk 40-year-old might be less than 200. Another fact to be aware of is that being too high or being too low may not be equally significant. The "usual clinical range" tends to be a "center cut." This leaves values on both the high-end and the low-end outside the "normal range," but for some lab tests, there is no clinical significance to the low extreme. It is really only the high end that is of interest or concern.

And just as underwriters must be prepared to ask whether a result flagged as abnormal (outside the range) is really *significantly* abnormal, another phenomenon to be aware of is that some lab results may be abnormal *for that individual* but hidden in the usual clinical range and, therefore, not advertised by an asterisk(*) or a HI/LO symbol. Because the usual clinical range is often a population-based reference range drawn from large numbers of people, the spread of the range can be quite broad. A creatinine for a 100 lb. woman might "normally" (healthily) be .3 mg/dl, but a value of 1.2 mg/dl (4 times "normal" for her and indicative of perhaps only 1/4 of her "normal" kidney function) would lie *within* most usual clinical ranges and *not* be associated with a telltale asterisk. Where available, a person's own prior test values might provide a more valid reference scale, but in the absence of such data, a population reference range is at least one useful yardstick for comparison.

Reference Ranges

Where do these reference ranges themselves come from? Reference ranges are constructed from groups of presumably healthy people, but such reference ranges may, in fact, include individuals with diseases that are subclinical or presymptomatic. We would like the reference range to represent only healthy people so that we can use it for comparison with our own results, in order to find the *ab*normals and detect *un*healthy people. But invariably any reference range has *some* unhealthy people in it, which is *one* reason why graphs of the test results of healthy and diseased populations commonly overlap instead of cleanly separating. Another reason for overlap is that the "diseased" population includes a spectrum of disease, with mild and early cases often toward the low end and more severe cases at the upper end.

Even though the reference group is the "presumably healthy" group, not even all of *its* members are considered "normal." Rather, the "normal" part of the reference range is defined as the central 95% of the reference population. This gives rise to another phenomenon of testing, namely the possibility that someone who is *healthy* may yet have an *abnormal* test result due to "chance" alone, and not due to any significant condition. In other words, if you did a single "routine" (not for a medical indication) lab test on an individual, and "normal" was defined as the middle 95% of "usual" test results, then you would have a 95% chance of the result coming back normal and a 5% chance of it coming back abnormal, due to just chance alone. What would happen if you ran a battery of tests? Well, if you did a chemistry *profile* (say a Chem-15) and if each of these tests were mutually *independent*, then the chance of being *normal* on *all* 15 is only .95 to the 15th power ($.95^{15}$) or only 46%! Turn it around, and the chance of being *abnormal* on *at least one* test is one minus 46% or 54%! (54% represents the sum of those individuals who are abnormal on at least one of those fifteen tests — they may be abnormal on just one or on various combinations of more than one.)

Abnormal results due to "chance" effects of multiple simultaneous testing tend to be *borderline* abnormalities that hover near the margins of the "normal" range. (Such results tend to "normalize" (regress toward the mean) on repeat testing, but they can cause considerable head-scratching when you first see them.)

Is it Hopeless, Then?

Given all of the angles and pitfalls to the interpretation of test results mentioned above, you may rightly wonder how it is ever possible to make a meaningful interpretation of a test result. But **interpretation relies upon the totality of the information available and not just the *one* fact of an isolated abnormality**. Doctors and underwriters ask themselves these questions: "How extreme is the abnormality? Is it solitary or not? Is it new or previously known? If previously known, is there any trend to it?" And most importantly of all, "what is the clinical context?" Do the history, physical exam, medications and previous diagnoses allow for a meaningful interpretation?

Validity

Understanding a bit about reference ranges and meaningful abnormals, the next thing to ask oneself is "how good is the test that is being performed?" A good test is one that has been standardized for the purpose in question, obtained for a worthwhile reason, and one which is a good discriminator for the condition in question. This raises the subject of *validity*. A *valid* test is one which is **appropriate for the intended purpose**. GGTP as a test for *diabetes* performs poorly; this is because it is not valid for that purpose. *Glucose* measurements as tests for diabetes, however, have validity for that purpose.

It is worth noting here that *few* tests are *unique* discriminators of one condition. While GGTP is most often used for assessing liver function, it is present in other tissues and can occasionally be elevated due to renal disease, for instance. If the GGTP *were* elevated due to renal disease, it would then be a "false positive" test for *liver* disease. This is not because there is anything "false" about the GGTP having been elevated, but rather because it is giving false and misleading information about the condition under consideration — liver abnormality. Or to take a slightly different example, while an elevated GGTP level commonly arouses suspicion for alcohol-related impairment, taken alone it is not *unique to* or even highly *predictive of* alcohol-related disease because elevations can and do occur for other reasons. This does *not* make it a bad test for underwriting purposes, because most of the *other* causes of significant elevations of GGTP *also* have important morbidity or mortality concerns to the underwriter; but for *establishing alcoholism* alone, the GGT by itself is of limited value.

There certainly are test results so extreme that you can say, "It's immaterial *what* the exact cause is; this is an unacceptable risk." But to do so, you must understand the test. Where possible, a medical director or underwriter would like to be able to name not only those underlying conditions that *could* cause an abnormal result, but what condition *is* causing it, and then "rate for cause."

Sensitivity and Specificity

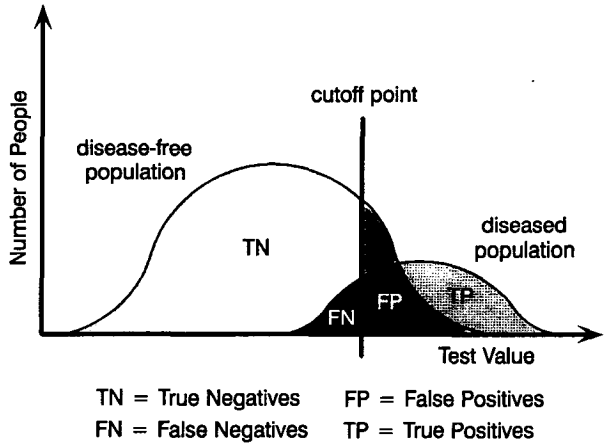
There are two more terms which it is important to clearly understand. Just as the word "normal" can mean several different things, the terms "sensitivity" (SN) and "specificity" (SP) have exact technical definitions that are slightly different from their common everyday usage. Sensitive can mean touchy, responsive or compassionate. But when the term **sensitive** is applied to a *test*, the exact definition to have in mind is "*positive in disease* (PID)." A highly sensitive test is one which, with high frequency, gives a positive result when the disease is known to be present.

Likewise, the term **specific** in everyday usage might mean precise, definite or particular. However, when we say that a test is specific, we mean that it is "*negative in health* (NIH)": that in healthy people (those without the disease or condition in question) the test is usually negative. One thing to note here is that in loose usage when someone says a test was "specific" for diabetes, it sounds as if he means that it is diagnostic of the disease being present (but note that this is more closely what the term *sensitive* — strongly associated with the presence of disease — technically means). Remember, a sensitive test, by definition is appropriately positive (i.e., *positive when disease is present*) and a specific test is appropriately negative (i.e., *negative when the disease in question is absent*).

True Positives, False Positives, etc.

If we were to graph the test results of a group of disease-free people against the range of values they could have for that test, we would get a curved distribution of results. This would be the reference population or the population that is presumed to not have the disease in question.

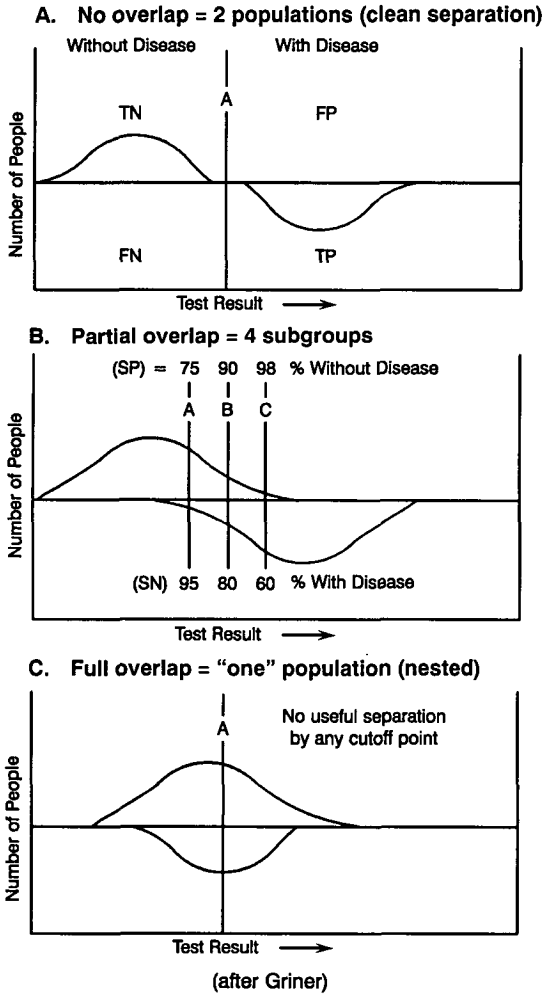
Figure 3



A second group of people, the "diseased population," will also have a certain distribution, and the *two curves* will typically *overlap* (Figure 3). If we now define a particular value of the lab test as the "cut-off point" *above which* we will classify the result as *abnormal*, we can see that we now get *four groups of people*. The cut-off line divides the non-diseased population into two parts, those below the cut-off who are the *true negatives* (TN = the lab result is negative, and the people are truly non-diseased) and *false positives* (FP = those *non*-diseased people who would falsely be called abnormal because their results are above the cut-off). Likewise, the diseased population is divided by the cut-off line into two parts, the *true positives* (TP

= those *with* the disease who are testing positive or abnormal, above the cut-off point) and the *false negatives* (FN = those members of the diseased population who are being missed because they fall below the cut-off point and are being classified as negative for the disease.) As you can see, the relative sizes of these four groups of people will depend upon: the sizes of the two populations, the degree of overlap of the two populations, and *where* the cut-off line is placed.

Figure 4



(after Griner)

If you move the cut-off point (Figure 4B) to the right, you will progressively eliminate the false positives, but at a price. That price is that you necessarily increase the number of false *negatives*. (You might want to try drawing similar graphs for yourself and move the cut-off level around.) Likewise, you could eliminate the false negatives by moving the cut-off level farther and farther to the left, but the price you would pay is increasing the number of false *positives*.

You will note that if the two curves did *not* overlap (Figure 4A), you could place the cut-off point *between* the two curves and now you would end up with just two groups — true positives and true negatives. But few diseases segregate so cleanly that a test can discriminate this well. Likewise, if the diseased population nestled *completely within* the hump of the non-diseased population (Figure 4C), it would be difficult to select a useful cut-off point since the degree of overlap of the two populations would be 100%.

Look at Figure 4 again.

The upper graph shows two populations with no overlap such that decision point A clearly separates the two populations. There are no false positives or false negatives. *Everyone* is properly and "truly" classified. This would be a "perfect test" for distinguishing the disease in question.

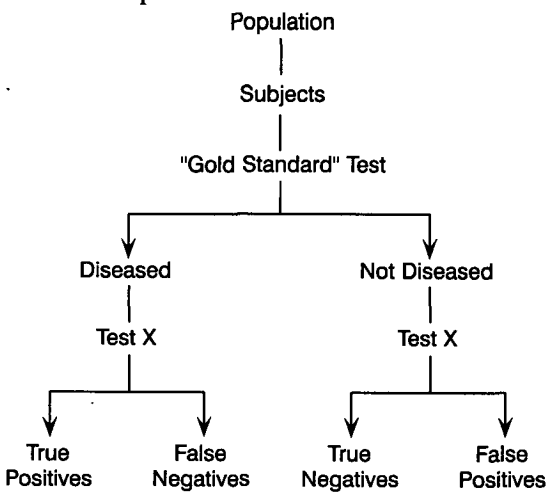
The bottom graph has the diseased population as a complete subset of the larger non-diseased population. No value of the test — not A, not anywhere — helpfully separates the groups. This would be a "worthless test." It would have no value for distinguishing between these two populations and would not be valid for that purpose.

But in the middle graph, there is significant overlap, and decision points could be placed at A, B, or C. This is a "typical test." Decision point C would result in high specificity but low sensitivity. Decision point A would have high sensitivity but at the price of low specificity. This is the *inevitable trade-off* when dealing with tests used to discriminate conditions in populations which have overlap: there is no free lunch, and it is hard to get both high specificity and high sensitivity out of a given cut-off level for a particular test.

This, then, is how sensitivity and specificity behave. But where do the numbers come from that tell you *how* sensitive or specific a test is for a particular disease? Well, somebody conducts the study and finds out. Typically, it works like this (Figure 5):

Figure 5

Relationships Between Test Results and Disease



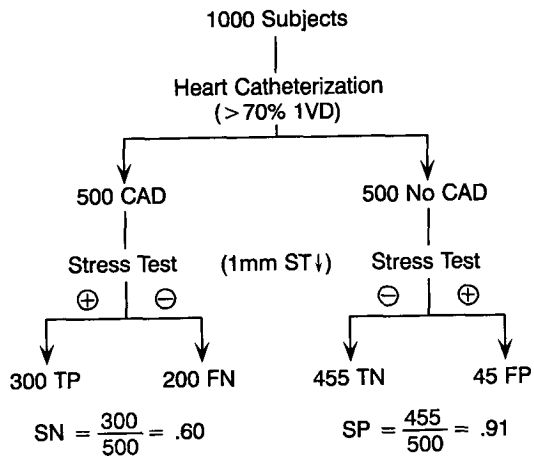
Out of a population, a group of subjects is chosen (according to some criteria) and subjected to a "gold standard" test which allows for them to be divided into two groups — those *with* the disease in question and those *without* the disease. For the condition "coronary artery disease," the gold standard might be coronary angiography.

Having established those "with" and those "without" disease, *another* test — more generally suitable, less invasive — (one, that could ultimately be used on a more general population as a surrogate for the gold standard) is applied. For coronary artery disease, this could be the treadmill test. The results of the treadmill, when applied to the "diseased" and "not diseased" populations, will then be either positive or negative. In the population with *already known* disease, the results will either be *true positives* or *false negatives*. In the population *known not to have* the disease, the results will be *false positives* and *true negatives*.

Let's paint in some numbers (Figure 6).

Figure 6

Example

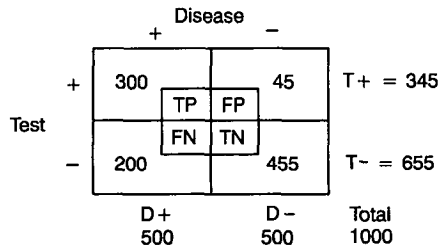


Suppose 1,000 subjects are chosen and subjected to coronary angiography. Suppose the decision point for angiography (i.e., the point at which we say "yes, they've got the disease") is >70% obstruction of at least one coronary artery." Let's assume we assemble a study group of 1,000 subjects, 500 *with* coronary disease and 500 *without* coronary disease, according to this definition. Now we apply the treadmill test, and we use a decision point for positive and negative outcomes of "1 mm. horizontal ST-segment depression." The results could look as follows: out of 500 patients with coronary disease, 300 have true positives and 200 have false negatives, according to the stress test; likewise, 455 have true negatives and 45, false positives.

Now recall that *sensitivity* (SN) means "positivity in disease." This means the ratio of true positives to the total number with disease. (This latter number — the total with disease — as you can see, is the sum of the true positives and the false negatives.) Thus, the sensitivity of this test in this situation is 300/500 or .60. Likewise, the *specificity* (SP) of the disease is its "negativity in health" (or negativity in absence-of-disease, according to how the "disease" may be defined). Thus, the specificity here is the true negatives over the total number without-disease. Thus, the specificity is 91%. These results aren't bad and are fairly typical of treadmill testing in such circumstances.

We can display the same data in a 2x2 table (Figure 7).

Figure 7



$$\text{Prevalence} = \frac{D+}{\text{Total}} = 50\%$$

$$SN = \frac{TP}{D+} = \frac{TP}{TP + FN} = 60\%$$

$$SP = \frac{TN}{D-} = \frac{TN}{TN + FP} = 91\%$$

$$PVP = \frac{TP}{T+} = \frac{TP}{TP + FP} = 87\%$$

$$PVN = \frac{TN}{T-} = \frac{TN}{TN + FN} = 69\%$$

The usual **convention** for doing so is to have "disease" at the *top* and "test" outcome at the *left* side, and to have the left upper box be the "positive-positive" box. Therefore, according to this convention, sensitivity and specificity are sort of "vertically" defined.

Sensitivity falls out of the first column, TP/D+ (D+ = disease-positive), and specificity falls out the the second column, TN/D-. Now just replace the abbreviations TP, FN, FP, and TN with the numbers from figure 6, and presto! You have changed the representation of the data from an algorithm to a 2x2 table. Each is a perfectly good representation of the same data, but the latter form may be more useful to those who think algebraically. Note that in addition to being "positivity in disease," sensitivity could also be called "few false negatives," because the fewer the FN, the higher the sensitivity. Likewise, specificity relies on "few false positives" in order to be as high as it can be.

So, sensitivity and specificity get defined for a *particular disease* (itself defined according to a gold standard) and for a *specific decision point* (such as 1 mm ST depression) defined according to clinical appropriateness as a discriminator. Are the sensitivity and specificity permanent attributes of the test? The answer is "no". Recall from Figure 3 that *where* the cutoff point or decision level is placed determines the proportion of results that are true positive and false positive, true negative and false negative, and since the sensitivity and specificity reflect these proportions (Figure 4B), they necessarily vary according to the cut-off point chosen. For example, a 2 *mm* ST segment depression is more "significant" than 1 mm depression. Experience teaches that it is more likely to define a "surer" positive (someone with coronary disease) than someone whose level of positivity on a treadmill may only be manifested at the 1 mm mark. This is the same as saying that in moving to a *stricter* cut-off point (2 mm), one increases the specificity but reduces the sensitivity of the treadmill test. Or to put it another way, one has reduced the number of false positives by allowing more false negatives.

One final thing to note is that this "standardization" study which gives us the sensitivity and specificity values was conducted on a population in which, quite artificially, there were equal numbers of people "with" and "without" the disease. It is pretty common to do it this way, with equal numbers of cases and controls. However, this gives you an inflated "prevalence" of disease of 50% for the subject population (D+/total number of subjects). The significance of this factor will become evident when we next discuss the "**P**redictive **V**alue of a **P**ositive test (PVP)." The baseline prevalences of diseases in the general population almost never run this high, but more commonly run between .1% and 10%, and may often be <0.1%.

Predictive Value

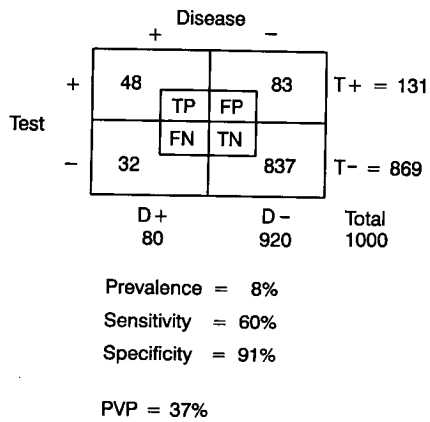
The *predictive value* of a test is how well it predicts the presence of the disease when the test is positive, or how well it predicts the absence of the disease when the test is negative. This is what one really wants to know, after all. All those other things are well and good, but predictive value is the bottom line.

Looking again at Figure 7, the "predictive value of the positive test result (PVP)" is defined "horizontally" on the 2x2 table, i.e., TP/TP + FP (or TP/T+). What this says is "what is the likelihood of the disease being present, if the test result is positive?" Likewise, the predictive value of a negative test result is the lower pair of boxes going across, i.e., TN/TN + FN (TN/T-) and this expresses the likelihood of the disease being absent when the test outcome is negative. (Remember, "positive" and "negative" test results refer to exceeding or being below whatever your cut-off point is — e.g., 1 mm or whatever, if you're talking about a treadmill.)

In our last example, where the sensitivity was 60% and the specificity 91%, the PVP = 300/345 = 87%. Notice that it's not 100% — that it doesn't predict the presence of disease with *certainty* but only with probability. (You would have to have 0 false positives for the equation to give you *that* result.) But 87% is quite good and makes coronary disease "highly probable."

Well, let's suppose that instead of starting with a population prevalence of 50% CAD, we instead are performing our treadmill test on a population with a pretest probability of disease of 8%. This would correspond to the baseline prevalence of CAD in a group of asymptomatic 50 year-old, American males. Let's start with a group of 1000 such people and create Figure 8.

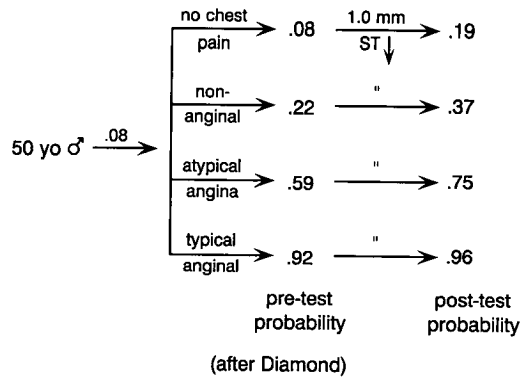
Figure 8



Of the 1000, 8% of them will have the disease (D+ = 80). The remaining 92% will be disease-free (D- = 920). The test sensitivity and specificity remain the same, assuming that we are using the same decision point of 1 mm ST segment depression, so sensitivity is still 60% and the specificity 91%. So, to finish filling in the boxes, TP = 60% x 80 = 48; and FN = the remainder of the 80, or 32. TN = 91% x 920, or 837; and FP = the remainder of the disease-free, or 83. That means T+ = 48 + 83 = 131, and T- = 837 + 32 = 869. The predictive value of the positive test, given this sensitivity and this specificity, and given a population prevalence of 8%, calculates out to be 48/131, or only 37%. Now, this is not a sufficient likelihood for a doctor to make an unequivocal diagnosis of coronary artery disease, but 1/3 chance of having ischemic heart disease might still be more uncertainty than an underwriter might care to live with. This is the problem inherent in screening tests — doing treadmills, etc. on asymptomatic individuals — as opposed to doing it for a medical indication such as substernal chest pain: low initial prevalence (low "pre-test probability") invites high false positive rate (low PVP) and requires further investigation to differentiate true positives from false positives. However, when a test is done after *risk factor* assessment, in light of *complaints*, and after *exam findings* have **raised** the index of suspicion higher, then it is being applied in a situation of much higher pre-test likelihood of disease than the mere population prevalence would suggest (recall Figure 1). You can begin to see then that a test of a given sensitivity and specificity can give you different probabilities of disease depending on what the pretest prevalence or likelihood was in the first place.

Figure 9

Probability of CAD After ETT
(effects of different kinds of pre-test pain)



You remember we started this whole discussion with a hypothetical applicant. He was a 50 year-old, American male with an abnormal ("positive") treadmill test.

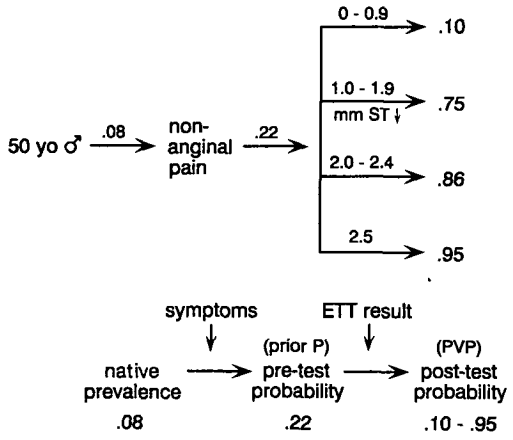
Figure 9 shows what the pre-test probability of disease would have been, depending on the presence and type of chest pain prior to the positive treadmill. (ETT means exercise tolerance test and is one of the abbreviations used for treadmills. Please note the pre- and post-test likelihoods in Fig. 9 and 10 are adapted from tables developed by Diamond, which is one place such estimates can be obtained. They are based on

somewhat different sensitivity and specificity assumptions than some of our other examples, like Fig. 8 and 11, so keep this in mind.) As you can see, our hypothetical applicant has a baseline prevalence of 8% of having coronary artery disease just by virtue of being 50, male and American. If he were having no chest pain prior to his treadmill, this remains his "pre-test probability." If he were having nonanginal chest pain, his pre-test probability becomes 22%; if atypical angina, 59%; and if typical angina, 92%. Following a treadmill, positive for 1 mm ST depression, the post-test likelihood of disease ranges from 19% to 96%.

Now we see why the PVP of his abnormal treadmill could be anywhere from 20% to 90%! The attending physician wasn't just being coy, he was being Bayesian! He was telling us that the likelihood of the disease was a function of its conditional probablility: that the probability of disease reflected both the *fact* that the test was abnormal *and* the *pre-test likelihood* of disease.

Figure 10

Probability of CAD After ETT
(effects of extreme-ness of abnormality)



(after Diamond)

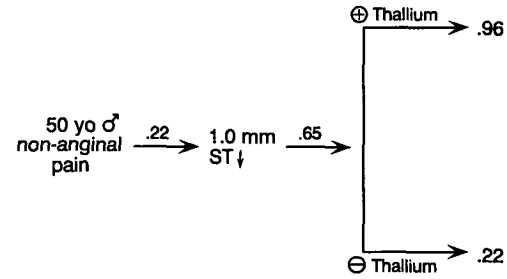
Suppose that rather than screening an asymptomatic patient, the attending physician is using the treadmill to help evaluate a chest-pain complaint. Let's suppose the applicant was having non-anginal chest pain. What would we have considered the likelihood of CAD after the treadmill if the treadmill abnormalities were something *other* than 1 mm? Figure 10 shows what different cut-off points for positivity on the treadmill might give you. As you can see, a treadmill result of less than 1 mm basically reclassifies Mr. Bigg from a pre-test likelihood of 22% for his nonanginal chest pain to 10%, which was close to his original "population probability" of having CAD just by virtue of being 50 and male. On the other hand, a treadmill result of 2.5 mm or greater would give him a 95% likelihood of coronary artery disease.

Suppose again that our hypothetical applicant winds up with a 1 mm treadmill and a 75% post-test likelihood of having coronary artery disease (Fig. 10).

Is this the closest one can get to a final risk assessment, or is there a way of trying to further resolve this question? You know the answer to this: look to results of further (sequential) testing using some independent means (that is, some other testing method that investigates for ischemia using a different end-point from electrical repolarization abnormality). Typically, thallium perfusion scans are used in this fashion as a means of further evaluating a treadmill abnormality. (An alternative to a Thallium scan might be an RVG — radionuclide ventriculogram — also known as a MUGA or **multigated** blood pool scan — to investigate wall motion abnormality.) Let's suppose the hypothetical applicant undergoes thallium testing next (Figure 11).

Figure 11

Probability of CAD After ETT and Thallium Test
(effect of multiple independent tests)



(Assume Thallium SN = 71% SP = 94%)

If his thallium test is positive for a reperfusing abnormality, he goes from a 75% likelihood of disease to a 96% likelihood; and if his thallium scan is negative, his chance of having CAD reverts to about 22%, (which you will recall was the pre-test likelihood of CAD in a 50 year-old male with nonanginal pain before the treadmill was positive!). I think you begin to see the pattern here. (Note that a negative thallium scan does not tell you that there is a *zero* percent chance of CAD, but instead, it essentially cancels the effect that the abnormal treadmill had on the likelihood of CAD, i.e., takes the applicant back near his pre-test 20% likelihood.)

Figure 12 is a table compiled from different articles from the medical literature that gives you some overall sense of variability of sensitivity and specificity according to the disease being assayed for, the particular gold standard used, and the cut-off point used to call the test positive. Thus, you can see that when coronary artery disease is defined as "greater than a 50% narrowing of at least one vessel," it has a higher sensitivity and lower specificity than when a stricter gold standard (70% one vessel disease, or more) is used, even with the same 1 mm ST depression end-point on the treadmill. (Does this imply that some gold-standards in use are really "gold-plated"? The reader can decide!)

Likewise, when the gold standard for CAD is a 70%-or-more narrowing of at least one vessel, the sensitivity and specificity for coronary artery disease depends on whether an endpoint of 1 mm or 2.5 mm is chosen. The sensitivity decreases as the endpoint becomes stricter, but the specificity increases. The thallium scan has a better sensitivity and specificity for coronary artery disease against a gold standard of "70% obstruc-

Figure 12
Variability of SN % SP

Test	Endpoint	For	SN%	SP%	Gold Std.
Treadmill	1.0 mm	CAD	75	85	> 50% 1VD
	1.0 mm	CAD	60	91	> 70% 1VD
	> 2.5 mm	CAD	20	99.5	> 70% 1VD
Thallium	reperfusing defect	CAD	71	94	> 70% 1VD
GGTP	3xULN	ETOH(L)	88	85	liver BX(H)
	3xULN	ETOH(L)	52	(85)	liver BX(A)
ELISA	OD⊕	HIV	98.8	99.6	HIV
WB	3 lanes⊕	HIV	99.6	99.8	HIV
Treadmill + Thallium	1.0 mm + reperfusing defect	CAD	42	98	> 70% 1VD
ELISA (2) + WB	O.D.⊕ + 3 lanes⊕	HIV	99.995	99.9	HIV

(other studies have yielded other results)

tion" than a 1 mm treadmill abnormality does. (But it is a much more expensive test to conduct and is generally reserved for sequential testing in those cases where the treadmill is first abnormal, and its success depends on being able to achieve an exertional pulse rate sufficient to produce a reperfusing abnormality [i.e., two *resting* Thallium scans are not going to have much information about ischemia].)

Because liver test abnormalities and HIV testing arouse similar concerns about false positives and real PVPs, some representative numbers from the literature are displayed on these as well. For GGT an end point of 3 times the upper limit of normal was studied for alcoholic liver disease (not alcoholism, and not just any liver disease, but alcohol disease of the liver). The gold standard in the first case was liver biopsy in hospitalized patients (H) and in the second instance liver biopsy in ambulatory patients (A). Liver biopsy is not a gold standard that many people willingly volunteer for, so it is appropriate to try to have a noninvasive substitute for such an invasive gold standard. The sensitivities and specificities for GGTP *for this narrow purpose* do not generate impressive PVPs (if you work through the arithmetic as on the treadmill example). But for underwriting purposes, when we use GGTP for "all conditions of substandard implication" instead of merely for detection of alcoholic liver disease, the overall sensitivity and specificity are better, and so is the predictive value.

And now we come to "Bayes' Theorem" — that famous equation arrived at by a country cleric playing around with numbers in 1763. Here it is in all its glory:

$$PVP = \frac{(P)\ (SN)}{(P)\ (SN) + (1\text{-}P)\ (1\text{-}SP)}$$

Hopefully, it will no longer be frightening or intimidating. It can be expressed using a variety of symbols or algebraic forms, but the one shown here is one of the more useful ways of rendering it. P is the pre-test probability of the disease being present. (1-P) is 1 minus this probability. SN and SP are the sensitivity and specificity, respectively, for the condition in question. Quite simply, Baye's Theorem tells you the predictive value of a positive test — the likelihood that the disease or condition in question is present, given the fact that the test result is positive.

As you can now see, there is *much* more to it than meets the eye. For the test to have been positive means many things — how was it done? What cut-off point was used?, etc. The sensitivity and specificity, likewise, require you to know a little bit about what gold standard was used to define the presence of the condition in question. But now, within those few constraints, all you have to do is fill in the blanks, crank out an answer, and you can do estimations of the likelihood of disease *all on your own*. Rather a powerful little tool!

With just a little experimentation with different numbers and scenarios, you can discover a multitude of uses and ways in which these concepts can help you understand and interpret test results.

Appreciation is hereby expressed to the Home Office Life Underwriters' Association (HOLUA), on whose behalf some of this material was presented as workshops, and to Robert J. Pokorski, MD, for his ideas and encouragement.

SOME USEFUL REFERENCES

BASIC

* DeTorre, AW. "The Evaluation of Abnormal Laboratory Results," J. Insur. Med. (1988) 20: 5/23.

Diamond, GA, Forrester, JS. "Analysis of Probability as an Aid in the Clinical Diagnosis of Coronary Artery Disease," NEJM. (1979) 300: 1350-1358.

Galen, RS, Gambino, SR. *Beyond Normality*. NY: Wiley, 1975.

* George, Hank. "Analyzing and Underwriting The Blood Chemistry Profile." Lincoln National Management Services.

* Pokorski, RJ. "Overview of Test Theory." *Triennial Course*, The Wigwam, Litchfield Park, Arizona, 1988.

* Riegelman, RK. *Studying a Study & Testing a Test*. Boston: Little Brown, 1981.

* Sackett, DL, et al. *Clinical Epidemiology*. Boston: Little Brown, 1985.

Siest, G, et al. *Interpretation of Clinical Laboratory Tests*. Foster City, CA: Biomedical Publications, 1985.

Speicher, CE, Smith, JW. *Choosing Effective Laboratory Tests*. Philadelphia: Saunders 1983.

Wallach, J. *Interpretation of Diagnostic Tests*. Boston: Little Brown, 1986.

INTERMEDIATE

Bahn, AK. *Basic Medical Statistics*. NY: Grume & Stratton, 1972.

* Colton, T. *Statistics in Medicine*. Boston: Little Brown, 1974.

* Griner, PF, et al. "Selection & Interpretation of Diagnostic Tests & Procedures," Annals Int. Med. (1981) 94 (4, Part 2): 453-600.

* McNeil, BJ, Sherman H. "Primer on Certain Elements of Medical Decision Making." NEJM (1975) 293: 211-215.

Murphy, E.A. *The Logic of Medicine*. Baltimore: Johns Hopkins Press, 1976.

MORE ADVANCED

Feinstein, AR. *Clinical Biostatistics*. St. Louis: Mosby, 1977.

* Ingelfinger, JA, et al. *Biostatistics in Clinical Medicine*. NY: Macmillan, 1987.

Weinstein, MC, Fineberg, HV. *Clinical Decision Analysis*. Philadelphia: Saunders, 1980.

(Asterisked references are especially helpful with some of the concepts).

APPENDIX

The discussion of treadmill testing was necessarily simplified for the purposes under discussion. It is important to remember that the information content of a treadmill extends well beyond that of an arbitrarily chosen index like "1 mm flat or downsloping ST depression." Questions to consider in *fully* evaluating treadmill reports would include such things as: "why was it done and why was it stopped?" Was the context "chest pain" and a high index of suspicion for coronary disease (in which case PVCs, ST segment depression more than 80 msec past the J point, and other findings may represent ischemic equivalents)? Was it *stopped* because of chest pain and if so, at what exercise level? Was the subject able to achieve 85% of his maximal pulse or not? Was the resting cardiogram abnormal, and if so, in what way? (Repolarization abnormalities with exercise are difficult to interpret meaningfully in the presence of left bundle branch block [LBBB] at rest.) Is there a normal heart exam? How was the exercise test done: 3 leads or 12 leads? According to Bruce protocol or some other protocol? If ST changes do occur with exercise, do they occur in one lead or more than one lead?

A treadmill test contains much valuable information and generally should be "milked" for *all* of the underwriting information that can be extracted from it.

In the hypothetical case, for his result to have been characterizable as *false* positive, there was presumably either negative sequential testing, allowing revision of his CAD-likelihood back toward baseline prevalence, or the total universe of information available about his CAD risk, including the specifics of his treadmill "positivity," allowed for such an inference. Underwriting, of course, is the science of applying methods *other than* wishful thinking to the task of risk-assessment!