IN THIS TUTORIAL WE will examine how to evaluate a diagnostic test. Initially we will consider the case when there is a binary measure (two categories: disease present / disease absent). We will then look at how to define a suitable cut-off for an ordinal or continuous measurement scale and we will finish with a short discussion contrasting diagnostic tests with screening tests.

When evaluating any diagnostic test one should have a definitive method for deciding whether the disease is present in order to see how well the test performs. For example, to diagnose a cancer one could take a biopsy, to diagnose depression one could ask a psychiatrist to interview a patient, and to diagnose a walking problem one could video a patient and have it viewed by an expert. This is sometimes called the ‘gold standard’. Often the gold standard test is expensive and difficult to administer and thus a test is required that is cheaper and easier to use.

**BINARY SITUATION**

Let us consider first the simple binary situation in which both the gold standard and the diagnostic test have either a positive or negative outcome (disease is present or absent). The situation is best summarised by a 2 x 2 table (table 1). In writing this table, always put the gold standard on the top and the results of the test on the side.

The numbers ‘a’ and ‘d’ are the numbers of true positives and true negatives, respectively. The number ‘b’ is the number of false positives, because although the test is positive the patients don’t have the disease, and similarly ‘c’ is the number of false negatives. The prevalence of the disease is the proportion of people diagnosed by the gold standard and is given by \( a + c \) / \( n \), although this is often expressed as a percentage.

In order to assess how good the test is we can calculate the sensitivity and specificity, and the positive and negative predictive values. The sensitivity of the test is the proportion of people with the disease who are correctly identified as having the disease. This is given by \( a / (a + c) \) and is usually presented as a percentage. Suppose a test is 100 per cent sensitive. Then the number of false negatives is zero and we would expect table 2.

From table 2 we can see that if a patient has a negative test result we can be certain that the patient does not have the disease. Sackett et al.1 refer to this as SnNout, i.e. for a test with a high sensitivity (Sn), a Negative result rules out the disease.

The specificity of a test is the proportion of people without the disease who are correctly identified as not having the disease. This is given by \( d / (b + d) \) and as with sensitivity is usually presented as a percentage. Now suppose a test is 100 per cent specific. Then the number of false positives is zero and we would expect table 3.

From table 3 we can see that if a patient has a positive test we can be certain the patient has the disease. Sackett et al. refer to this as SpPin, i.e. for a test with a high specificity (Sp), a Positive test rules in the disease.

**USEFUL MNEMONIC**

<table>
<thead>
<tr>
<th>Sensitivity = 1 – proportion false negatives</th>
<th>Positives [p in each side]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity = 1 – proportion false positives</td>
<td>Negatives [n in each side]</td>
</tr>
</tbody>
</table>

What patients really want to know is ‘if I have a positive test, what are the chances I have the disease?’

**What patients really want to know is ‘if I have a positive test, what are the chances I have the disease?’**

It should be noted that whilst sensitivity and specificity are independent of prevalence, positive and negative predictive values are not. Sensitivity and specificity are characteristics of the test and will be valid for different populations with different prevalences. Thus we could use them in populations with high prevalence such as elderly people as well as for low prevalence such as for young people. However, the PPV is a characteristic of the population and so will vary depending on the prevalence.

To show this, suppose that in a different population the prevalence of the disease is double that of the current population (assume the prevalence is low, so that \( a \) and \( c \) are much smaller than \( b \) and \( d \) and thus the results for those without the disease are much the same as the earlier table). The situation is given in table 4.

The sensitivity is now \( 2a / (2a + 2c) \) = \( a / (a + c) \) as before. The specificity is unchanged. However the positive predictive value is given by \( 2a / (2a + b) \) which is greater than the earlier value of \( a / (a + b) \).

**LIKELIHOOD RATIO**

It is common to prefer a single summary measure, and for a diagnostic test this is given by the likelihood ratio for a positive test (LR+) as defined below:

\[
LR^+ = \frac{P(+) \mid D^+}{P(+) \mid D^-} = \frac{P(+) \mid D^+}{P(-) \mid D^-}
\]

Probability of positive test given the disease = Probability of positive test without disease

\[
\text{Sensitivity} = \frac{a}{a + c}
\]

\[
1 - \text{Specificity} = \frac{b}{a + c}
\]

One reason why this is useful is that it can be used to calculate the odds of having the disease given a positive result. The odds of an event are defined as the ratio of the probability of the
event occurring to the probability of the event not occurring, i.e. $p/(1-p)$ where $p$ is the probability of the event. Before the test is conducted the probability of having the disease is just the prevalence, and the odds are simply $\frac{(a + c)/n}{(b + d)/n} = \frac{a + c}{b + d}$. The odds of having the disease after a positive test are given by

$$\frac{odds \text{ after positive test}}{odds \text{ before test}} = \frac{a}{c} \cdot \frac{a+b}{b+d} = \frac{a}{c} \cdot \frac{1}{LR(-)}$$

We can also get the odds of disease after a positive test directly from the PPV since the odds of disease after a positive test is $PPV/(1−PPV)$.

**EXAMPLE**

A recent study by Kroenke et al. surveyed 965 people attending primary care centres in the US. They were interested in whether a family practitioner could diagnose Generalised Anxiety Disorder (GAD) by asking two simple questions (the GAD2 questionnaire): ‘Over the last two weeks, how often have you been bothered by the following problems? (1) Feeling nervous, anxious or on edge; (2) not able to stop or control worrying’. The patients answered each question from ‘not at all’, ‘several days’, ‘more than half’ and ‘nearly every day’, scoring 0, 1, 2 or 3, respectively. The scores for the two questions were summed and a score of over 3 was considered positive. Two mental health professionals then held structured psychiatric interviews with the subject over the telephone to diagnose GAD. The professionals were ignorant of the result of the GAD2 questionnaire. The results are given in table 5.

![Table 1](image1.png)

**TABLE 1. Standard table for diagnostic tests.**

![Table 2](image2.png)

**TABLE 2. Results of a diagnostic test with 100 per cent sensitivity.**

![Table 3](image3.png)

**TABLE 3. Results of a diagnostic test with 100 per cent specificity.**

![Table 4](image4.png)

**TABLE 4. Standard situation but with a doubling of the prevalence.**

The prevalence of the disease is given by $(a + c)/n = 73/965 = 0.076 = 7.6$ per cent.

The sensitivity of the test is given by $a/(a + c) = 63/73 = 0.86 = 86$ per cent.

The specificity of the test is given by $d/(b + d) = 740/892 = 0.83 = 83$ per cent.

The positive predictive value (PPV) is $a/(a + b) = 63/215 = 0.29 = 29$ per cent.

The negative predictive value is $d/(c + d) = 740/750 = 0.987 = 98.7$ per cent.

Thus before the test the chances of having GAD were $7.6$ per cent. After the test they are either $29$ per cent or $1.3$ per cent (i.e. $100 × (1−0.987)$) depending on the result. Note that even with a positive test the chances of having GAD are still less than $1/3$.

For the GAD example we find that $LR(+) = 0.86/(1−0.83) = 5.06$ and the odds is $0.29/(1−0.29) = 0.41$.

**ROC CURVES**

For a diagnostic test that produces results on a continuous or ordinal measurement scale, a convenient cut-off level needs to be selected to calculate the sensitivity and specificity. For example the GAD2 questionnaire has possible values from 0 to 6. Why should one choose the value of 3 as the cut-off? For a cut-off of 2 the sensitivity is 0.95, the specificity is 0.64 and the $LR(+) = 2.6$. One might argue that since a cut-off of 3 has a better $LR(+) = 7.6$ then one should use it. However, a cut-off of 2 gives a higher sensitivity, which might be important. It should be noted that a sensitivity of 100 per cent is always achievable by stating that everyone has the disease, but this is at the expense of a poor specificity (similarly a 100 per cent specificity can be achieved by stating no-one has the disease. If the prevalence is low, this tactic will have a high accuracy, i.e. it will be right most of the time, but sadly wrong for the important cases). A discussion of the different scenarios for preferring a high specificity or sensitivity is given in the next section.

A simple graphical device for displaying the trade-offs between sensitivity and specificity for tests on a continuous or ordinal scale is a receiver operating characteristics (ROC) curve (the unusual name originates from electrical engineering). This is a plot of sensitivity versus one minus specificity for different cut-offs of values of the diagnostic test. ROC curves for two theoretical tests are shown in figure 1, together with the line of equality which is what we would expect if a test had no power to detect disease. A perfect diagnostic test would be one with no false negatives (i.e. sensitivity of 1) or false positives (i.e. specificity of 1) and would be represented by a line starting at the origin, travelling vertically up the Y-axis to a sensitivity of 1 and then horizontally across to a false positive rate of 1. Any diagnostic test that was reasonable would produce a ROC curve in the upper left-hand triangle of figure 1. The selection of the optimal cut-off will depend upon the relative medical consequences and costs of false positive and false negative errors. ▶

**TABLE 4. Standard situation but with a doubling of the prevalence.**
ROC curves are particularly useful for comparing different diagnostic tests and when more than one test is available they can be compared by plotting both on the same plot. A test for which the plot is consistently nearer the left-hand side and the top is to be preferred. In addition the area under the curve (AUC) for each plot can be calculated. For the perfect test outlined above the AUC is 1 and represents the total area of the panel (i.e. $1 \times 1$). For the two curves displayed it is obvious that the best test is the one with the line represented by the dashed line on the left of the figure. This has an AUC value of 0.95 compared to the other much poorer fitting line which as an AUC value of 0.59.

**DISTINCTION BETWEEN DIAGNOSIS AND SCREENING**

It is important to understand the difference between diagnosing a disease and screening for it. In the former case there are usually some symptoms, and so there may already be a suspicion that something is wrong. If a test is positive some action will be taken. In the latter case there are usually no symptoms and so if the test is negative the person will have no further tests. Recalling Sackett’s mnemonics SpPin and SnNout, for diagnosis we want a positive test to rule people in, so we want a high specificity. For screening we want a negative test to rule people out so we want a high sensitivity. Thus mass mammography will have a fairly low threshold of suspicion, to ensure a high sensitivity and reduce the chances of missing someone with breast cancer. The subsequent biopsy of positive results will have a high specificity to ensure that if, say, mastectomy is to be considered, the doctor is almost certain that the patient has breast cancer.

**SUMMARY**

This tutorial has summarised the methods used for examining the suitability of a particular test for diagnosing disease. In addition it has highlighted the difference between diagnostic and screening tests. In reality the same methods are used to evaluate both diagnostic and screening tests, the important difference being the emphasis that is placed on the sensitivity and specificity. Further details are given in Chapter 4 of Campbell et al.3

**FIGURE 1.** Example ROC curves showing also the line of equality.

**TABLE 5.** Results from Kroenke et al.

<table>
<thead>
<tr>
<th>GAD2</th>
<th>Diagnosis by mental health worker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>≥3</td>
<td>63</td>
</tr>
<tr>
<td>&lt;3</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
</tr>
</tbody>
</table>

**REFERENCES**