Interpreting Diagnostic Tests Using Probability

Ralph G. O’Brien, PhD
Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation

Introduction

These notes detail my 40-minute lecture given 11 June 1995 as part of the Seventh Annual Intensive Review of Internal Medicine, a symposium conducted by the Cleveland Clinic. I was asked to review how probability is used with diagnostic testing. Our main question is:

How can diagnostic results and other information about a patient be used to estimate the chance that he/she has a particular “disease?”

First, however, we must address:

In general, how well does the test perform as a diagnostic tool for this disease?

Use probabilities to state diagnostic uncertainty.

Everyone thinks about probabilities when making everyday decisions. Usually, this is done roughly: “There’s a good chance of rain today, so I’ll ...”. But when we are required to think probabilistically in our professional work, we must do so using the methods and information that science gives us.

Figure 1 depicts the flow of information and decision-making in the management of the solitary pulmonary nodule (SPN). Please take a minute to study it. My point here is that the entire process is driven by the need to estimate the probability that the SPN is malignant. As decision makers, both clinicians and patients want such estimates to be “very low” or “high” in order to avoid The Tough Call. What probability values are actually “very low” or “high” will vary from case to case. Some patients might opt for resection if the chance of malignancy exceeds 10%; others will take this action only if the chance is considerably higher. Regardless, our goal should be to estimate this probability as best we can.

Figure 1.

Estimating the probability of a malignancy is a key component in managing the solitary pulmonary nodule.
Probability is rooted in mathematics and therefore has its own sub-language, just as all scientific fields do, including medicine. I really need to use some technical notation or else I will be even harder to follow. Let D stand for “disease,” a term we take as quite generic. A patient who truly has the disease (whether we know this or not) is a D+ patient; one who truly does not is D−. Before we have specific information about the patient, such as diagnostic tests, we would estimate that their chance of disease is

Pr[D+], the prevalence of the disease in the population of interest.

Let T stand for “test,” which like “disease” is a generic term; in fact, “test” could represent some risk factor, such as smoking status. A positive test is T+ (“abnormal”), which is positively associated with D+. A negative test is T− (“normal”). Of course, many tests are not simply T+ or T−, but instead have many possible values. We will use the more general term T* to denote any such value. (The simplest case has T+ and T− as the only two values for T∗.)

If we get a T+ result on a patient, we would revise Pr[D+] to become

Pr[D+ | T+], the probability that disease is present given that the test is positive.

Later, we will learn that this is called the positive predictive value of the test. If the test has any value at all, P[D+ | T+] > P[D]. The negative predictive value of the test is

Pr[D− | T−], the probability that disease is absent given that the test is negative.

In general, we will have the result T*, and we are interested in

Pr[D+ | T*], the probability that disease is present given that the test result is T∗.

**SPN example, part 1.** Consider an example based on diagnosing malignancy in solid pulmonary nodules. “Jack” is a 45-year-old male who has never smoked. For such a person, assume that the prevalence of a lung malignancy is about Pr[D+] = 0.2%, i.e. 1 in 500. Jack gets a nagging cough and is given a standard chest x-ray. This shows “something” abnormal; call this first test result T1*. It seems unrelated to the cough, but could this be a malignancy? Probably not: Pr[D+ | T1*] = 1.0%. But the concern is enough that Jack undergoes a CT scan. The images show a single pulmonary nodule, 5 cm in diameter (T2*), with no occult calcification (T3*). These results replace the standard chest x-ray result. Now, Pr[D+ | T2*, T3*] = 13.0%. These values were motivated by the work of Lillington and colleagues[1][2][3] and will be revisited later.

What should be the next step in this diagnosis? A definitive benign biopsy result (T4−) would cut the chance of malignancy to Pr[D+ | T2*, T3*, T4−] = 0.5%, with the “T2*, T3*, T4−” notation signifying that the CT information would still be relevant because such biopsy results could be wrong. A definitive positive biopsy would make the CT information mostly irrelevant, so we can rely on the biopsy’s positive predictive value: Pr[D+ | T4+] = 98%[4]. Of course the biopsy results could be indeterminate.

**Question 1.** Let’s see if you can use this notation and reason some logic about probabilities. First, let me say that “hyperutolatemia” and the “CCF950611 gene” are fictitious. Now consider the following question:

In the general population, 6% suffer from hyperutolatemia (D+). But this rate is 57% among those carrying the CCF950611 gene. The prevalence of CCF950611 has not been determined.

Which of the following statements are false?

(a) Pr[D+] = 6%; Pr[D+ | CCF950611+] = 57%
(b) Pr[D−] = 94%; Pr[D− | CCF950611+] = 43%
(c) Pr[CCF950611+ | D+] > Pr[CCF950611+]
Answer (d) if all statements are true.
Study the answer now, given at the end of these notes.

How well does the test perform?
Before using a test in clinical practice, one must understand its strengths and weaknesses. How well does it distinguish between people with and without the disease? This section covers some classic definitions and concepts about overall test performance. *The terminology is confusing.* (I did not create it. It is now too universally used to try to change it.)

**A hypothetical diagnostic test.** We need to work with some actual numbers, so I created the hypothetical example displayed in Figure 2. Pretend that it depicts the distribution of results for all 702 people known to be D+ (“cases”) and all 441 known to be D− (“controls”). Each circle is one person’s score on the value of this test, V, a continuous measure. A real continuous diagnostic measure familiar to you is plasma glucose level (mg/dl) in the oral glucose tolerance test for diabetes.

![Figure 2. Hypothetical example of diagnostic test with continuous measure.](image)

The upper distribution shows only D− people. Taking people with values exceeding 7.5 to be “abnormal” (T+), we have #D−T− = 419 *true negatives* and #D−T+ = 22 *false positives*. Likewise, there are #D+T− = 83 *false negatives* and #D+T+ = 619 *true positives*.

**Sensitivity and specificity.** One common way to summarize how well this test discriminates between D+ and D− people is to compute how well the test does for the two groups. The accuracy of the test for D+ people is

\[ \text{sensitivity: } \Pr[T^+ | D^+] = \frac{\#D^+T^+}{\#D^+} = \frac{619}{702} = .88. \]

The accuracy of the test for D− people is

\[ \text{specificity: } \Pr[T^- | D^-] = \frac{\#D^-T^-}{\#D^-} = \frac{419}{441} = .95. \]

These terms are confusingly similar. I distinguish them by thinking that we want tests to be *sensitive* in discovering D+ people, but *specific* enough to not misdiagnose D− people.
**Question 2.** This question was motivated by an ordeal suffered by a CCF staff physician I know who trained at UC San Francisco around 1980 and was no doubt in contact with HIV+ blood in the normal course of his work. When he later applied for life insurance, some “non-normal” blood values and his UCSF history caused some problems. Please answer this question:

In the early days of the AIDS epidemic, diagnostic tests for HIV infection lacked adequate ______________, causing needless anguish and hardship for those who were falsely diagnosed as being infected.

(a) sensitivity  
(b) specificity  
(c) statistical significance  
(d) likelihood ratios

**Choosing the cut-point: trading specificity for sensitivity.** Figure 3 and Table 1 show that if we lower the cut-point to increase the test’s sensitivity, then this in turn reduces its specificity.

**Figure 3.**
Lowering the cut-point raises sensitivity but reduces specificity.

**Table 1.**
Sensitivities and specificities for the four cut-points illustrated in Figure 3.

<table>
<thead>
<tr>
<th>“Abnormal”</th>
<th>#D^T+</th>
<th>Sensitivity</th>
<th>#D^-T^-</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>V &gt; 6.5</td>
<td>689</td>
<td>.98</td>
<td>362</td>
<td>.82</td>
</tr>
<tr>
<td>V &gt; 7.0</td>
<td>669</td>
<td>.95</td>
<td>394</td>
<td>.89</td>
</tr>
<tr>
<td>V &gt; 7.5</td>
<td>619</td>
<td>.88</td>
<td>419</td>
<td>.95</td>
</tr>
<tr>
<td>V &gt; 8.0</td>
<td>526</td>
<td>.75</td>
<td>433</td>
<td>.98</td>
</tr>
</tbody>
</table>

#D^+ = 702  
#D^- = 441
What cut-point should one use? This depends on many factors, mostly concerning the costs and benefits (not just financial, of course) that are associated with making either of the two correct diagnoses (D+T+ and D–T–) or the two incorrect ones (D–T+ and D+T–). This is a most complex and subjective matter, well beyond the scope of this lecture.

**ROC curves.** With so many potential cut-points, how do we judge a given test’s sensitivity and specificity? How do we summarize these values, both absolutely and relative to competing tests? The ROC\(^1\) curve plots the false positive rate (= 1 - specificity) versus the sensitivity for every relevant cut-point. Figure 4 shows the ROC curve derived from our hypothetical data, including the four cut-points we just discussed. An error-free diagnostic test would have no overlap whatsoever for the D+ and D– distributions, thus giving a false positive rate of 0.00 and a sensitivity of 1.0 for every cut-point, i.e. all points would lie at the very top of the graph. The *area under this curve* (AUC) would be 1.0. AUC is the most common statistic for summarizing the overall sensitivity and specificity. For our data, the AUC is 0.98. Also shown is the ROC curve of an inferior test that nevertheless is still predictive. AUC is often a primary outcome measure in clinical studies that compare competing diagnostic tests. Finally, the figure shows a “chance” line, which would indicate that the D+ and D– distributions were identical, so that using any cut-point would be no better than just guessing.

**Figure 4.**
ROC curves for the hypothetical data, for an inferior diagnostic test, and for simple chance.

**Likelihood ratio of test result.** The likelihood ratio of a particular test outcome is\(^2\)

\[
LR(T^*) = \frac{\Pr[T^* \mid D^+]}{\Pr[T^* \mid D^-]}.
\]

---

\(^1\)ROC: “Receiver Operator Characteristic,” a term coined in electronics.

\(^2\)Technically, for a truly continuous variable, the probability of any particular outcome is infinitesimal, so we have to use the ratio of the “likelihoods,” a topic beyond this lecture. This is the basis of the term “likelihood ratio.”
Taking $T^* = T^+$, and using our hypothetical data with a cut-point of $V > 7.5$, we get

$$LR(T^+) = \frac{Pr[T^+ | D^+]}{Pr[T^+ | D^-]} = \frac{\text{sensitivity}}{\text{false positive rate}} = \frac{619/702}{227/441} = 17.7.$$ 

The $LR(T^+)$ values for all four cut-points are given in Table 1.

**Table 2.**
Likelihood ratios for the four cut-points illustrated in Figure 3.

<table>
<thead>
<tr>
<th>“Abnormal”</th>
<th>Sensitivity</th>
<th>False Positive Rate</th>
<th>LR(T^+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V &gt; 6.5$</td>
<td>.98</td>
<td>.18</td>
<td>5.5</td>
</tr>
<tr>
<td>$V &gt; 7.0$</td>
<td>.95</td>
<td>.11</td>
<td>8.9</td>
</tr>
<tr>
<td>$V &gt; 7.5$</td>
<td>.88</td>
<td>.05</td>
<td>17.7</td>
</tr>
<tr>
<td>$V &gt; 8.0$</td>
<td>.75</td>
<td>.02</td>
<td>41.3</td>
</tr>
</tbody>
</table>

Aside: Another general way to define LR is to make it the slope of ROC curve where the cut-point corresponds to $V > T^*$. To see this, note that the first part of the curve is quite straight (linear) and has a slope of approximately

$$LR(T^*) = \frac{\text{sensitivity}}{\text{false positive rate}}$$

at any point in this range. Estimating and testing ROC curves and LRs is a specialized area of biostatistics.

**Positive/negative predictive values.** A diagnostic test’s utility is also described by its predictive values. The accuracy of the test for $T^+$ people is the *positive predictive value*,

$$PV(T^+) = \text{Pr}[D^+ | T^+] = \frac{#D^+T^+}{#T^+} = 619/641 = .97.$$ 

The *negative predictive value* is

$$PV(T^-) = \text{Pr}[D^- | T^-] = \frac{#D^-T^-}{#T^-} = 419/502 = .83.$$ 

These values are only applicable to this particular study because they do not use the actual prevalence of $D^+$, but instead use the one in the study sample. As a result, the .97 and .83 values could be quite misleading when applied clinically—unless the prevalence of disease in the study matches the prevalence in the clinical population of interest. Clinical studies typically study $D^+$ people in much higher proportions than they exist in the clinic population at large. In our sample, $\text{Pr}[D^+] = 702/1143 = .61$. If in the clinic setting $\text{Pr}[D^+]$ was much less than .61, say .04 instead, then the “clinical” $PV(T^+)$ would be much less than the “study” value of .97. Likewise, the clinical $PV(T^-)$ would be greater than .83. Later, we will see how to use LRs to easily compute $PV(T^*)$ with a clinically relevant $\text{Pr}[D^+]$ value.

**Estimating whether a particular patient is $D^+$**

Understanding and using probabilities and likelihood ratios can help physicians and patients make critical decisions. Of course we must have access to prevalence and LR values that are published in sound places. But when we do, using them in clinical practice only requires the use of some
simple formulas and/or a convenient graphical device. Stay with me: Nothing here is mathematically complex or arduous.

**SPN example, part 2.** In community surveys of males with a solitary pulmonary nodule, it was estimated that about \( P[D^+] = 9\% \) would be malignant.\(^1\) Lillington\(^3\) supplied a set of LR\(\text{s} \) that are useful in revising this rate up or down depending on the values of various risk factors and the outcomes of CT imaging. Here are some of them:

- **Diameter of nodule (cm) via CT:** \( \text{LR(<1.5)=0.1; LR(1.5-2.2)=0.5; LR(2.3-3.2)=1.7; LR(3.3-4.2)=4.3; LR(4.3-5.2)=6.6; LR(5.3-6.0)=29.4.} \)
- **Type of lesion edge via CT:** \( \text{LR(I)=0.2; LR(II)=0.5; LR(unknown)=1.0; LR(III)=5.0; LR(IV)=14.0.} \)
- **Occult calcification via CT:** \( \text{LR(yes)=0.02; LR(unknown)=1.0; LR(no)=2.15.} \)
- **Age:** \( \text{LR(\leq35)=0.1; ... LR(45-49)=0.7; ... LR(70-83)=5.7.} \)
- **Smoking:** \( \text{LR(never)=0.15; ... LR(>41 cigs/day)=3.9.} \)

We see that the risk increases with nodule size and type of edge, and in the absence of calcification. Older patients and smokers are also at increased risk. We will discuss how to use these numbers soon.

**Odds and probability.** Another common way to quantify chance is through odds, which is directly related to probability,

\[
\text{Odds}[\bullet] = \frac{\text{Pr}[\bullet]}{1 - \text{Pr}[\bullet]},
\]

where \( \text{Pr}[\bullet] \) means probability of “whatever.” Going back the other way is also simple:

\[
\text{Pr}[\bullet] = \frac{\text{Odds}[\bullet]}{1 + \text{Odds}[\bullet]}.
\]

If \( \text{Pr}[\bullet] = .20 \), then \( \text{Odds}[\bullet] = .25. \) “Even odds” is \( \text{Odds}[\bullet] = 1.00 \), which is \( \text{Pr}[\bullet] = .50. \)

**Updating odds and probabilities.** We need to cover Bayes’s Rule. It won’t hurt. In essence, all it does is use new information (\( T^* \)) to update “old odds.” Here is the formula:

\[
\text{Updated Odds}[D^+] = \text{Old Odds}[D^+] \times \text{LR}(T^*).
\]

Of course, one can convert such updated odds into updated probabilities. The proper terms for “old” odds and probability are *prior odds* and *prior probability*. Likewise, *posterior odds* or *posterior probability* are the proper terms for “updated” odds or probability.

**SPN example, part 3.** Now we can interpret and use the LR values given by Lillington. Discovering that the nodule is 5 cm increases the odds of malignancy by a factor LR(4.3-5.2) = 6.6. If we use the prevalence for 45-year-old non-smoking males, \( \text{Pr}[D^+] = .002 \), to set the prior odds, we get

\[
\text{Odds}[\text{Malignancy} | 5 \text{ cm}] = \frac{.002}{.998} \times 6.6 = .0132.
\]

which converts to

\[
\text{Pr}[\text{Malignancy} | 5 \text{ cm}] = \frac{.0132}{1.0132} = .013.\]
On the other hand, if the CT scanning shows calcification \([\text{LR(yes) } = 0.02]\), then the odds of malignancy are cut by a factor of 50. Thus if our 5 cm nodule were to show calcification, then the new odds would be \(0.0132 \times 0.02 = 0.00026\).^{3} Thus, \(\Pr[\text{malignancy} \mid 5 \text{ cm, no calcification}] < 0.00026\), which is less than the prevalence rate we began with.

**Multiple predictors.** This last example motivates the general form of Bayes’s Rule when there are several \((m)\) predictors:

\[
\text{Odds} [D^+ \mid T_1^*, T_2^*, \ldots, T_m^*] = \text{Odds} [D^+] \times \text{LR}(T_1^*) \times \text{LR}(T_2^*) \times \ldots \times \text{LR}(T_m^*).
\]

**SPN example, part 4.** Let us discuss Jack again, the 45-year-old who has never been a smoker. A potential SPN is discovered via a standard chest x-ray performed for reasons wholly unrelated to cancer. Subsequent CT imaging shows a nodule 5 cm in diameter, with no calcification. The attending radiologist was unaware of Lillington’s system for classifying the lesion’s edge, so this is unknown. The LR\$s for these values are given above. With \(P[D^+] = 0.09\) as the prevalence rate, we have \(\text{Odds}[D^+] = 0.1\). These give:

\[
\begin{align*}
\text{Odds}[\text{malignancy} \mid \text{male/45/never} & \& \text{5/no calc./unknown}] \\
= 0.1 \times (0.7 \times 0.15) \times (6.6 \times 2.15 \times 1.0) \\
= 0.1 \times (0.105) \times (14.2) \\
= 0.1 \times 1.49 \\
= 0.149.
\end{align*}
\]

Even though Joe’s age and never-smoked status reduce the risk of cancer \([\text{LR(age + smoking) } = 0.105]\), the CT results are troublesome \([\text{LR(CT) } = 14.2]\). With this information, the probability is \(0.149/1.149 = .13\) that this SPN is malignant. This probably makes Jack’s case a Tough Call at this point.

**Fagan’s nomogram**

Figure 5 is my redrawing of a marvelous nomogram^{4} developed by Fagan.\[5\] Just line up the prior probability with the LR of interest and you get the posterior probability. The two lines I placed on there show how to get the posterior probabilities for (a) the example just completed: prior probability of .09 and overall LR of 1.49; and (b) the same example albeit changing the prior odds to Lillington’s “clinical setting” value of 0.7, a prior probability of .7/1.7 = .41. You could even do this iteratively to handle multiple LRs, but it is probably easier to just do the multiplications and conversions on a simple calculator.

**Question 3.** Use the nomogram to solve this problem.

Based on history-and-physical information, your (subjective) prior probability was low, say .03, that your patient, “Jill,” has congenital hyperutolatemia. However, a genetic screening now reveals that Jill carries the CCF950611 gene. This has a likelihood ratio of 150 in predicting congenital hyperutolatemia. What is your revised estimate of the chance that Jill is so afflicted?

---

^{3}Because of the way Lillington assembled his LRs, the validity of this calculation depends on assuming that nodule size and calcification status are conditionally independent predictors. This point is addressed in [1], but I find their discussion somewhat wanting.

^{4}Definition from *American Heritage Dictionary*: A graph consisting of three coplanar curves, each graduated for a different variable so that a straight line cutting all three curves intersects the related values of each variable.
Predictive value with clinic-based prevalence values. With Bayes’s Rule (and Fagan’s nomogram) we can easily get valid clinic-based estimates of predictive values. To see this for
positive predictive value, let us return to the hypothetical data and recall that $LR(V>7.5) = 17.7$. Suppose that the clinic-based prevalence is 0.04, which translates to odds of $0.04/0.96 = 0.042$. Therefore, the posterior odds are $0.042 \times 17.7 = .7375$, which translates back to a probability of $PV(V>7.5) = 0.42$. This is much different from the sample-based value of 0.97. Try other base rates using the nomogram and you will easily see how much predictive values are influenced by one’s choice of prevalence value. Prevalence values are just one kind of initial prior probability—they should all be subjected with some scrutiny.

**What is left out**

With but 40 minutes to have your attention, I had to leave a lot of things out. The most important matter is the whole topic of doing good studies to properly estimate the various types of statistical values discussed here. Good clinical research has good internal validity: We want the answers we get to be highly likely to be correct, and our estimates to be “on target” even though they are not “bull’s eyes” when applied in clinical situations similar to those studied in the research project. Good research also has defined generalizability, that is, we know how widely we can apply our new knowledge. Studies with high external validity can be applied widely. Is a diagnostic imaging “reliability” study with outstanding internal validity but conducted solely within the Cleveland Clinic of any value to radiology departments elsewhere? If we decide to do a study with several dissimilar departments world-wide, will this compromise internal validity?

How good are our estimates for ROC curves, prevalence values, and likelihood ratios? Are they from studies with high internal and external validity? How do we decide which study patients are really $D^+$ and $D^-$ in such studies? Who defines the “gold standard?” Was the study sample large enough and the sampling scheme well designed and carried out? How much statistical expertise, care, and collaborative energy went into the data analyses?

This morass of issues is beyond the scope of this lecture. To start on your own, I recommend Harris’s provocative paper,[6] “The hazards of bedside Bayes.”

**Answers to questions**

1. All statements are true. (a) just expresses the information in our notation. (b) expresses the same information, albeit in terms of $T^-$ instead of $T^+$. (c) is true because there is some positive association between hyperutolatemia and the CCF950611 gene.
2. “Likelihood ratio” is marginally correct, but “specificity” fits perfectly. (By the way, my friend eventually did test negative for HIV and now has plenty of life insurance.)
3. I got something close to 0.80, maybe a little above.

**REFERENCES**