An Introduction to Diagnostic Tests

Stephen D. Simon

The Children's Mercy Hospitals and Clinics

What is a diagnostic test?

A diagnostic test is a procedure which gives a rapid, convenient and/or inexpensive indication of whether a patient has a certain disease. Some examples of diagnostic tests are:

QTc dispersion.

A standard electrocardiogram can produce a measure called QTc dispersion. In a study of 49 patients with peripheral vascular disease (Darbar 1996), all were assessed for their QTc dispersion values. These patients were then followed for 52 to 77 months. During this time, there were 12 cardiac deaths, 3 non-cardiac deaths, and 34 survivors. A value of QTc dispersion of 60 ms or more did quite well in predicting cardiac death.

Yale-Brown obsessive-compulsive scale.

The Yale-Brown obsessive-compulsive scale, a simple yes/no answer to the following question: *Do you often feel sad or depressed?* In a study of stroke patients at the Royal Liverpool and Broadgreen University Hospitals (Watkins 2001), this test was shown to perform well compared to a more complex measure, the Montgomery Asberg depression rating scale.

Rectal bleeding.

Patients with rectal bleeding will sometimes develop colorectal cancer. In a study at a network of practices in Belgium (Wauters 2000), 386 patients presented with rectal bleeding between 1993 and 1994. After following these patients for 18 to 30 months, only a few developed colorectal cancer.

To assess the quality of a diagnostic test, you need to compare it to a gold standard. This is a measurement that is slower, less convenient, or more expensive than the diagnostic test, but which also gives a definitive indication of disease status. The gold standard might involve invasive procedures like a biopsy or could mean waiting for several years until the disease status becomes obvious.

You classify patients as having the disease or being healthy using the gold standard. Then you count the number of times that the diagnostic test agrees and disagrees with the gold standard of disease and the number of times that the diagnostic test agrees and disagrees with the gold standard of being healthy.

This leads to four possible categories.

- TP (true positive) = # who test positive and who have the disease,
- FN (false negative) = # who test negative and who have the disease,
- FP (false positive) = # who test positive and who are healthy, and
- TN (true negative) = # who test negative and who are healthy.

See the figure below for a graphical layout of these results.

		Test Positive (T+)	Test Negative (T-)
Disease Present	(D+)	True Positive (TP)	False Negative (FN)
Disease Absent	(D-)	False Positive (FP)	True Negative (TN)

A good diagnostic test will minimize the number of false negative and false positive results.

What are the economic consequences of a bad diagnostic test?

The New York Times had an excellent article on newborn screening tests (Kolata 2005). It discusses a recent push to standardize and expand the screening tests for newborns to include 29 different diseases. It seems like such an obvious thing to do: let's screen for these conditions, because the more we know, the better we are able to care for these children.

> *Proponents say that the diseases are terrible and that an early diagnosis can be lifesaving. When testing is not done, parents often end up in a medical odyssey to find out what is wrong with their child. By the time the answer is in, it may be too late for treatment to do much good.*

Opponents, however, point out that false positive results may present more problems.

> *But opponents say that for all but about five or six of the conditions, it is not known whether the treatments help or how often a baby will test positive but never show signs of serious disease. There is a danger, they say, of children with mild versions of illnesses being treated needlessly and aggressively for more serious forms and suffering dire health consequences.*

The article also offers a historical perspective by citing phenylketonuria (PKU) testing as an example. An infant with PKU cannot metabolize phenylalanine, and the build up of this amino acid can lead to serious neurological damage. The treatment, a diet low in phenyalanine, is very effective, but only if the condition is diagnosed early. The PKU testing done today is very good, but tests performed 45 years ago had problems.

Back then, any infant who tested positive would be put on this special diet. When phenylalanine is withdrawn from the diet of a healthy infant, that infant suffers from even more serious neurological problems and can even die. Many infants who falsely tested positive were put on this diet and their harms outweighed the benefits of PKU screening. As researchers learned more, they were able to refine the test to prevent most false positives, but the damage had already been done.

Another New York Times article (Kolata 2003), documented the patient demand for diagnostic tests even when they have no rational basis.

> *Even doctors who know all about the evidence-based guidelines for preventive medicine say they often compromise in the interest of keeping patients happy. Dr. John K. Min, an internist in Burlington, N.C., tells the story of a 72-year-old patient who came to him for her annual physical, knowing exactly what tests she wanted. She wanted a Pap test, but it would have been useless, Dr. Min said, because she had had a hysterectomy. She wanted a chest X-ray, an electrocardiogram. Not necessary, he told her, because it was unlikely that they would reveal a problem that needed treating before symptoms emerged. She left with just a few tests, including blood pressure and cholesterol. Dr. Min was proud of himself until about a week later, when the local paper published a letter from his patient - about him. "Socialized medicine has arrived," she wrote. Admitting defeat, he called her and offered her the tests she had wanted, on the house. She accepted, Dr. Min said, but after having the full physical exam, she never returned.*

How does prevalence affect performance?

Prevalence is the proportion of patients who have the disease in the population you are testing. This can vary quite a bit in real situations. For example, the prevalence of a disease is often much higher in a tertiary care center than at a primary care physician's office. Prevalence can also vary sometimes by seasons of the year. It can also vary sometimes by race or gender.

Prevalence plays a large role in determining how effective a diagnostic test is. In general, when the prevalence of the disease you are testing is rare, it becomes harder to positively diagnose that disease.

This is the source of controversy over many screening tests such as mammograms. There is no controversy over these tests for older women, or for women at higher risk for breast cancer because of a specific genetic marker or a family history of the disease.

The controversy over mammograms occurs with younger women (40-50 years old) who have no known risk factors for breast cancer. A careful analysis of the controversy is beyond my skills, but I can outline the issues that have to be evaluated.

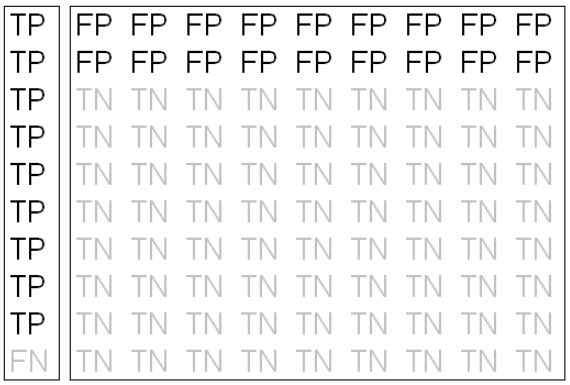
First, what is the cost of misdiagnosis in the mammogram? A false negative result will prevent a woman from seeking treatment for breast cancer. You won't prevent treatment forever, because sooner or later, the cancer is going to become overtly noticeable through other diagnostics, such as a breast self-exam. The loss is the lost time. The cost of a false positive is the economic cost of the unnecessary biopsy, plus the psychological cost of the anxiety produced by the false positive test. You need to tally the two costs, adjusted by the relative proportion of false positives and false negatives.

Now tally the cost of failing to seek a mammogram. Tally up the increase in costs when the true positives under the mammogram become false negatives under the no test option. Tally up the decrease in costs when the false positives become true negatives (it's impossible to have a false positive if you never test). I can't tell you which way the scales would tip, of course, because I am not an expert on breast cancer.

Let's look at a hypothetical situation. In the graph below, patients on the left have the disease and patients on the right are healthy.

This situation represents a disease with high prevalence. A positive test is reasonably definitive because the number of true positives is much larger than the number of false positives.

Let's consider a different hypothetical situation.



In this situation, the prevalence of the disease is much lower. Since there are more healthy patients, their false positive results swamp the true positive results.

What is sensitivity?

The sensitivity of a test is the probability that the test is positive when given to a group of patients with the disease. Sensitivity is sometimes abbreviated *Sn*. The formula for sensitivity is

$Sn = TP / (TP + FN)$

where *TP* and *FN* are the number of true positive and false negative results, respectively. You can think of sensitivity as 1- the false negative rate. Notice that the denominator for sensitivity is the number of patients who have the disease.

The following table summarizes these calculations.

	Test Positive (T+)	Test Negative (T-)
Disease Present (D+)	True Positive (TP)	False Negative (FN)
Disease Absent (D-)	False Positive (FP)	True Negative (TN)

Sensitivity (Sn) = TP / (TP + FN)

A large sensitivity means that a negative test can rule out the disease. David Sackett coined the acronym "SnNOut" to help us remember this.

Example: serum pepsinogen.

In a study of 5,113 subjects checked for gastric cancer by endoscopy (Kitahara 1999), serum pepsinogen concentrations were also measured. A pepsinogen I concentration of less than 70 ng/ml and a ratio of pepsinogen I to pepsinogen II of less than 3 was considered a positive test. There were 13 patients with gastric cancer confirmed by endoscopy. 11 of these patients were positive on the test. The sensitivity is 11/13 = 85%.

What is specificity?

The specificity of a test is the probability that the test will be negative among patients who do not have the disease. Specificity is sometimes abbreviated *Sp*. The formula for specificity is

$Sp = TN / (TN + FP)$

where *TN* and *FP* and the number of true negative and false positive results, respectively. You can think of specificity as 1 - the false positive rate. Notice that the denominator for specificity is the number of healthy patients.

The following table summarizes these calculations.

	Test Positive (T+)	Test Negative (T-)
Disease Present (D+)	True Positive (TP)	False Negative (FN)
Disease Absent (D-)	False Positive (FP)	True Negative (TN)

Specificity (Sp) = TN / (TN + FP)

A large specificity means that a positive test can rule in the disease. David Sackett coined the acronym "SpPIn" to help us remember this.

Example: urine latex agglutination test.

In a study of the urine latex agglutination test (reference misplaced, sorry!), children were tested for H. influenzae using blood, urine, cerebrospinal fluid, or some combination of these. Of all the children tested, 1,352 did not have H. influenzae in any of these fluids. Only 9 of these patients tested positive on the urine latex agglutination test, the remaining 1,343 tested negative. The specificity is 1343 / 1352 = 99.3%.

What is a positive predictive value?

The positive predictive value of a test is the probability that the patient has the disease when restricted to those patients who test positive. This term is sometimes abbreviated as *PPV*. You can compute the positive predictive value as

$$PPV = TP / (TP + FP)$$

where *TP* and *FP* are the number of true positive and false positive results, respectively. Notice that the denominator for positive predictive value is the number of patients who test positive.

The following table summarizes these calculations.

		Test Positive	(T+)	Test Negative	(T-)
Disease Present	(D+)	True Positive	(TP)	False Negative	(FN)
Disease Absent	(D-)	False Positive	(FP)	True Negative	(TN)

Positive Predictive Value (PPV) = TP / (TP + FP)

Do not calculate the positive predictive value on a sample where the prevalence of the disease was artificially controlled. For example, the PPV is meaningless in a study where you artificially recruited healthy and diseased patients in a one to one ratio.

Example: rectal bleeding.

In a study of patients in a network of sentinel practices in Belgium (Wauters 2000), 386 patients presented with rectal bleeding. These patients were followed from 18 to 30 months and 27 of them developed colorectal cancer. The positive predictive value for rectal bleeding is 27 / 386 = 7%.

What is the negative predictive value?

The negative predictive value of a test is the probability that the patient will not have the disease when restricted to those patients who test negative. This term is sometimes abbreviated as *NPV*. You can compute the negative predictive value as

$$NPV = TN / (TN + FN)$$

where *TN* and *FN* are the number of true negative and false negative results, respectively. Notice that the denominator for negative predictive value is the number of patients who test negative.

The following table summarizes these calculations.

		Test Positive	(T+)	Test Negative	(T-)
Disease Present	(D+)	True Positive	(TP)	False Negative	(FN)
Disease Absent	(D-)	False Positive	(FP)	True Negative	(TN)

Negative Predictive Value (NPV) = TN / (TN + FN)

Do not calculate the negative predictive value on a sample where the prevalence of the disease was artificially controlled. For example, the NPV is meaningless in a study where you artificially recruited healthy and diseased patients in a one to one ratio.

Example: depression.

In a study of depression among 79 patients hospitalized for stroke (Watkins 2001), 34 patients responded "no" to the question: *Do you often feel sad or depressed?* Among these 34 patients who tested negative, 6 had clinical depression as defined by a more complex measure, the Montgomery Asberg depression rating scale. Since 28 did not have depression, the negative predictive value is 28 / 34 = 82%.

What is the likelihood ratio?

You can summarize information about the diagnostic test itself using a measure called the likelihood ratio. The likelihood ratio combines information about the sensitivity and specificity. It tells you how much a positive or negative result changes the likelihood that a patient would have the disease.

The likelihood ratio incorporates both the sensitivity and specificity of the test and provides a direct estimate of how much a test result will change the odds of having a disease (see the appendix for an explanation of odds). The likelihood ratio for a positive result (LR^+) tells you how much the odds of the disease increase when a test is positive. The likelihood ratio for a negative result (LR^-) tells you how much the odds of the disease decrease when a test is negative.

The positive likelihood ratio is

$$LR^+ = Sn / (1 - Sp).$$

You want to see a large value for LR^+. This can occur if the numerator of the fraction is large, or the denominator is small. Since it is impossible to get the numerator any larger than one, the only practical way to get a large value for LR^+ is to make the denominator small. This occurs when *Sp* is close to one. This is consistent with the David Sackett acronym **SpPIn** (if the **sp**ecificity of test is large, then a **p**ositive test will help rule **in** the diagnosis.

The negative likelihood ratio is

$$LR^- = (1 - Sn) / Sp.$$

You want to see a small value for LR^-. This can occur when the numerator of the fraction is small or the denominator is small. Since the denominator cannot get any larger than one, the only practical way to to get a small value is to make the numerator small. This occurs when *Sn* is close to 1. This is consistent with the David Sackett acronym SnNOut (if the **sen**sitivity of a test is large, then a **n**egative test will help rule **out** the diagnosis).

What's a good value for a likelihood ratio? There are no absolute boundaries, but here are some general rules. For a positive likelihood ratio, anything less than 2 is worthless. A good likelihood ratio should be 10 or higher. Anything bigger than 50 represents an excellent diagnostic test. For a negative likelihood ratio (LR^-), the corresponding boundaries are 0.5 (1/2), 0.1 (1/10), and 0.02 (1/50).

Some diagnostic tests will have a good LR^+, but a poor LR^-. This might be entirely appropriate if the cost of a false positive is far greater than the cost of a false negative. In a setting where a false negative is a bigger concern, a mediocre LR^+ might be acceptable if combined with a robust LR^-.

You combine the likelihood ratio with information about

1. the prevalence of the disease,
2. characteristics of your patient pool, and
3. information about this particular patient

to determine the post-test odds of disease.

If you want to quantify the effect of a diagnostic test, you have to first provide information about the patient. You need to specify the pre-test odds: the likelihood that the patient would have a specific disease prior to testing. The pre-test odds are usually related to the prevalence of the disease, though you might adjust it upwards or downwards depending on characteristics of your overall patient pool or of the individual patient.

This process of specifying pre-test odds is very important because you have to adapt the diagnostic test to the patient rather than the patient to the diagnostic test.

A simple example using likelihood ratios.

An early test for developmental dysplasia of the hip. The test has 92% sensitivity and 86% specificity in boys (AJPH 1998; 88(2): 285-288). This paper does not compute likelihood ratios, so you have to do a few calculations yourself.

$LR^+ = Sn / (1 - Sp) = 0.92/0.14 = 6.6.$
$LR^- = (1 - Sn) / Sp = 0.08/0.86 = 0.09.$

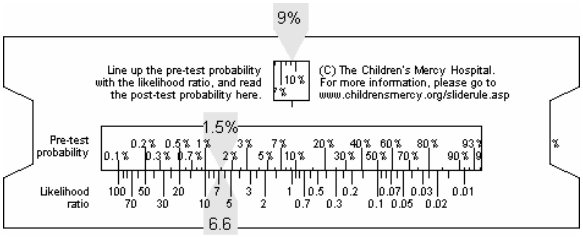
Suppose one of our patients is a boy with no special risk factors. The diagnostic test is positive. What can we say about the chances that this boy will develop hip dysplasia? The prevalence of this condition is 1.5% in boys. This corresponds to an odds of one to 66. Multiply the odds by the likelihood ratio, you get 6.6 to 66 or roughly 1 to 10. The post test odds of having the disease is 1 to 10 which corresponds to a probability of 9%.

Suppose we had a negative result, but it was with a boy who had a family history of hip dysplasia. Suppose the family history would change the pre-test probability to 25%. How likely is hip dysplasia, factoring in both the family history and the negative test result? A probability of 25% corresponds to an odds of 1 to 3. The likelihood ratio for a negative result is 0.09 or 1/11. So the post-test odds would be roughly 1 to 33, which corresponds to a probability of 3%.

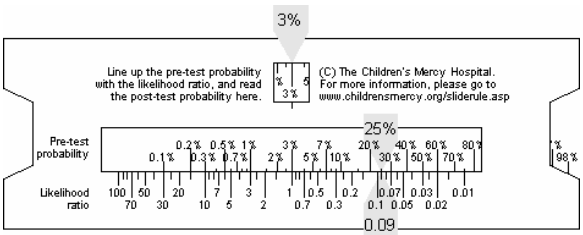
The use of likelihood ratios requires a bit of tedious calculations. I have developed a simple slide rule that will do likelihood ratio calculations for you.

Slide the insert up or down until the pre-test probability in the left window lines up with the likelihood ratio. Read the post-test probability in the right window.

Let's show how the slide rule would work for the hip dysplasia example. The prevalence of this condition is 1.5%, and since there are no unusual risk factors, we will use this as the pre-test probability. Line up this with the value for LR^+ (6.6) and read a post-test probability of 9%.



Suppose that instead the patient had a family history that raised the pre-test probability to 25%. The test, thankfully, is negative. For this test, you line up the 25% pre-test probability with the value of LR^- (0.09) to get a post-test probability of 3%.



Notice that the change is more dramatic for the second case rather than the first case. There are two things that account for this.

First, a diagnostic test is most useful and shows the largest change in disease probability when that probability is in the middle (somewhere between 20% and 80%). When the pre-test probability is very close to 0% or very close to 100%, it is hard to move the probabilities very much.

The second factor at work here is that this test was already slightly better at ruling out a diagnosis than ruling it in since LR^- (0.09 or 1/11) is more extreme than LR^+ (6.6).

The likelihood ratio slide rule that I developed was inspired by the Fagan nomogram (Fagan 1975). It is a bit more complex to make, but it calculates more rapidly, and it is small enough to fit in your shirt pocket.

How do you estimate the pre-test odds?

The likelihood ratio is a measure of how much the odds of disease change when you get a particular result for a diagnostic test. In a practical setting, you first specify the pre-test odds. This is the odds that you assess that a patient has the disease prior to any testing.

When you are estimating a pre-test odds for a diagnostic test, take three steps:

1. find an estimate of the prevalence of the disease in the general population,
2. modify this estimate based on characteristics of your particular practice, and
3. further modify this estimate based on characteristics of the individual patient that is currently sitting in front of you.

The estimate of prevalence is often found in the literature. For certain diseases that are seasonal, you may wish to use a different estimate of prevalence in the winter months compared to the summer months.

Where you practice can sometimes make a big difference in the pre-test odds, because of the way patients are filtered and funneled. Someone practicing in a secondary or tertiary care setting will often see high rates of certain diseases because someone ahead of you in the queue will remove many of the obvious cases of non-disease.

You should also adjust the pre-test odds based on the patient sitting in front of you. If you discover an important risk factor while taking the patient's history that is known to influence the disease, adjust the odds up or down.

If your patient has diabetes, you should increase the pre-test probability estimates of arteriosclerosis, retinopathy, and renal disease. If your patient has a long history of cocaine abuse, you should increase the pre-test probability of various sinus and nasal diseases. If your patient has a sister who was diagnosed with breast cancer at the age of 45, you should increase the pre-test probability of breast cancer for this patient.

If you notice something unusual during the physical exam, try to take this into account as well. These risk factors will vary depending on the disease that you are trying to diagnose. There are some efforts, such as the Rational Clinical Exam series in JAMA that try to quantify risk factors collected during the history and physical using likelihood ratios, and you should be familiar with these.

There is nothing wrong, however, in using a bit of subjective judgment in assessing pre-test odds. A big criticism of evidence-based medicine is that it does not allow for your personal clinical judgment or the characteristics of the individual patient to enter into the equation. Well, here's your chance to use your judgment and avoid the application of "one-size-fits-all" medicine.

There's an amusing story about screening for alcohol abuse. It turns out (not surprisingly) that alcohol abuse is very much dependent on the age and gender of the patient which is important for your Step-3 adjustment, but another interesting fact is that the rate of abuse in an outpatient setting is about twice that of the rate in the general population. For an inpatient setting, the rate is four times higher. These are Step-2 adjustments.

Why is it that patients in an outpatient or inpatient setting have a much greater probability of alcohol abuse? Does being around doctors so much drive people to drink? Are they depressed because they are stuck in the hospital so much? The answer, of course, is actually quite logical. People who abuse alcohol tend to have more health problems than the general population and tend to be overrepresented in outpatient and inpatient settings.

Evaluating a post-test odds/probability

So what do you do with the post-test odds or probabilities? To answer that question, you need to first specify what the costs of a false positive diagnosis is and what the cost of a false negative diagnosis is.

The costs of a false positive depend on what the next step would be if you have convincing evidence that the patient has the disease you are looking for. In some cases, the next step if a diagnostic test is positive is to run a more expensive diagnostic test. The risk of the additional diagnostic test is usually small, so the only major consideration in this setting is the wasted resources by running an unnecessary additional test, plus the temporary unwanted anxiety produced by leaving the patient's status on hold while you wait for the second diagnostic test.

When the costs of a false positive are very low, then you should take the next step even if the post-test probability is as small as 10 or 20%.

In other situations, the costs of a false positive are quite large. When the next step if you are convinced that the patient has the disease is a high-risk operation, then you want to be more conservative. After all, one of the worst things you can do is to cut open a patient who is completely healthy.

When the next step following a positive diagnosis would be very risky, then you should not take such a step unless the post-test probability is larger than 70 or 80%.

There are costs associated with false negatives, as well, and sometimes these costs dominate the consideration. Consider a diagnostic test for head and neck injuries that is intended to rule out the possibility of a cervical fracture. The rationale is to minimize the cost and risks associated with unnecessary x-rays. In such a setting, false negatives represent patients that you skip the x-ray on and send home even though they have a fracture in a rather important part of the body. These are the people who show up a day later paralyzed and with an army of lawyers ready to litigate.

When the cost of a false negative is high, a small post-test probability may warrant taking the next step.

When the costs of a false positive and a false negative are both very high, you make sure that you stay current on your malpractice insurance. Actually, the key here is not the absolute costs, but the relative costs. A simple rule is that if the ratio of costs of a false positive to a false negative is x, then set the post-test odds (not probability) threshold to the same value. For example if a false positive is twice as serious as a false negative, then you should take the next step if the post-test odds of disease are 2 to 1. This corresponds to a post-test probability of 0.67.

Keep in mind that the costs of false positives and false negatives are derived using the patient's perspectives and values not yours. Some of the risks and costs associated with misdiagnosis are highly variable and you need to take the time to understand what your patient values as important.

Suppose you are testing for allergic reaction to foods. If the test is a false positive, it would mean that you advise your patients unnecessarily to avoid foods that are actually safe for them to eat. In my value system, I would place a very low cost if the food I had to avoid was Brussels sprouts, but a much higher cost if the food I could no longer eat was chocolate. Another patient, of course, might miss Brussels sprouts more than they would miss chocolate.

This analysis is somewhat simplistic. Still it is a valuable exercise to go through and even a simplistic evaluation of costs is preferable to using a "one-size-fits-all" approach to diagnostic testing.

When is a diagnostic test unnecessary?

A diagnostic test is unnecessary when the pre-test probability is so low that even a positive test result will not shift you to a post-test probability large enough to warrant taking the next step OR if the pre-test probability is so high that even a negative test result will not shift you to a post-test probability small enough to discourage you from taking the next step.

This should make good intuitive sense. If you are already quite certain that the patient has the disease or quite certain that the patient does not have the disease, then a diagnostic test becomes irrelevant. Tests are valuable for those patients in the uncertain middle.

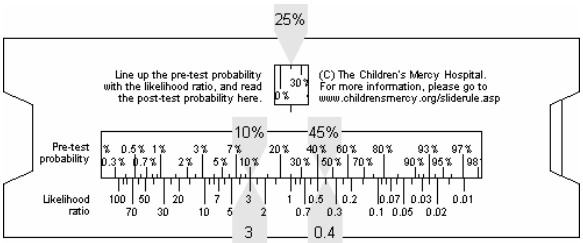
Your calculations also need to take into account the likelihood ratios for the test. If a test has very strong likelihood ratios, then it can be persuasive even if you have a very strong prior belief about the probability that this patient has the disease. If the test has very weak likelihood ratios, then it is only persuasive if you have substantial uncertainty.

You can use the likelihood ratio slide rule to calculate the range of uncertainty that would justify use of the diagnostic test.

Adjust the slide rule so that the threshold post-test probability appears in the upper window. Then note the pre-test probabilities that line up with the values of LR^+ and LR^- of the diagnostic test. These represent the range of pre-test probabilities where the diagnostic test can make a difference in your clinical decision.

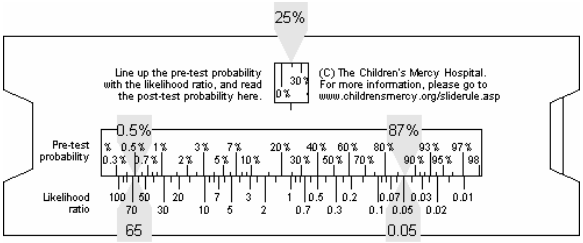
Here's a hypothetical example. Suppose the cost of a false positive is one third the cost of a false negative. You will treat any patient if their post-test odds are one to three, which is equivalent to a 25% post-test probability. The likelihood ratios for this diagnostic test are LR^+ = 3 and LR^- = 0.4. Note that if both of these values were just a bit weaker, we would call the diagnostic test worthless.

Line up the slide rule so that 25% shows in the post-test probability window. The likelihood ratios of 3 and 0.4 correspond to pre-test probabilities of 10% and 45% respectively.



If there is a moderate amount of uncertainty, this test can help. If, however, there is only a small chance that this patient has the disease (less than 10%), then a positive test result will not provide sufficiently persuasive evidence to justify taking the next step. If you have a moderate to strong belief prior to testing that this patient does have the disease (anything more than 45%), then a negative test result will not provide sufficiently persuasive evidence to talk you out of taking the next step.

Compare this to a diagnostic test that has LR^+ = 65 and LR^- = 0.05. These are very good values for a diagnostic test, especially for ruling in a diagnosis.



You should still have 25% in the post-test probability window. The likelihood ratios of 65 and 0.05 correspond to pre-test probabilities of 0.5% and 87% respectively. Unless your patient is at the extremes of certainty, you will find this diagnostic test very helpful.

Making prevalence adjustments

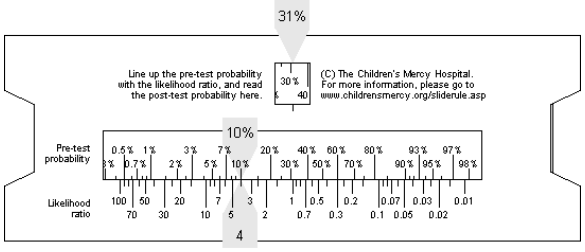
In Watkins (2001), a single question diagnostic test (the Yale-Brown obsessive-compulsive scale) was compared to a "gold standard" measure of depression, the Montgomery Asberg depression rating scale (MADRS). In this study, the values of *Sn* and *Sp* were 86% and 78%. From these values, you can compute LR^+ = 3.9 and LR^- = 0.18 (round these to 4 and 0.2 for simplicity).

Although the authors also computed PPV and NPV, the prevalence of depression in this population was unusually high (43%). The authors presented additional positive predictive values (*PPV*) and negative predictive values (*NPV*) for prevalence values ranging from 10% to 90% (see below).

Prevalence	PPV	NPV
10%	30%	98%
20%	49%	96%
30%	63%	93%
40%	72%	89%
50%	80%	85%
60%	85%	79%
70%	90%	70%
80%	94%	58%
90%	97%	38%

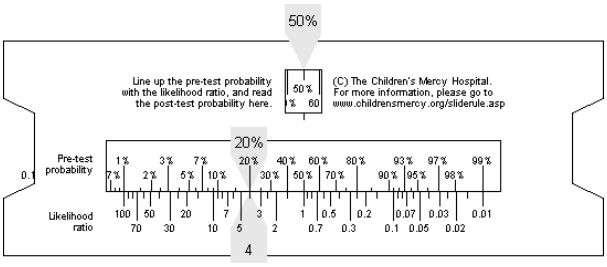
Since the *PPV* is simply the post-test probability after a positive test, we can use the likelihood ratio slide rule to re-create their calculations.

To compute the positive predictive value when the prevalence of the disease is 10%, line up the 10% pre-test probability with the likelihood ratio of 4 (the unlabelled tick mark between 3 and 5).



In the right side window, the post-test probability should be slightly more than 30%, which matches the value computed by Watkins.

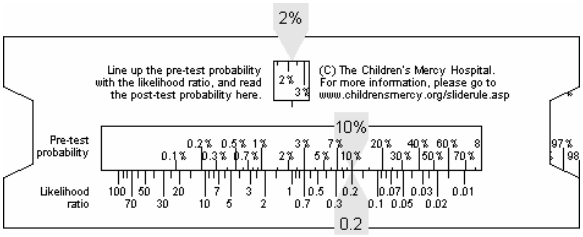
Slide the insert so the pre-test probability of 20% lines up with the likelihood ratio of 4.



The post-test probability should be around 50% which also matches the value in Watkins.

Repeat this for 30%, through 90% and see if you can estimate the remaining PPV values.

To adjust NPV for prevalence, line up the prevalence of 10% with the likelihood ratio of 0.2.



Read off the post-test probability of 2%. Since there is only a 2% chance of having the disease, there is a 98% of being healthy, which matches the NPV computed by Watkins.

Repeat this for 20% through 90% and see if you can estimate the remaining NPV values. Remember that the value shown on the slide rule is probability of disease. Subtract it from 100% to get the NPV.

The value of the likelihood ratio slide rule is that it allows you to rapidly recompute the PPV and NPV values for varying prevalence rates.

What about tests with multiple levels?

In a diagnostic test with multiple endpoints, the initial question most people ask is what the cut-off should be between a positive test result and a negative test result. The correct answer, in many cases is to allow multiple cut-offs.

Example: serum ferritin.

In a meta-analysis of studies of diagnosing anemia (Guyatt 1992), serum ferritin was discovered to be the most effective test. Here are the results of this test

Serum ferritin level	Iron Deficient	Not Iron Deficient
>= 100 ug/l	48	1320
45-100 ug/l	76	398
35-45 ug/l	36	43
25-35 ug/l	58	50
15-25 ug/l	117	29
<= 15 ug/l	474	20
Total	809	1860

So what should the cutoff be for this diagnostic test? 15? 25? 35?

To provide some perspective, let's look at the column probabilities.

Serum ferritin level	Iron Deficient	Not Iron Deficient
>= 100 ug/l	5.9%	71.0%
45-100 ug/l	9.4%	21.4%
35-45 ug/l	4.4%	2.3%
25-35 ug/l	7.2%	2.7%
15-25 ug/l	14.5%	1.6%
<= 15 ug/l	58.6%	1.1%
Total	100.0%	100.0%

Normally, I would round these values a bit more, but I am saving an extra decimal place to facilitate some future calculations.

These percentages allow you to compute *Sn* and *Sp* for any possible cut-off. For example, if you define a low serum ferritin level as 35 or less, then

Sn = 58.6 + 14.5 + 7.2 = 80.3%, and
Sp = 71.0 + 21.4 +2.3 = 94.7%.

This process, however, is inefficient, because it treats values of 30, 20, and 10 as if all of them provided the same degree of evidence that the patient has anemia.

A better approach is to treat each discrete category as providing its own level of evidence for or against anemia. You can define a likelihood ratio for each category. This would simply be a ratio of the probability of a specific category given two percentages shown in the table above.

LR^{100+} = 5.9 / 71.0 = 0.08,
$LR^{45\text{-}100}$ = 9.4 / 21.4 = 0.44,
$LR^{35\text{-}45}$ = 4.4 / 2.3 = 1.9,
$LR^{25\text{-}35}$ = 7.2 / 2.7 = 2.7,
$LR^{15\text{-}25}$ = 14.5 / 1.6 = 9.1, and
$LR^{0\text{-}15}$ = 58.6 / 1.1 = 53.3.

This process allows extremely small values of serum ferritin to provide very strong evidence in favor of anemia, extremely large values to provide strong evidence against anemia, and intermediate values to provide smaller but sometimes still important degrees of evidence for or against anemia.

If instead you insisted on a binary classification, you could get a good value for LR^+ at the expense of a mediocre value for LR^- by choosing a small cut-off. Or you could get a good value for LR^- at the expense of a mediocre value for LR^+ by choosing a large cut-off. Or you could choose a middle value for a cut-off and get the worst of both worlds.

Binary choices are simpler, but statisticians generally avoid them when they can because they throw away information about the gray regions.

How do you evaluate the credibility of research studies on diagnostic tests?

There is a lot of controversy about diagnostic testing, and I have mentioned some of these controversies in other weblog entries. I wanted to review what the experts say about diagnostic testing. The definitive resource for evaluating any medical controversy is **Evidence-based Medicine How to Practice and Teach EBM.** David L. Sackett, Scott W. Richardson, William Rosenberg, Brian R. Haynes (1998) Edinburgh: Churchill Livingstone.

There's a newer edition, published in 2005, but I don't think the material I am quoting has changed all that much. The material in Sackett et al that I am quoting was published earlier (Jaeschke 1994a Jaeschke 1994b) and is available on the web (see bibliolgraphy for details.

Suppose you are reviewing a research paper that touts a new diagnostic test. Before you decide whether to use this diagnostic test, you have to assess whether the research findings are valid. You need to ask yourself three questions:

1. Was there an independent, blind comparison with a reference standard?
2. Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?
3. Did the results of the test being evaluated influence the decision to perform the reference standard?

If the research findings are valid, then you have to assess whether the diagnostic test is clinically significant.

If the diagnostic test is valid and clinically significant, you have to assess whether you can extrapolate the results of the study to the particular patient who is in your office right now. You need to ask whether the results in the particular study are applicable to the patients that I normally see.

Finally, you need to know if you have enough information to apply the results in your particular setting. You need to ask yourself three more questions.

1. Is the diagnostic test available, affordable, accurate, and precise in your setting?
2. Can you generate a clinically sensible estimate of your patient's pre-test probability?
3. Will the resulting post-test probabilities affect your management and help your patient?

Let's consider this advice in detail.

Was there an independent, blind comparison? Any research study evaluating a diagnostic test is going to compare it to a more expensive or invasive test that produces a definitive diagnosis of disease. The test that provides a definitive diagnosis is referred to as the "gold standard." Blinding is important in any research study, but it is especially important when there is subjectivity in the interpretation of results. Most diagnostic tests require some level of judgment and if the person applying the diagnostic test is aware of the results of the gold standard or vice versa, that can influence the results. Usually lack of blinding will produce overly optimistic results for the diagnostic test. If the diagnostic test and the gold standard are produced by an automated system with little or no operator intervention and with little or no ambiguity in the reading of results, then blinding is less critical.

Did the study have an appropriate spectrum of patients. Some research designs will include only patients with obvious and overt manifestations of disease. By excluding the milder cases (the shades of gray), the resulting black versus white comparison will result produce overly optimistic results for the diagnostic test. An

appropriate spectrum of patients is also important in insuring that the research results can be extrapolated to your patients (see below).

Did the diagnostic test results influence the decision to perform the reference standard? The gold standard is by definition more expensive or more invasive, so there is a natural reluctance to apply the reference standard. The ideal research study would require every patient to endure both the diagnostic test and the gold standard, but sometimes this is difficult. Suppose the gold standard involves surgery. What do you tell the patients who test negative on the diagnostic test (we suspect that everything is okay, but we want you to submit to this surgery to preserve the credibility of our research findings).

Are the results for the diagnostic test clinically significant? A diagnostic test is clinically significant if knowledge of the results of the diagnostic test can substantially alter your belief about whether your patient has a particular disease. The likelihood ratio will help you answer this question. A likelihood ratio for a positive result smaller than 2 or a likelihood ratio for a negative result larger than 0.5 is pretty much worthless.

Can you extrapolate the results? Medical research is often conducted in an idealized setting that makes the research easier to run but which makes it difficult to generalize the results to your particular patients. Look at the inclusion and exclusion criteria in the study and see if the research population is drawn more narrowly than your patients. Also examine the table of demographics to see if they are comparable to the demographics of your patients (e.g., comparable ages and comparable mixes of race, ethnicity, and gender).

Is the diagnostic test available, affordable, accurate, and precise in your setting? Does the diagnostic test require special skills in its application? Does it require equipment that you do not have? Does the mix of patients that you see raise special issues? For example, do your patients experience developmental problems that make communication difficult?

Can you generate a clinically sensible estimate of your patient's pre-test probability? To apply a diagnostic test, you first need an estimate of the pre-test probability. Do you have records in your practice regarding how often patients who come to you complaining of a particular problem actually have the disease that you are testing for? Are there regional or national surveys that estimate prevalence of the disease? You'd have to adjust this estimate, of course, because the patients who come to see you are more likely to have the disease than the typical probability you'd get by an "on the street" survey. If your patients are similar to the research studies, then the prevalence of disease in that study might be a reasonable estimate. If your patients are dissimilar, but in a way that leads to a predictable increase or decrease in the pre-test probability, make the appropriate adjustment. If you have personal experience through many years of practice, you might be able to provide a "seat of the pants" estimate. Just be sure that your estimate is not colored by your most recent case or your most embarrassing case.

Will the resulting post-test probabilities affect your management and help your patient? A diagnostic test is useless if the likelihood ratio does not shift the probability by a sufficient amount to cause you to cross a treatment threshold. You don't have to do a formal likelihood ratio calculation for every patient that you see, however. Just run a few examples that are typical for a reasonable range of patients (e.g., calculate the results using pre-test probabilities from 45 year old, 65 year old, and 85 year old patients, both smokers and non-smokers).

Appendix. What are odds?

The experts on this issue live just south of Kansas City in a town called Peculiar, Missouri. The sign just outside city limits reads "Welcome to Peculiar, where the odds are with you."

Odds are just an alternative way of expressing the likelihood of an event such as catching the flu. Probability is the expected number of flu patients divided by the total number of patients. Odds would be the expected number of flu patients divided by the expected number of non-flu patients.

During the flu season, you might see ten patients in a day. One would have the flu and the other nine would have something else. So the probability of the flu in your patient pool would be one out of ten. The odds would be one to nine.

There is some ambiguity in how people describe odds using words rather than numbers. Usually, though, it is obvious from the context. For example, the odds of winning a lottery might be a million to one. That either means there are a million people who win the lottery for every one person who loses or there are a million people who lose the lottery for every person that wins.

It's easy to convert a probability into an odds. Simply take the probability and divide it by one minus the probability. Here's a formula.

$$odds = probability / (1 - probability)$$

If you know the odds in favor of an event, the probability is just the odds divided by one plus the odds. Here's a formula.

$$probability = odds / (1 + odds)$$

You should get comfortable with converting probabilities to odds and vice versa. Both are useful depending on the situation.

Here are a few examples of odds calculations. If both of your parents have an Aa genotype, the probability that you will have an AA genotype is .25. The odds are

$$odds = 0.25 / (1 - 0.25) = 0.333$$

which can also be expressed as one to three.

With the same parents, the probability that you will be Aa is .50. The odds are

$$odds = 0.5 / (1 - 0.5) = 1$$

We will sometimes refer to this as even odds or one to one odds.

When the probability of an event is larger than 50%, then the odds will be larger than 1. With the same parents, the probability that you will have at least one A gene is .75. This means that the odds are

$$odds = 0.75 / (1 - 0.75) = 3$$

which we can also express as 3 to 1 in favor of inheriting that gene. Let's convert that odds back into a probability. An odds of 3 would imply that

$$probability = 3 / (1 + 3) = 0.75$$

Well that's a relief. If we didn't get the same answer back, that would leave me open to all sorts of lawsuits.

Suppose the odds against winning a contest were ten to one. We need to re-express as odds in favor of the event, and then apply the formula. The odds in favor would be one to ten or 0.1. Then we would compute the probability as

$$probability = 0.1 / (1 + 0.1) = 0.09$$

Notice that in this example, the probability (0.09) and the odds (0.1) did not differ too much. This pattern tends to hold for rare events. On the other hand, when the probability is large, the odds will be quite different from the probability.

Bibliography

Sensitivity and specificity of QTc dispersion for identification of risk of cardiac death in patients with peripheral vascular disease. Darbar D, Luck J, Davidson N, Pringle T, Main G, McNeill G, Struthers AD. British Medical Journal 1996; 312(7035): 874-8. The full free text of this article is available at bmj.com/cgi/content/full/312/7035/874.

Nomogram for Bayes theorem. Fagan T. New England Journal of Medicine 1975; 293: 257.

Laboratory diagnosis of iron-deficiency anemia: an overview. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W and Patterson C. J Gen Intern Med 1992; 7(2): 145-53.

Panel to Advise Testing Babies for 29 Diseases. Kolata G. The New York Times, February 21, 2005. The full text of this article is available at www.nytimes.com/2005/02/21/health/21baby.html.

Annual Physical Checkup May Be an Empty Ritual. Kolata G. The New York Times, August 12, 2003. The full text of this article is available at query.nytimes.com/gst/fullpage.html?res=9505E5D81031F931A2575BC0A9659C8B63.

Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jaeschke R, Guyatt G, Sackett DL. Jama 1994: 271(5); 389-91. The full text of this article is available at www.cche.net/usersguides/diagnosis.asp.

Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? Jaeschke R, Guyatt G, Sackett DL. Jama 1994: 271(5); 389-91. The full text of this article is available at www.cche.net/usersguides/diagnosis.asp.

Accuracy of screening for gastric cancer using serum pepsinogen concentrations. Kitahara F, Kobayashi K, Sato T, Kojima Y, Araki T, Fujino MA. Gut 1999; 44(5): 693-97.

Evidence-based Medicine How to Practice and Teach EBM. David L. Sackett, Scott W. Richardson, William Rosenberg, Brian R. Haynes (1998) Edinburgh: Churchill Livingstone.

Accuracy of a single question in screening for depression in a cohort of patients after stroke: comparative study. Watkins C, Daniels L, Jack C, Dickinson H, van den Broek M. BMJ 2001; 323(7322): 1159. The full text of this article is available at bmj.com/cgi/content/full/323/7322/1159

Rectal bleeding and colorectal cancer in general practice: diagnostic study. Wauters H, Van Casteren V, Buntinx F. British Medical Journal 2000; 321(7267): 998-9. The full text of this article is available at bmj.com/cgi/content/full/321/7267/998.

Additional resources including a link to a PDF version of this handout can be found at www.childrensmercy.org/stats/diagnostic.asp.