5 | Use of the Laboratory

CRAIG KAPLAN

Ordering and interpreting diagnostic tests are fundamental skills. Surprisingly, however, evidence indicates that many of us are poorly trained in this vital area. Studies have shown that physicians commonly order more laboratory tests than required, use them for the wrong purposes, and ignore or misinterpret their results. While these errors have obvious implications for the quality of patient care, there are large socioeconomic implications as well. Health care expenditures in the United States in the 1980s exceeded $400 billion yearly, approximately 10% of the gross national product. Of this cost, a significant proportion can be attributed to laboratory tests. In one study of costs incurred during hospitalization, laboratory charges accounted for an average of 26% of the hospital bill. It has been estimated that clinical laboratories account for more than $11 billion in health care resources nationally, of which $3 billion was for clinical chemistries alone!

Diagnostic tests obviously add up to significant expenditures. If many of these tests are inappropriate, superfluous, or misinterpreted, then learning how to use the laboratory properly might have significant benefits for both individual patients and the economy as a whole. The term *diagnostic test* does not refer only to costly "big ticket" imaging or monitoring procedures such as magnetic resonance imaging, computerized axial tomography, electronic fetal monitoring, or cardiac catheterization. It refers in addition to the countless laboratory tests ordered every day on patients, tests such as electrolytes, serum chemistries, coagulation profiles, or complete blood counts. While individually inexpensive, these "little ticket" tests account in the aggregate for as much total health care expenditures as the more expensive items, perhaps more.

The rationale behind this chapter is to improve the quality of care you offer to your patients while decreasing its cost. Both aims can be achieved by a rational approach to diagnostic test ordering and interpretation. After reading this chapter, you should have an understanding of the following:

1. The probabilistic nature of diagnosis and its implications for ordering and interpreting laboratory tests
2. The five operating characteristics that define a laboratory test: reliability (precision), accuracy, sensitivity, specificity, and predictive value
3. The different purposes for which laboratory tests are ordered (diagnosis, monitoring therapy, and screening) and the operating characteristics required for each purpose
4. The "normal" test result and its meaning
5. Frequently used test ordering strategies and their limitations

The Probabilistic Nature of Diagnosis

Diagnosis is an uncertain art. By this I do not mean that physicians are too hesitant or equivocating but that, short of the autopsy room, we can rarely be absolutely certain of a diagnosis. This is particularly true in the early phases of diagnostic evaluations. Studies of physician reasoning have demonstrated that, as physicians evaluate patients, they keep a limited set of five to seven diagnostic hypotheses in mind, each of which is assigned a relative probability, such as "very likely," "possible," or "unlikely." These relative probabilities (or likelihoods) are adjusted upward or downward, depending on new information gained through history, physical examination, and diagnostic tests. With enough information, one diagnostic possibility becomes likely enough for the physician to stop further investigations and declare that possibility as the "diagnosis."

As an example of this process, imagine that you are asked to evaluate a patient with hypertension. Before you even see the patient, several diagnostic possibilities come to mind: essential hypertension, renovascular disease, primary hyperaldosteronism, pheochromocytoma, coarctation of the aorta, and chronic renal disease, to name a few. Based on large referred series of hypertensive patients, prior to any information other than the fact that the patient is hypertensive, you can speculate that the likelihood of essential hypertension is between 90 and 95% and that the cumulative likelihood of the secondary forms is between 5 and 10%, the majority being accounted for by renovascular disease. The "initial" probabilities might thus be stated: essential hypertension 90% (very likely), renovascular 5% (possible), and others less than 5% (unlikely).

Imagine, then, that you walk into the examining room and see that the patient is a 16-year-old white female. Based on the high proportion of such patients with renovascular disease, this information alone causes you to revise your probabilities. The "revised" probability of renovascular disease would increase, while those of essential hypertension and the others would decrease. Your adjusted or revised set of probabilities might then look like this: essential hypertension 70% (likely), renovascular disease 25% (likely), and others each less than 1% (extremely unlikely). If on physical examination you hear a flank bruit, you would be increasingly certain of renovascular disease, further raising its probability and lowering those of the other possibilities. This diagnostic process would then proceed with an intravenous pyelogram or radionuclide flow scan through arteriography, with probability revisions occurring after each step.

The important point here is that, whether done explicitly or implicitly, physicians approach diagnosis by using this process of successively revising probabilities. There are several problems, however. Evidence also exists that physicians do not gain maximum benefit from information available from history, physical, and diagnostic tests; that is, they do not always revise diagnostic probabilities to the degree warranted by the data. When compared to formal mathematical models of probability revision, physicians fail properly to adjust their disease likelihoods, either under- or overestimating the diagnostic value of laboratory tests. Physicians

tend to treat diagnostic tests as if they were perfect information, failing to recognize that laboratory data, in and of themselves, are probabilistic. These human failings in the face of uncertainty lead to the misuses of diagnostic information described earlier.

Much of this chapter and the one that follows is devoted to the implications of the probabilistic nature of diagnosis and the interpretation of laboratory tests. In them, some formal rules for diagnostic test interpretation are set forth.

Operating Characteristics of Laboratory Tests

Several characteristics of tests are crucial to their interpretation. Two of these operating characteristics enable you to judge the test per se: reliability and accuracy. The other characteristics enable you to judge a test with respect to how its results affect your diagnostic probabilities: sensitivity, specificity, and predictive value (positive and negative).

The *reliability* (or precision) of a test is a measure of reproducibility obtained by running the test many times on the same specimen. An unreliable test is one that yields results that vary widely due to chance or technical error. Such tests are therefore very difficult to interpret because day-to-day variation in a given test may be due to something other than a true change in the patient's condition. Such changes in a reliable test, in contrast, are more likely to reflect true patient change. Test *accuracy* reflects the degree to which the test result reflects the "true" value. Imagine

that we are investigating a new, cheaper way of measuring hematocrit. If we assume that the "true" hematocrit is obtained by hand centrifugation, then the new test is accurate if it closely matches the centrifuged results. If the test is inaccurate, then it will systematically vary from the true value.

It is important to realize that test reliability and accuracy may vary independently of one another. A test may be inaccurate but totally reliable. This would be the case if the newly developed hematocrit in the above example was always exactly five percentage points higher than the centrifuged hematocrit. In this case, the new hematocrit would be reliable (always the same value), but inaccurate (always different from the true value). The different relationships between reliability and accuracy are shown in Figure 5.1.

The *sensitivity* of a test refers to the proportion of patients with a given disease who have a positive test. The *specificity* of a test refers to the proportion of patients without the disease who have a negative test. As an example, consider obtaining antinuclear antibodies (ANA) in a large number of patients with systemic lupus erythematosus (SLE). Each patient may have either a positive ANA, referred to as a *true positive*, or a negative ANA, referred to as a *false negative*. Sensitivity would be the proportion of these SLE patients who have positive ANAs. Similarly, the proportion of a set of patients without SLE who have negative ANAs, called *true negatives*, would be called the specificity of the ANA. Those non-SLE patients with positive ANAs would be called *false positives*.
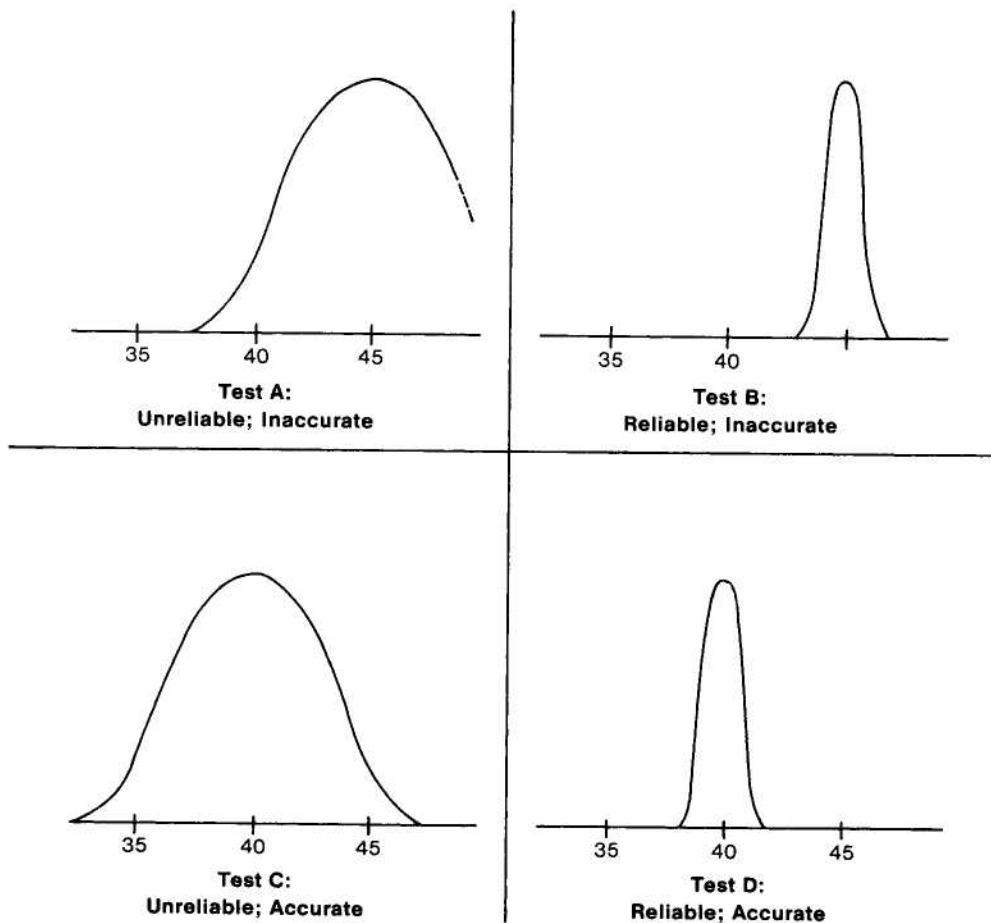

Figure 5.1
Operating characteristics of diagnostic tests. The different relationships between reliability and accuracy demonstrated for a measurement of hematocrit. Assume that the "true" hematocrit is 40%.

Predictive value, on the other hand, refers to the likelihood that a patient has or does not have the disease, given a positive or negative test result. Referring again to our SLE/ANA example, if you took a large number of people with positive ANAs and independently (without using the ANA results) determined which truly had SLE, the proportion of those ANA-positive patients who had SLE would represent the *positive predictive value* of the ANA. Conversely, the proportion of those negative-ANA patients who did not have SLE would represent its *negative predictive value*. The relationships between sensitivity, specificity, and positive and negative predictive values are demonstrated in Figure 5.2.

While these concepts are dealt with in greater depth in the following chapter, it is important to note here that the positive and negative predictive values of a test depend not only on its sensitivity and specificity, but also on the likelihood of disease *before* the test is done (referred to as *prior probability, prior likelihood,* or *prevalence* of disease). This means that, while the sensitivity and specificity of tests do not vary from patient to patient, the positive or negative predictive values will vary if the patients have different prior likelihoods of disease. The importance of prior likelihood in interpreting a test result cannot be overemphasized. In particular, a low prior likelihood frequently results in a low positive predictive value even with a relatively specific test. Take, for instance, a disease with a low prior probability, say 5%. If a test for this disease has a sensitivity of 90% and a specificity of 90%, what is the predictive value of a positive test? A little calculation reveals that the positive predictive value of this test in a patient with a 5% prior likelihood is only 32%. There is still a 68% chance that the patient is free of disease, even though the test is positive! In contrast, if the prior likelihood is 50%, the positive predictive value of the same test is 90%. Low prior likelihoods usually result in low positive predictive values.

If the above calculations surprised you, then you can understand that physicians do not always do well when asked intuitively to revise their own probabilities. For this reason, every physician should be able to perform simple probability calculations at the bedside. Understanding these operating characteristics will enable you to carry out the probability revisions necessary to the diagnostic process. First, a prior likelihood is determined from history and physical examination. Then a diagnostic test is obtained and, using sensitivity, specificity, and prior likelihood, the positive or negative predictive value of the test is calculated. This predictive value becomes the revised probability of disease. These calculations can be performed with a hand-held calculator, using no more effort than it takes to calculate antibiotic doses or creatinine clearances. The following chapter deals more specifically with the actual process of calculating predictive values.

Purposes for Ordering Laboratory Tests

The purposes for which tests are obtained have a great deal to do with both the choice of diagnostic test and its interpretation. One survey of physicians in a large teaching hospital found that three general reasons accounted for most laboratory test ordering: diagnosis (37%), monitoring therapy (33%), and screening for asymptomatic disease (32%).

Purpose 1: Diagnosis

In order to use a test for diagnostic purposes, the test has to be positive in a large proportion of patients with the disease (high sensitivity) and negative in a large proportion without the disease (high specificity). Ideally, for the test to be maximally useful for diagnostic use, both sensitivity and specificity should be 100%. This would mean that every patient with the disease would have a positive test (no false negatives) and result in a negative predictive value of 100%. Similarly, positive tests would occur only in patients with disease (no false positives), resulting in a positive predictive value of 100%. Unfortunately, tests with such high sensitivity, specificity, and predictive value are unheard of. In fact, tests with high sensitivity tend to have low specificity, and vice versa. A little thought, however, will allow you to choose the proper test for your particular diagnostic purpose.

One might say that there are two basic diagnostic uses for laboratory tests. The first is when you wish to rule out a disease absolutely and the second when you wish to confirm it. Examining these purposes more closely allows you to identify test characteristics necessary for each use. In order to be absolutely sure a patient does not have a disease (ruling it out), falsely negative tests have to be minimized, so a test with high sensitivity should be used. With a highly sensitive test, a negative result produces a very high negative predictive value, in essence meaning that the vast majority of patients with a negative result do not have the disease. Conversely, in order to be very confident that an individual does have the disease (confirming it), false positive tests should be minimized, necessitating a very specific test. With such a test, an individual with a positive result is very likely to have the disease (high positive predictive value).

In general, if you wish to rule out a particular diagnosis, choose a test with high sensitivity; if you wish to confirm a diagnosis, choose a test with high specificity. The particular diagnostic purpose for which you choose the test will determine the most important operating characteristic, sensitivity or specificity. There will be obvious tradeoffs, however. Since sensitivity and specificity tend to be inversely related, a test chosen for ruling out disease will not be par-

	Diseased	Nondiseased
Test +	TP	FP
	A	B
	C	D
Test −	FN	TN

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{A}{A + C}$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{D}{D + B}$$

$$\text{Pos. Predictive Value} = \frac{TP}{TP + FP} = \frac{A}{A + B}$$

$$\text{Neg. Predictive Value} = \frac{TN}{TN + FN} = \frac{D}{D + C}$$

Figure 2
Operating characteristics of diagnostic tests: sensitivity, specificity, predictive value.

ticularly valuable for confirming it if it is unexpectedly positive. Similarly, a negative test, if chosen to confirm the disease, is less helpful for ruling it out.

As an example, consider a disease for which there exists two tests: Test A, which is very sensitive but poorly specific, and Test B, which is very specific but poorly sensitive. Furthermore, assume that the likelihood of the disease is roughly 50% (a 50% chance that a patient has the disease). Table 5.1 presents the revised probabilities of the disease, given the results of each test. You can see that if Test A is used, it is possible to rule out the disease, but not possible to confirm it. If this test is positive, the likelihood of disease is still only 70%, whereas if it is negative, the likelihood drops all the way to 8%. Conversely, if Test B is used, it may be possible to confirm the diagnosis, but not possible to rule out the disease. If this test is positive, the likelihood of disease is greater than 90%, but if it is negative, disease likelihood remains as high as 30%. Remember, to confirm disease, choose a test with high specificity; to rule out disease, choose a test with high sensitivity.

Purpose 2: Monitoring Therapy

Examples of this use for diagnostic tests abound. Whenever a test is repeated in order to follow a therapeutic drug level or observe for side effects, it is being used for monitoring purposes. Following the serum potassium in patients using diuretics, the creatinine with use of nonsteroidal anti-inflammatory agents and aminoglycosides, or the hematocrit in patients with acute GI bleeding are all monitoring uses. Monitoring therapeutic drug levels is yet another instance. For this purpose, sensitivity, specificity, and predictive value have less relevance than do reliability and accuracy.

Monitoring a serum value ordinarily means measuring it repeatedly. For this purpose, reliability is critical. If serum potassium suddenly drops from 4.2 to 3.5 mmol/L, it is vitally important to be confident that this represents a true change in the potassium value, not simply technical error or random variation. For this, a highly reliable and accurate test is needed. It does little good to know that the potassium level is 4.0 mmol/L if the particular test you are using is inaccurate. In this situation, the true potassium might be as low as 3.0 or as high as 5.0.

For the purposes of monitoring then, the salient operating characteristics are reliability and accuracy.

Table 5.1
Effect on Revised Probabilities of Using Two Different Diagnostic Tests, One Which Is Highly Sensitive or One Which Is Highly Specific. The Prior Likelihood of Disease Is 50%

Test A: Sensitivity 95%
       Specificity 60%

Test B: Sensitivity 60%
       Specificity 95%

	Test A	Test B
Prior probability of disease	50%	50%
Revised probability if test positive	70%	92%
Revised probability if test negative	8%	30%

Note: If a test is negative, the revised probability is equal to 1 minus the negative predictive value.

Purpose 3: Screening

The object of the use of diagnostic tests for screening is to detect disease in its earliest, presymptomatic state when, presumably, it is less widespread and more easily treated or cured. Most screening programs, such as stool occult blood screening or mammography, are aimed at cancer detection, although other screening programs exist for disorders such as glaucoma, hypertension, or diabetes. There are many criteria for successful screening programs, but we focus here on operating characteristics that are necessary to produce a good screening test.

An ideal screening test would share many of the characteristics of the ideal diagnostic test in general. Sensitivity, specificity, and predictive values should all be close to 100%. As I mentioned above, such tests do not exist. We are ordinarily faced with a choice between tests with high sensitivity or high specificity. What are the implications of this choice when a screening test is desired?

The initial tendency would be to use a sensitive test. Such a test, because it has few false negatives, has a high negative predictive value, meaning that very few individuals with disease would be missed. The tradeoff, unfortunately, is that this type of test (high sensitivity, low specificity) has a poor positive predictive value, meaning that most positive individuals would be falsely positive. When such a test is used for screening large populations, a significant number of individuals have positive tests but only a small number of them are truly diseased. Remember that individuals in screening populations, by definition, have relatively low prior likelihoods of disease. In such low-likelihood populations, the predictive value of positive tests will be low. Such is the case for stool occult blood screening, where between 2 and 6% of all tested individuals are positive but only 5 to 10% of these positive individuals have colon cancer. With a positive predictive value of only 5 to 10%, many individuals are subjected to costly, uncomfortable, or potentially dangerous procedures in order to detect those few with colon cancer. But using a highly specific test has its drawbacks too. Highly specific tests will have higher positive predictive values, but lower negative predictive values. A higher proportion of patients who have positive tests will actually have cancer, but fewer individuals with cancer will be detected. In the absence of a relatively inexpensive test that is both highly sensitive *and* highly specific, the choice is between a sensitive test that results in a high proportion of unnecessary diagnostic evaluations, or a specific test that results in a sizeable proportion of diseased individuals going undetected. Unfortunately, there is no general rule to depend on here. Your choice will depend on whether or not patients will benefit from early disease detection, as well as the relative costs and risks of subsequent diagnostic evaluation.

Table 5.2 summarizes general guidelines to use when

Table 5.2
Selection of Diagnostic Test According to Purpose

Purpose	Most important operating characteristic(s)
Diagnostic	
Confirmation	Specificity
Ruling out	Sensitivity
Monitoring	Reliability
Screening	Sensitivity/specificity[a]

[a]See text for explanation.

choosing diagnostic tests for one of these three specific purposes. Remember that it is important to have a specific purpose in mind when choosing a test because very few tests have the operating characteristics required to satisfy every purpose.

The Normal Test Result and Its Meaning

How do most laboratories define the normal values they provide for their results? Most commonly, this set of normal values is obtained by determining test results on a large number of presumably healthy volunteers and plotting a frequency distribution of the results. The normal range is then determined by calculating the mean and standard deviation of this distribution and declaring as normal any value falling within two standard deviations of the mean (Fig. 5.3). By definition statistically, this range encompasses roughly 95% of the normal individuals. When normals are defined in this way, 5% of normal individuals are then paradoxically declared abnormal! This is an important point to remember when you are faced with an unexpectedly abnormal test result in a patient.

The fact that a normal person, on any given test, has a 5% chance of being declared abnormal has important implications, especially when one recalls that diagnostic tests are often obtained in groups (or batteries) of 12, 20, or even more. In such large diagnostic batteries, the overall likelihood that any one test will be abnormal rises as the number of tests in the battery rises. A look at Table 5.3 demonstrates this. The likelihood that any one test in a 5-test battery will be abnormal is 23%. This means that roughly 1 in 4 healthy patients will have an abnormal test result. The percentage of chance abnormals is even more impressive as the number of tests rises. In a 40-test chemistry "screen," the likelihood of a set of completely normal results is only 13%! It is important to consider the possibility of such chance deviations when you are deciding whether or not to investigate further unexpectedly abnormal test results on large test batteries. Such investigations may account for a large proportion of the unnecessary costs and risks incurred through inappropriate laboratory testing.

There are, of course, causes for unexpected abnormals other than chance and the presence of unsuspected disease.
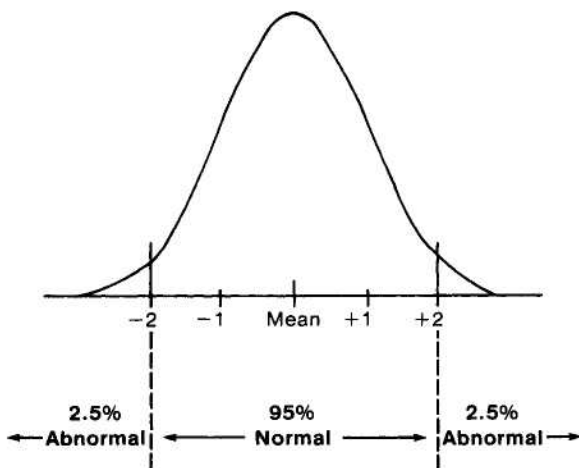
Table 5.3
Likelihood of Abnormal Tests, by Chance Alone, on Large Multiple-Test Batteries

Number of tests in battery	Likelihood of at least one test abnormal by chance[a]
1	5%
2	10
3	14
4	19
5	23
10	40
20	64
40	87

[a]Assumes normal range chosen as ± 2 standard deviations from mean of normal population.

Certain subsets of individuals differ systematically from the normal range of the general population. The very young and the elderly, for instance, have their own particular normal ranges for many diagnostic tests. If the usual normal ranges are applied, significant proportions of these individuals may be declared abnormal. This may also be true for persons of different sex and for persons of varying racial backgrounds.[1] What, then, is one to do when faced with an unexpectedly abnormal result? My suggestion is to reconsider and reexamine the patient. By taking a more focused and detailed history, as well as repeating certain portions of the physical examination, you may uncover information that, in the light of the unexpected result, may increase your suspicion of underlying disease. If this process does uncover new information, perhaps the unexpected result may reflect a true abnormality in the patient. If not, however, then perhaps the best strategy is cautiously to observe the patient over time, repeating the pertinent history and physical as well as the single abnormal test result. If the patient develops symptoms or signs, or if the test result becomes progressively more abnormal, then further investigation is warranted. It is difficult, of course, to recommend a specific action for every unexpected abnormal result. Instead, my intent is to create awareness that many, perhaps the majority, of such test results are due to some factor other than underlying disease. The best course of action may often be a cautious "wait and see," as opposed to blindly pursuing every abnormal value to its costly and often frustrating end point.

A normal test result, on the other hand, may provide more information than you think. It may, in fact, affect diagnostic probabilities just as much as an abnormal test result. As an example, imagine a patient in whom you are considering two disease possibilities, A and B. In disease A, the erythrocyte sedimentation rate (ESR) is almost always elevated (80% of the time), whereas in disease B, it is uncommonly elevated (20% of the time). How will a normal ESR affect the diagnostic likelihoods if the original probabilities of diseases A and B were both 50%? A look at Table 5.4 demonstrates that the normal result has had a marked effect on diagnostic probabilities. The revised probabilities show that disease B, given a normal ESR, is now almost four

Figure 5.3
Determining the normal range of a diagnostic test as ± 2 standard deviations from the mean.

1. There are exceptions, albeit rarely. The normal ranges for hemoglobin and hematocrit are different for men and women. In addition, many tests have different normal ranges for pediatric and adult age groups. The term *adult*, however, compromises a large age range in itself, for which normal lab values vary considerably.

Table 5.4
Value of Normal Test Results

Disease	Likelihood of elevated ESR	Prior probability of disease	Revised probability of disease
A	80%	50%	20%
B	20	50	80

times more likely than disease A. Normal results can markedly alter diagnostic probabilities.

Even normal individuals have a range over which test results are distributed. This is also true for abnormal individuals, and it is unfortunately the case that these ranges overlap. Once this situation is understood, it becomes apparent why the sensitivity and specificity of many tests cannot be simultaneously maximized. Figure 5.4 demonstrates this. In Figure 5.4A, the overlapping ranges of normal and abnormal individuals for a given lab test are displayed. If we define the "cutoff point" as that value below which the test result is said to be abnormal, then a certain percentage of normal individuals will be declared abnormal (the false positives) and a certain percentage of abnormal individuals will be declared normal (the false negatives).

What would happen if an attempt was made to improve the sensitivity of the test by changing the cutoff point? This could be done by moving the cutoff point to the right, as in Figure 5.4B, decreasing the numbers of false negatives and increasing sensitivity. Unfortunately, by moving the cutoff point to the right, the number of false positives would increase, decreasing specificity. As long as the distributions of normal and abnormal overlap, there is no way to increase sensitivity without decreasing specificity.[2] The reverse is also true. Figure 5.4C represents an attempt to increase specificity by moving the cutoff point to the left in order to decrease false positives. This results in decreasing sensitivity by simultaneously increasing the number of false negatives.

It is sadly true that as long as normal and abnormal distributions overlap, any attempt to improve sensitivity or specificity by changing the cutoff point will result in lowering the other characteristic. This is the case for most laboratory tests. There is, however, a brighter side. Recall that the required operating characteristics differ with different purposes. Rule-out tests require high sensitivity, whereas confirmation tests require high specificity. One therefore might imagine changing the cutoff point in order to make the test suitable for a particular purpose. If the test is being used for rule-out purposes, then the cutoff point could be moved to the right in order to increase sensitivity; if used for confirmatory purposes, the cutoff point could be moved

2. Recall that sensitivity equals

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

If the number of false negatives decreases, then sensitivity increases. Similarly, specificity equals

$$\frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

If false positives increase, then specificity decreases. The reverse is also true: increasing false negatives decreases sensitivity, while decreasing false positives increases specificity.
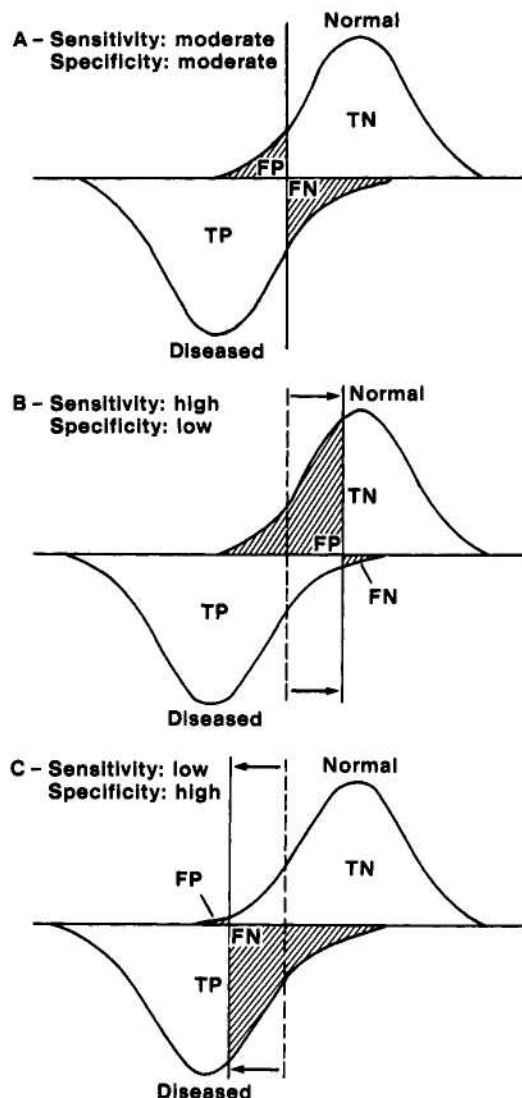
Figure 5.4
Effects on proportions of false positives (FP) and false negatives (FN) of moving "cut point" of a diagnostic test.

to the left to increase specificity. The test can be made suitable for either purpose simply by adjusting the cutoff point.

In summary, there is a lot more to a normal test result than first meets the eye. Unexpectedly abnormal results should be viewed with suspicion, especially when encountered in large test batteries obtained in apparently healthy individuals with low likelihoods of underlying disease. Normal test results, on the other hand, frequently allow revisions of diagnostic probabilities. Finally, the normal–abnormal cutoff point for any given test may be altered in order to make the test suitable for specific purposes.

Test Ordering Strategies and Their Misuse

Several diagnostic test ordering strategies routinely used by physicians are discussed in this section. Some have been shown to be inefficient and costly.

Strategy 1: "Routine" Admission Laboratory Testing

Since the advent of multichannel chemistry laboratory equipment, it has become fashionable "routinely" to order laboratory tests on every patient admitted to the hospital. The tests usually include a complete blood count, coagulation profile, urinalysis, serum electrolytes, some combination of serum chemistries, an electrocardiogram and a chest x-ray. These tests are ordered routinely, that is without regard to the actual admitting diagnosis or underlying probabilities. The two reasons usually cited to justify this process are that these tests are helpful in screening for asymptomatic disease and that they are useful in defining a "database," a baseline set of data in the light of which future changes can be evaluated.

Routine screening admission tests have been looked at critically by several investigators. In general, as I have described above, the problem is that as the number of routine tests increases, the likelihood of an unexpectedly abnormal test rises dramatically. In addition to the anxiety aroused in both the physician and the patient, such abnormal tests can result in further expense if they are repeated or investigated using more costly or invasive confirmatory tests. Furthermore, such abnormal results, even when persistent, sometimes have little impact on patient care. One study found that abnormal screening tests are often ignored by physicians and that, when pursued, lead only infrequently to positive diagnosis. Another study investigated 200 patients with unexpectedly abnormal screening tests, finding that in only 3 cases did such abnormalities lead to the diagnosis of presymptomatic disease. In the other cases, abnormalities could have been predicted in retrospect by careful consideration of the clinical status of the patient, or the tests were normal on repeated testing. Durbridge (1976) evaluated the benefits of routine screening admission testing in 500 patients, comparing them to two other similarly sized groups in which screening tests were either not obtained or were obtained but were withheld from attending physicians unless requested or deemed necessary by an independent advisory committee. No benefit was found in spite of a 64% increase in total laboratory costs and a 5% increase in total hospital bill. Most recently, Hubbell (1985) looked at the impact of routine chest x-rays on patient care, finding that results affected treatment only 4% of the time, in spite of the fact that the x-rays were obtained in a population of patients with a high prevalence of cardiopulmonary disease.

Studies of routine testing have therefore found that this strategy rarely leads to "new" diagnoses or to dramatic changes in patient care, in spite of the increased cost. It is important to note, however, that I have carefully defined "routine" as a test obtained in patients in whom you have little evidence to suspect underlying disease (i.e., patients in whom prior likelihood of disease is low because they have no symptoms or signs to suggest disease). In such patients, one would expect the predictive value of a positive test to be low, leading to the findings in the studies cited above. Specific admission testing should be done on patients in whom you have cause to believe that the prior likelihood of disease is higher, examples being serum potassium or magnesium in patients on diuretics, chest x-rays in patients with uncontrolled malignancies, or electrocardiograms in patients with chest pain. Tests obtained in such patients are not obtained for routine or screening purposes; instead, they are obtained for monitoring therapy or for diagnostic purposes in a population of patients with a high likelihood of disease.

The database concept is closely linked to that of routine screening. Many hospitals, especially those associated with training programs, have a defined database consisting of a battery of diagnostic tests obtained on every patient admitted. When this database is intended to screen for disease, the strategy is subject to the same criticisms discussed above. Many physicians, however, use a database as a "baseline" to which possible changes resulting from known disease and/ or therapy are compared. A few examples include measuring renal function in patients who are to undergo therapy with nephrotoxic antibiotics or cytotoxic agents, measuring serum electrolytes in patients who are to be placed on diuretics, or obtaining an electrocardiogram in patients at high risk for subsequent development of ischemic heart disease. In many such cases, a database serves a valuable purpose. Note, however, that the database is not "routine." It is used specifically in patients or groups of patients in whom the likelihood of present or future abnormalities is higher than the "usual" patient, a specific instance of monitoring.

If you decide to use a database, you should remember several points. First, the full database does not have to be repeated every time a patient is admitted. Unless events have occurred since the last admission that give you reason to believe certain database values have changed, the originally obtained values may suffice. Second, the same database does not have to be obtained for every patient, especially ambulatory patients. A healthy 26-year-old man who presents with an inguinal hernia does not require the same database as a 76-year-old man who presents with a severe pneumonia, simply because they differ both in likelihoods of underlying disease and in expected monitoring needs. The decision to obtain or repeat a database is therefore one that should be individualized with respect to the patient, and to what happens to the patient over time.

To summarize, "routine" admission or repeated database testing has not confirmed the high hopes of those who began advocating its use several decades ago when multichannel testing equipment was developed. Admission or database testing should never be routine; instead, it should be tailored to the differing needs of individual or selected groups of patients over time.

Strategy 2: "Rule-out" Testing

This strategy consists of ordering one or two laboratory tests in order to "rule out" every disease in a long list of diagnostic possibilities. Most often practiced when physicians are faced with complex diagnostic problems with broad differential diagnoses such as a fever of unknown origin, rule-out testing falls prey to many probabilistic pitfalls.

The major problem with rule-out strategies is that they ignore prior probabilities. A test is ordered to investigate every disease possibility regardless of how likely or unlikely that possibility is. Because of the relatively large number of tests ordered using this strategy, the likelihood that one or more tests will be abnormal by chance is rather high. In addition, given the poor specificity of many lab tests, the positive predictive value of any one abnormal test will be low, especially if the prior likelihood of disease in question was low to begin with. Rule-out testing, when utilized in this fashion, often leads to more, rather than less, diagnostic uncertainty.

A better strategy might be developed by formulating a

differential diagnosis based on the preliminary assessment of the patient, and then ordering laboratory tests only for those diseases deemed most likely. By choosing tests for this "most likely" subset of possibilities, a diagnosis may be confirmed without the unnecessary expense and confusion involved in investigating the entire list.[3] If by such testing all possibilities are ruled out, then the revised probabilities of this subset drops and those of a second subset rise. The second stage of investigation involves ordering laboratory or diagnostic tests to evaluate this second list. If none of these possibilities is confirmed, progress to a third subset, and so on. This strategy might be called the "ordered subset" strategy. Evaluating successive subsets of an originally lengthy differential diagnosis avoids the pitfalls of multiple testing and poor predictive value involved in the less efficient and more costly rule-out strategy.

Strategy 3: Combination or Parallel Testing

Two or more tests, varying in sensitivity and specificity, are available for many diseases. The choice of test, either highly specific or highly sensitive, depends on whether one wants to confirm or rule out the disease. Why not use both tests together? This strategy is termed *combination* or *parallel* testing.

Imagine a disease for which the prior likelihood is 50%. There are two tests, A and B, each differing in sensitivity and specificity. Test A is very sensitive (95%), but has lower specificity (90%); Test B is poorly sensitive (80%), but highly specific (95%). You decide to use both tests. Furthermore, you decide that you will diagnose disease if *either* Test A *or* Test B is positive. Table 5.5 shows the sensitivity and specificity of this parallel combination of tests. When used in this fashion, the sensitivity of the combination is higher than either test alone, but the specificity is much lower. Using tests in such combinations generally results in increased sensitivity and lower specificity. The negative predictive value of these combinations, therefore, will generally be very high, but the positive predictive value will be very low when compared to any one single test.[4]

Whether or not you select this strategy is determined by your purpose. If you wish to rule out disease, parallel combinations may be useful because of their high negative predictive values. If your purpose is to confirm disease, the low positive predictive value of parallel combinations may be unsuitable. Combination testing is common in clinical medicine, examples including "liver function tests" (SGOT, total bilirubin, alkaline phosphatase, and albumin), or "rule-out myocardial infarction" (CPK, SGOT, LDH) batteries.

Whatever your purpose, remember that combinations are *always* more expensive than single tests. The decision to use a combination is therefore affected not only by probability considerations but by costs as well.

Table 5.5
Effect of Test Combination on Sensitivity and Specificity

Test(s)	Sensitivity	Specificity
A only	95%	90%
B only	80	95
A or B	99	85

Source: From Galen, 1975, 43.

Table 5.6
Rules for Repeating Laboratory Tests

	Test Results		
Rule	Initial	Repeat	Action
Believe the positive	+	+	Further diagnostic or therapeutic action
	+	−	
	−	+	
	−	−	No further action
Believe the negative	+	+	Further diagnostic or therapeutic action
	+	−	No further action
	−	+	
	−	−	

Source: Adapted from Politser, 1982.

Strategy 4: Repeating Tests

There are two situations in which the same lab test is repeated: when a test result is unexpectedly abnormal, and when multiple tests are routinely ordered, as with serial stool occult blood examinations or serial blood cultures. In some respects, this strategy is analogous to combination testing except that multiple repetitions of the same test are used rather than combinations of different tests.

When using this strategy, the clinician is faced with a choice between two rules: "believe the positive" or "believe the negative." With the first rule, any positive test in the repeated set is used to confirm the diagnosis, whereas with the second rule any negative in the set is used to rule out the diagnosis. These rules are illustrated in Table 5.6 for tests that are repeated twice. Deciding which rule to use may be considerably more difficult than it seems. In general, the "believe the positive" rule will tend to increase sensitivity at the expense of specificity since, over multiple repetitions, false positive results carry more weight than false negative results. Similarly, the "believe the negative" rule tends to increase specificity at the expense of sensitivity. The often reflex act of "simply" repeating an abnormal test may not be as simple as we would like to believe![5]

Conclusion

Much of the costly misuse and misinterpretation of diagnostic tests is due to the fact that physicians are not always able to manage intuitively the probabilistic information that most laboratory tests provide. The diagnostic process is one

---

3. Considerations other than just probability may enter here. Perhaps the most relevant is the need to diagnose certain diseases early in their course, when the consequences of even short delays can be life-threatening. The possibility of such a disease may prompt diagnostic testing even if the likelihood is fairly low.

4. This analysis is based on the assumption of conditional independence. Basically, this assumption dictates that the sensitivity and specificity of any given test are independent of the results of any of the other tests. This assumption may not always be justified.

5. Politser (1982) discusses this issue in detail, and it can be quite more complex. If one considers test reliability in both diseased and nondiseased populations, these rules may either increase or decrease both sensitivity and specificity!

of successive probability revision. Knowledge of the operating characteristics of laboratory tests (i.e., sensitivity, specificity, and predictive value) can greatly facilitate this process. Knowledgeable use of these operating characteristics allows test choices to be tailored to the specific purposes of diagnosis, monitoring, and screening. In addition, such use allows proper interpretation of normal test results, as well as discriminate use of common strategies such as screening, rule-out testing, repeating tests, and combination testing.

References

Berwick DM, Fineberg HC, Weinstein MC. When doctors meet numbers. Am J Med 1981;71:991–98.

Borak J, Veilleux S. Errors of intuitive logic among physicians. Soc Sci Med 1982;16:1939–47.

Bradwell AR, Carmalt MHB, Whitehead TP. Explaining the unexpected abnormal results of biochemical profile investigations. Lancet 1974;1:1071–74.

Casscells W, Schoenberger A, Graboys TB. Interpretation by physicians of clinical laboratory results. N Engl J Med 1978;299:999–1001.

Christensen-Szalanski JJ, Bushyhead JB. Physicians' misunderstanding of normal findings. Med Decis Making 1983;3:169–75.

Collen MF, Feldman R, Siegelaub AB, et al. Dollar cost per positive test for automated multiphasic screening. N Engl J Med 1970;283:459–63.

Connelly D, Steele B. Laboratory utilization: problems and solutions. Arch Pathol Lab Med 1980;104:59–62.

Detmer DE, Fryback DG, Gassner K. Heuristics and biases in medical decision-making. J Med Ed 1978;53:682–83.

Dixon RH, Laszlo J. Utilization of clinical chemistry services by medical housestaff. Arch Intern Med 1974;134:1064–67.

Durbridge TC, Edwards F, Edwards RG, et al. Evaluation of benefits of screening tests done immediately on admission to hospital. Clin Chem 1976;22:968–71.

Elstein AS, Shulman LS, Sprafka S, et al. Medical problem solving: an analysis of clinical reasoning. Cambridge: Harvard University Press, 1978.

Elveback LR, Guillier CL, Keating FR. Health, normality, and the ghost of Gauss. JAMA 1970;211:69–75.

Fineberg HV. Clinical chemistries: the high cost of low-cost diagnostic tests. In: Altman S, Blendon R, eds. Medical technology: the culprit behind health care costs? DHEW Publication (PHS)79–3216. Washington D.C.: GPO, 1979.

Ford HC. Use and abuse of clinical laboratories. NZ Med J 1978;88:16–18.

Freeland MS, Schendler CE. National health expenditure growth in the 1980's: an aging population, new technologies, and increasing competition. Health Care Financing Review 1983;4:1–58.

Galen RS, Gambino SR. Beyond normality: the predictive value and efficiency of medical diagnosis. New York: John Wiley, 1975.

Gorry GA, Pauker SG, Schwartz WB. The diagnostic importance of the normal finding. N Engl J Med 1978;298:486–89.

Griner PF, Glaser RJ. Misuse of laboratory tests and diagnostic procedures. N Engl J Med 1982;307:1336–39.

Griner PF, Liptzin B. Use of the laboratory in a teaching hospital. Ann Intern Med 1971;75:157–63.

Griner PF, Mayewski RJ, Mushlin AI, et al. Selection and interpretation of diagnostic tests and procedures. Ann Intern Med 1981;94:553–600.

Hubbell FA, Greenfield S, Tyler JL, et al. The impact of routine admission chest x-ray films on patient care. N Engl J Med 1985;312:209–13.

Kassirer JP, Gorry GA. Clinical problem-solving: a behavioral analysis. Ann Intern Med 1978;89:245–55.

Moloney TW, Rogers DE. Medical technology—a different view of the contentious debate over costs. N Engl J Med 1979;301:1413–19.

Politser P. Reliability, decision rules, and the value of repeated tests. Med Decis Making 1982;2:47–69.

Schneiderman LJ, DeSalvo L, Baylor S, et al. The "abnormal" screening laboratory result. Arch Intern Med 1972;129:88–90.

Schoen I, Brooks SH. Judgment based on 95% confidence limits. Am J Clin Pathol 1970;53:190–95.

Schottenfeld D. Fundamental issues in cancer screening. In: Winawer S, Schottenfeld D, Sherlock P, eds. Colorectal cancer: prevention, epidemiology, and screening. New York: Raven Press, 1980:167–74.

Simon JB. Occult blood screening for colorectal carcinoma: a critical review. Gastroenterology 1985;88:820–37.

Snapper KJ, Fryback DG. Inferences based on unreliable reports. J Exp Psychol 1971;87:401–4.

Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: W.B. Saunders, 1980.

Weintraub WS, Madeira SW, Bodenheimer MM, et al. Critical analysis of the application of Bayes' theorem to sequential testing in the noninvasive diagnosis of coronary artery disease. Am J Cardiol 1984;54:43–49.

Wertman BG, Sostrin SV, Pavlova Z, et al. Why do physicians order laboratory tests? JAMA 1980;243:2080–82.

Wong ET, Lincoln TL. Ready! Fire!. . . Aim! An inquiry into laboratory testing. JAMA 1983;250:2510–13.