10

How Do I Know When a Diagnostic Test Works?

MARTIN BUNNAGE

Diagnostic validity is a crucial component of effective practice in clinical neuropsychology. Healthcare practitioners are in the privileged position of being able to help members of society with some of their health concerns. To be able to do this as efficiently and effectively as possible, it is necessary to have valid procedures by which to make decisions. Decision-making in healthcare is often imperfect, but some approaches are more overtly scientific than others (Ruscio, 2007; Straus et al., 2011).

In the pursuit of evidence-based practice, it is appropriate to acknowledge the weaknesses and biases human decision-makers display and instead combine clinical skills and observations with assessment tools and decision-making methods that maximize expertise and diagnostic validity (Haynes et al., 2002). Diagnostic validity is a special application of criterion-related validity, which has long been discussed within applied psychology as the most direct method by which to validate tests (e.g., see Strauss and Smith, 2009; Faust, 2003).

Within the practice of clinical neuropsychology, diagnostic decision-making occurs at many levels. Sometimes the focus is on diagnosis when it relates to deciding upon the likely presence or absence of a disease. At other times, the focus is on making prognostic statements that rely on the diagnosis of specific signs or symptoms. In each scenario, efforts are made to bring some order to the wealth of information available during a clinical encounter. Clinicians formulate meaningful questions and then try to answer them in a valid way that allows for inferences of interest to be made that benefit the patient (Schoenberg & Scott, 2011). The questions may be very varied, for example, whether the person has suffered from a traumatic brain injury, whether the person has a dyspraxia or a memory disorder, which may then be used to infer damage within the cerebrum, or whether from a neuropsychological perspective the person being assessed is safe to return to their former work role as a skilled professional.

When trying to understand the scientific basis of diagnostic decision-making, practitioners are often discomfited by the apparent mechanistic and mathematical presentation of the relevant principles, even though research suggests that careful, objective techniques are likely to increase judgement accuracy (Grove & Meehl, 1996). In the current chapter, mathematical formulae have been kept to an absolute minimum, and wherever possible, ideas and principles are illustrated using scenarios and tests common to clinical neuropsychologists.

Consider the following example. A clinician has a test of memory and wants to know whether or not someone has a memory problem sufficient to meet the suggested diagnostic criteria of mild cognitive impairment (MCI; see Albert et al., 2011). The diagnostic question being asked is whether or not someone has a memory problem. The presence or absence of a memory problem is something that is not known in direct or absolute terms, if it were there would be no need to test for it. Instead, the clinician has a conceptual understanding of what memory is and there are some operational definitions of how a problem with memory manifests compared to what healthy memory function displays.

Specifically, "memory" can be defined as the ability to learn, retain, and recall new information. Healthy memory function might be defined statistically as a score on a test of this ability that falls within the range of scores observed in the relevant reference group, in this case, members of the community without known deficit (Wechsler, 2010). The clinician could then adopt, for example, the criteria suggested by Albert et al. (2011) to define abnormal memory within the context of possible MCI, that is, a score on the memory test of 1–1.5 standard deviations below the mean for a patient's age- and education-matched peers.

In practice, when considering this question, the clinician might conceptualize memory as the ability to learn, retain, and recall stimulus material via prose recall or verbal paired-associate learning. Defined in this way normal memory would be indicated by a score on these tests above the 7th percentile (that is, 1.5 standard deviations below the mean) when compared with age- and education-matched peers and consider a memory deficit to be indicated by a score below the 7th percentile.

If the clinician applied these tests to patients, some of whom had a memory disorder and some of whom did not, the clinician would get a range of performance on the memory tests that had some degree of relationship with the presence or absence of the underlying memory disorder. Strauss and Smith (2009) provide detailed discussion of the more general research strategy of criterion-related validity. Many of the patients who perform below the 7th percentile on the test will demonstrate evidence of a memory disorder in their everyday life. These people can be described in the terminology of criterion-related validity as "true positives." The test scores says these people have a "memory disorder," and they appear to demonstrate problems with memory in day-to-day life.

A second category includes the people who perform below the 7th percentile on the test but nonetheless do not demonstrate any memory problems in their everyday life. These people may be described as "false positives."

Table 10.1 Categorical descriptions reflecting the association between the results of a test for impairment in a cognitive ability used to detect the "condition of interest" (COI) shown in the rows, and the real-life presence of the "condition of interest" shown in the columns. Also shown are the metrics of sensitivity and specificity.

	COI Is Present	**COI Is Not Present**
Test says "yes" to COI	True Positives A	False Positives B
Test says "no" to COI	False Negatives C	True Negatives D
	Sensitivity = A/A + C	**Specificity = D/D + B**

A third category of patients will be those who perform above the 7th percentile on the test and do not demonstrate any problem behavior in their everyday life that would suggest the presence of a memory disorder. These people are our "true negatives." The test says "no memory disorder," and the persons so identified appear to have no memory disorder.

Finally, a fourth category of patients will be those who perform above the 7th percentile on the test but nonetheless demonstrate behavior in their everyday life that would suggest the presence of a memory disorder. These people are "false negatives." The test says "no memory disorder" but they appear to have a memory disorder. The association between the results of a test for the "condition of interest" and the real-life presence of the condition of interest are represented in Table 10.1.

While it would be excellent for the practice of clinical neuropsychology to rest firmly on the basis of tests without any false negative or false positive results, this is unfortunately not the case, nor is it the case for most diagnostic tests (Straus et al., 2011). All tests are imperfect to some extent, and as a consequence, classification accuracy of every test can be quantified in terms of the four cells shown in Table 10.1, that is, true positives, false positives, false negatives and true negatives (Straus et al., 2011).

SENSITIVITY AND SPECIFICITY

The relationship between the true positive, true negative, false positive, and false negative results on a test can be expressed as the test sensitivity and specificity in relation to a specific criterion. The comparison criterion is usually termed the "external validity criterion" or "gold standard" for a criterion with the best available validity (Straus & Smith, 2009). For example, the sensitivity and specificity of a memory test for detecting memory impairment in everyday life could be compared to an external validity criterion defined by an informant's objective ratings of a person's performance in their everyday life. Another example of an external validity criterion would be expert clinician-panel consensus ratings of the presence of a diagnosis.

To continue the MCI example, sensitivity reflects how many people with a memory disorder in their everyday life have a positive test result. In this case, the number of people with a memory disorder in everyday life who have a memory test score below the 7th percentile. Usually there is an imperfect relationship between these two sources of classification. Consequently, whilst, with a good test, most of the people with a memory disorder in everyday life will score below the 7th percentile on this memory test (i.e., the test is sensitive to the presence of a memory disorder), there will be some people who have a memory disorder in everyday life but who score better on the test. This scenario reflects a false negative (i.e., the test says there is no memory disorder when in fact there is). Also there will be some people who score below the 7th percentile on the test who have no apparent memory disorder in everyday life. This scenario reflects a false positive (i.e., the test says there is a memory disorder when in fact there is not). These further metrics are shown in Table 10.1.

"Sensitivity" is calculated from the number of true positives as a percentage of the total number of "positives" in the population. In Table 10.1, this value would be reflected by the equation **A/A + C.**

"Specificity" reflects how many people without a memory disorder in real life have a negative test result, in this case, defined as a memory score above the 7th percentile. As before, whilst most people without a memory disorder will score above the 7th percentile on this memory test, there will also be some people who do not have a memory disorder but who score poorly on the test. The latter scenario reflects a false positive error. Specificity is calculated from the number of true negatives as a percentage of the total number of "negatives" in the population. In Table 10.1, this would be reflected by the equation **D/D + B.**

Sensitivity and specificity are often expressed as percentages or decimal proportions reflecting the outcomes of the two equations above. Tools for calculating these values and other values described below are readily available on the Internet, for example, see http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc or http://www.cebm.net.

There is usually a tradeoff between sensitivity and specificity for any given test. No diagnostic test is perfect, consequently, as sensitivity increases, it is usually at the expense of specificity and vice versa, as shown in Figure 10.1. If a clinician is trying to capture all the people with the condition of interest, represented by the darker distribution in Figure 10.1, then the ability of the test to do so increases as the cut-score used moves from the left to the right in Figure 10.1. That is, from "A" to "B" and finally to "C". Using the cut-score of "C," almost all those in the darker distribution are below the cut-score and so would be correctly identified by the test. In this circumstance, the sensitivity of the test is high. However, as the cut-score changes from "A" to "B" and finally to "C," it can also be seen that the number of people within the lighter distribution (which represents people without the condition) who are correctly classified decreases because more of their scores fall below the cut-score as it moved to "C." That is, as the sensitivity of the test increases, the number of false positive test results also increases, which means the specificity of the test decreases.

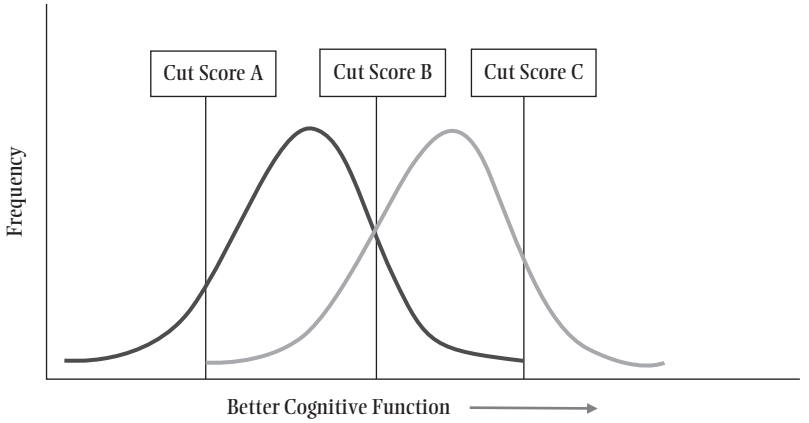

Figure 10.1 The trade-off between sensitivity and specificity when using three different cut-scores to distinguish between scores within the condition of interest distribution (darker curve on the left) versus those within the control group distribution (lighter curve on the right).

The opposite is also true: as the cut-score changes from "C" to "B" to "A," the number of false positives decreases, but so does the number of true positives identified by the test. In this latter circumstance, where the cutoff score moves from right to left in Figure 10.1, the specificity of the test increases, but at the expense of the sensitivity of the test.

Invariably, in clinical practice, the distribution of scores on tests between those who have, and those who do not have, the condition of interest overlaps. It is also the case in clinical practice that the cut-scores used to help guide the interpretation of test results need not be absolute or fixed. Consequently, the relationship between the sensitivity and specificity of a test result and the condition of interest will vary, depending upon the cut-score that is used. The choice of cut-off can also be used to help favor either sensitivity or specificity, depending upon the clinical question that is being asked. Sometimes, particularly when screening, it is usually more helpful to emphasize sensitivity over specificity (Straus et al., 2011). The reason for weighting sensitivity is that the goal of screening is usually to identify all the people who may have the condition of interest. It is more important not to miss people with the condition of interest (false negative errors) than it is to minimize potential false positive errors. Subsequently, the people whose scores are classified as positive at the first screening assessment can be reassessed with a test with a high specificity. This strategy is known as the "two-step diagnostic process" (Straus et al., 2011).

Alternatively, in some scenarios, it would be more important to emphasize specificity rather than sensitivity, that is, for decisions where the costs of false-positive errors might be high. Such a circumstance might apply with tests used to help identify people who are potentially feigning their cognitive problems. In this scenario, given the cost of wrongly diagnosing malingering, test cut-scores

are often weighted to emphasize high specificity, sometimes at the expense of sensitivity (Gervais et al., 2004).

POSITIVE AND NEGATIVE LIKELIHOOD RATIOS

A likelihood ratio is a way of estimating how much a test result should shift a clinician's index of suspicion in the direction of the "condition of interest" being present or absent. Strictly defined, likelihood ratios reflect the change in the likelihood of a diagnosis, after obtaining a positive or negative test result. In the memory disorder example above, the likelihood ratios are a direct way of indicating the change in the likelihood of a person having a memory disorder in everyday life, depending on whether that person obtained a positive or negative test result. In essence, the positive likelihood ratio shows how much more likely it is that the person tested has a memory disorder in everyday life when they obtain a positive score, namely, a score below the 7th percentile on the memory test. In the case of a negative test result, a negative likelihood ratio shows how much less likely it is that the person has a memory disorder in everyday life when they score above the 7th percentile on the memory test.

These likelihoods can be calculated by the following equations (see Grimes & Shultz, 2002):

$$\text{Positive Likelihood Ratio } (\text{LR}+) = \frac{\text{Probability that a person with the condition has a positive test result}}{\text{probability than an individual without the condition has a positive test result}}.$$

$$\text{Negative Likelihood Ratio } (\text{LR}-) = \frac{\text{Probability that a person with the condition has a negative test result}}{\text{probability that an individual without the condition has a negative test result}}.$$

In terms of the values in Table 10.1, these formulae can be written as:

$$\text{LR}+ = (\text{A}/\text{A}+\text{C}) \,/\, (\text{B}/\text{B}+\text{D})$$
$$\text{LR}- = (\text{C}/\text{A}+\text{C}) \,/\, (\text{D}/\text{B}+\text{D})$$

These equations can also be written in terms of sensitivity and specificity, namely:

$$\text{LR}+ = \text{sensitivity} \,/\, (1 - \text{specificity})$$
$$\text{LR}- = (1 - \text{sensitivity}) \,/\, \text{specificity}$$

As likelihood ratios for a positive test result increase significantly above 1, there is an increased probability of the condition of interest being present after a positive test result is obtained. Conversely, as the likelihood ratio for a negative test result decreases significantly below 1, there is a decreased probability of the condition

being present after a negative test result is obtained. As positive likelihood ratios increase above 10, for example, the probability of the condition being present is greatly increased when a positive test result is obtained. Conversely, as a negative likelihood ratio decreases below 0.1, for example, the probability of the condition being present is greatly reduced when the test result is negative.

The positive likelihood ratio is a way of thinking about a positive test result affecting the base-rate estimate to increase the likelihood of the diagnosis. Conversely, a negative likelihood ratio is a way of thinking about a negative test result affecting the base-rate estimate to reduce the likelihood of the diagnosis. In this way, the pre-test odds (base rate) are changed by the likelihood ratio, resulting in the post-test odds. Likelihood ratios are interpreted with reference to an estimated or known pre-test probability (also referred to as the "clinical prevalence" or "base rate"). A nomogram for interpreting diagnostic test results is shown in Figure 10.2. In the nomogram, a line is drawn from the pre-test probability through the likelihood ratio to estimate the post-test probability (see Fagan, 1975).

An online calculator is also available to estimate post-test probability using likelihood ratios, see http://araw.mede.uic.edu/cgi-bin/testcalc.pl

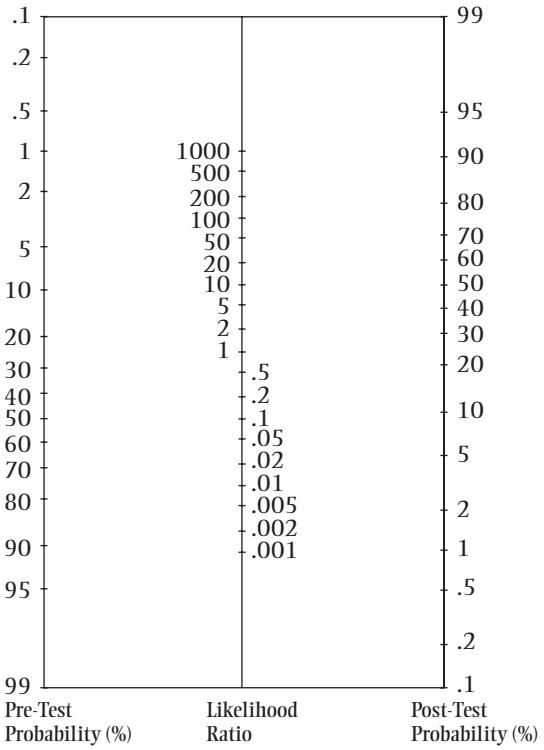

Figure 10.2 Nomogram for interpreting diagnostic test results.

Table 10.2 Examples showing how the likelihood ratio changes selected pre-test probability (base-rate) values, to re-estimate the probability of the diagnosis after obtaining the test result, termed the "post-test probability."

Pre-Test Probability (Base Rate)	Likelihood Ratio	Post-Test Probability
1%	.1	<0.1%
1%	.5	1%
1%	2	2%
1%	5	5%
1%	10	9%
10%	.1	1%
10%	.5	5%
10%	2	18%
10%	5	36%
10%	10	53%
25%	.1	3%
25%	.5	14%
25%	2	40%
25%	5	62%
25%	10	77%

To assist in the appreciation of the relationship between selected pre-test probability values, likelihood ratio values, and post-test probability, examples of the calculation are presented in Table 10.2. The interested reader could also plot the numbers presented in Table 10.2 on the nomogram presented at Figure 10.2 to see how the nomogram reduces the need for calculations to obtain an estimate of post-test probability. As can be seen from Table 10.2, the likelihood ratio allows re-estimation of the probability of the condition of interest after obtaining a positive or negative test result. In other words, the likelihood ratio acts on the base rate of the condition of interest within the population tested. This latter relationship is explored further in the following section.

POSITIVE AND NEGATIVE PREDICTIVE VALUE AND THE PREVALENCE OF THE CONDITION OF INTEREST (BASE RATE)

In clinical practice, a diagnostic test is applied to a population of interest. A clinician uses the test to help answer a question about the group membership of the person tested. For example, a memory test is used to help answer the question of whether or not the person tested is a member of the impaired-memory group or the unimpaired-memory group. Frederick and Bowden (2009) noted that, unlike the true positive and false positive classification rate of a test, the positive predictive power and negative predictive power of a test is affected by the base rate of the condition of interest within the population tested. The definition and

calculation of the positive and negative predictive values of a test are reported in relation to the four quantities in Table 10.1.

The "positive predictive value" refers to the number of people with a positive test result who actually have the condition of interest. In the memory example above, we would be asking how many people who obtain a memory test score below the 7th percentile when tested have a memory disorder in everyday life. This can be represented mathematically as **A/A + B** for the values in Table 10.1. The "negative predictive value" refers to the number of people with a negative test result who do not actually have the condition of interest. In the memory example, this would be the number of people with a memory test score above the 7th percentile who do not have a memory disorder in everyday life. This can be represented mathematically as **D/C + D** in Table 10.1.

The performance of a test in relation to a diagnostic decision is further affected by a property of the population being tested rather than of the test itself. This property is the "prevalence" (or base-rate or pre-test probability) of the condition of interest in the population being tested. In terms of the memory disorder example, this would be the number of people who actually have a memory disorder in everyday life, out of all the people being tested. This percentage is referred to as the base-rate or prevalence or pre-test probability of the condition of interest in the population.

If we considered a community-based example, the proportion of people with a memory disorder in everyday life among all those being tested would be much lower than if we considered a population of people attending a tertiary-referral dementia diagnosis clinic. In the latter setting, it would be reasonable to assume the number of people attending with a memory disorder in everyday life would be higher.

A specific example of the impact of the prevalence of the condition of interest on the diagnostic performance of a test can be seen in the study of Mioshi et al. (2006) of the diagnostic validity of the Addenbrooke Cognitive Examination–Revised. These researchers noted the sensitivity, specificity, likelihood ratios, and positive and negative predictive values of the test when making a dementia diagnosis using cut-scores of 82 and 88 on the Addenbrooke test at different levels of prevalence.

Consider the results of Mioshi et al. (2006) in relation to their cut-score of 88. At an estimated prevalence of 40% the positive predictive value of the test was 0.85, meaning that there was a .85 probability that a person with a score at 88 or below had dementia. This positive predictive value changed dramatically, however, when the presumed prevalence was 5%. At this level of prevalence, the positive predictive value of the test became 0.31, meaning that there was only a .31 probability that a person with a score at 88 or below had dementia. In other words, as the prevalence of the condition of interest, in this case dementia, lessened in the population tested, the validity of a positive test result indicating the presence of dementia also declined. At 5% prevalence, a positive test score was diagnostically accurate only 31% of the time. That is, at this 5% prevalence, a positive test score was a false positive 69% of the time. A positive test result is,

therefore, at this low level of prevalence, much more likely to be incorrect than it is correct at detecting dementia.

These ideas are further expanded below using hypothetical data applied to our previously discussed memory disorder example. In that example, the diagnosis being made was whether or not someone had a memory disorder in everyday life, and the cut-score on the memory test used was the 7th percentile.

Lets assume we started from an estimated prevalence of 50%, that is, the original research study used to derive the cut-score was made up of two equal-sized groups that were matched in terms of their demographic and clinical characteristics other than the presence of memory disorder. If a valid memory test is used, it might be expected to have a sensitivity of, say, 0.78 and a specificity of 0.82, to choose two arbitrary hypothetical values. Further suppose that the cut-score is derived from research to classify people into either the "memory disorder in everyday life group" (disorder present) or the "no memory disorder in everyday life group" (disorder not present). Suppose also, in this hypothetical example, that there are 125 people in each of these two groups. These properties of the test are represented in Table 10.3. The numbers of people in each cell (A to D) in Table 10.3 are determined by our hypothetical sensitivity and specificity values.

The sensitivity of the test shown in Table 10.3 is given by A/A + C, which in this example is 97/125 = 0.78 (rounded to two decimal places). The specificity of the test is given by D/D + B which in this example is 103/125 = 0.82. In this example, the base rate of the condition of interest was set to 50%, that is, the disorder is present in 125 people (A + C) and is not present in 125 people (B + D), so the base rate is 125/250.

The positive predictive value of the test is given by A/A + B, which in this example is 97/119 = 82%. This value indicates that there is a probability of .82 that a positive test result comes from a person with the condition of interest, in this example, memory disorder. The negative predictive value of the test is represented by D/C + D, which in this example is 103/131 = 79%. This value indicates that there is a probability of .79 that a negative test result comes from a person without memory disorder.

Table 10.3 Number of people classified into the diagnostic categories reflected by cells A to D for the hypothetical memory disorder example, with a study prevalence or base rate of 50%.

	Disorder Is Present	**Disorder Is Not Present**	**Total**
Test says "yes"	True Positives A 97	False Positives B 22	119
Test says "no"	False Negatives C 28	True Negatives D 103	131
Total	125	125	

The interpretation of the usefulness of the test changes, however, if the base rate of the condition changes, especially if the base rate falls to a lower level than was represented in the research study. Base rates will change if the population tested is different from the research study population, for example, if all the people from a community-based population are tested. When compared with the research study population, the community-based population would have a lower base rate of the condition of interest. The base rate is lower because the prevalence of the condition of interest will be diluted across many more people than in the research study, where the condition of interest was deliberately identified and concentrated into one of the groups tested.

If the same test and cut-off from the original research study are applied to the community-based population, the interpretation of the test results will be different. This is because, in the community-based population, the base rate of the condition of interest is lower. Let us suppose, in this example, that a base rate of 9% reflects the frequency of memory disorder in the community-based population. A reworking of the preceding calculations, but with the lower base rate, is shown in Table 10.4.

For the example in Table 10.4, the sensitivity of the test is given by A/A + C and stays the same as in the previous example, that is, it is 97/125 = 0.78. The specificity of the test is represented by D/D + B and stays the same as in the previous example, that is, it is 1030/1250 = 0.82. However, in Table 10.4, the base rate of the condition is now 9%, the disorder is present in 125 people and is not present in 1,250 people, that is, 125/(125 + 1250).

Therefore, the positive predictive value of the test, which is represented by A/A + B, is now in this example 97/317 = 31%. That is, the probability of a positive test result coming from a person with memory disorder is now only .31. To put it another way, out of all the positive test results obtained, 31% are true positives and reflect the presence of the condition of interest and 69% are false positives. The negative predictive value of the test is represented by D/C + D, which in this example is now 1030/1058 = 97%, that is, out of all the negative test results

Table 10.4 Number of people classified into the diagnostic categories reflected by cells A to D for the hypothetical memory disorder example, with a base rate of 9%.

	Disorder Is Present	**Disorder Is Not Present**	**Total**
Test says "yes"	True Positives A 97	False Positives B 220	317
Test says "no"	False Negatives C 28	True Negatives D 1030	1058
Total	125	1250	

obtained, 97% are true negatives and reflect the absence of the condition of interest and 3% are false negatives.

These calculations and the numbers in Table 10.4 show that, when the base rate of the condition of interest tends towards zero, we can become more confident in negative test results but less confident in positive test results. The negative predictive value increased from 79% to 97% as the base rate decreased in these two examples. A negative test result was more likely to be accurate and the number of false-negatives less as the base rate of the condition of interest decreased. However, the positive predictive value decreased from 81% to 31% as the base rate decreased from 50% to 9%. That is, a positive test result was less likely to be accurate and the number of false positives increased in proportion to the true positives as the base rate of the condition of interest decreased.

So, in general, when the base rate of the condition of interest is low, a test that appears to have good diagnostic properties (when calculated under the high base-rate conditions often found in published research studies) can actually perform so poorly that a positive test result is more likely to be wrong than it is right, that is, a positive test result is more likely to be a false positive than a true positive. The impact of prevalence upon the predictive power of diagnostic tests is discussed in detail in Baldessarini et al. (1983).

The implication of these calculations is that, when applying these principles to diagnostic decision-making in clinical practice, it is necessary to have some idea of the base rate of the condition of interest within the clinical setting in which the test is being used. This information allows for the calculations necessary to estimate how the test will perform when taken from the research setting to a clinical setting with a different base rate. Returning to the Mioshi et al. (2006) example, while sensitivity and specificity were high using the cut-score of 88, the positive predictive value of the test was shown to be poor at low base rates, it was 0.31 at a 5% base rate. In other words, at this low base rate, a positive test result was much more likely to be a false positive than a true positive. At the higher base rate of 40%, the positive predictive value was much higher, at 0.85, indicating a much reduced likelihood of false positive test results. Before using this test at the prescribed cut-offs to make diagnostic decisions, any clinician would be wise to estimate the base rate of the condition of interest in the population being tested to avoid the error of interpreting a false positive as a true positive.

Putting these ideas together allows for the calculation of the post-test probability at any specified base rate.

The post-test probability is calculated as follows:

$$\text{Post-Test Probability} = \text{Prevalence} / \left(1 - \text{Prevalence}\right) \times \text{Likelihood Ratio} / \left[\text{Prevalence} / \left(\left(1 - \text{Prevalence}\right) \times \text{Likelihood Ratio}\right) + 1\right]$$

In the example described here, with a prevalence of 9%, the post-test probability of the person with a memory test score below the 7th percentile having a memory problem in everyday life is calculated as follows, when using the data presented

in Table 10.4. LR+ = 4.41, Prevalence = 0.09, Post-Test Probability = 30% probability of a memory disorder in everyday life. When the prevalence is assumed to be 50%, the post-test probability becomes: LR+ = 4.41, Prevalence = 0.5, Post-Test Probability = 82% probability of a memory disorder in everyday life. See http://araw.mede.uic.edu/cgi-bin/testcalc.pl for an online calculator.

RELIABILITY OF MEASUREMENT

When considering the likely diagnostic validity of a test procedure, it is also crucial to consider the reliability of the test result. When thinking about diagnostic decision-making, we are essentially attempting to arrive at a decision, namely, is the condition of interest present or not? The reliability of our test result has a crucial bearing on the confidence we have in the decision we are making. Put simply, nothing can be more valid than it is reliable. On average, a test result cannot correlate better with some diagnostic outcome than it can correlate with itself. If a test score is not relatively reliable, it cannot have high validity and therefore cannot be of high diagnostic utility (Schmidt & Hunter, 1996).

When thinking about diagnostic validity as described above, it is necessary to turn the score on a test into a decision. This is usually achieved by considering whether the obtained test score falls above or below a cut-score that has been empirically derived to maximize the accuracy of classification. For example, when considering performance validity during cognitive testing, Schroeder et al. (2012) highlighted cut-scores of less than or equal to 6 or 7 on the Reliable Digit Span measure as being optimal, depending upon the population being tested. Any decision about whether or not a score falls above or below a specific cut-off needs to consider the reliability of the score itself. The reliability of the Digit Span subtest is high, which encourages confidence in the obtained result but if, for example, the test score were very unreliable, one could only have limited confidence that the score obtained at one assessment would reflect the person's true score. If the test score is unreliable and a patient is tested again, their score might vary, and thus the patient could be classified as being above or below the cut-off at different points in time merely as a consequence of poor measurement reliability. Such measurement unreliability fundamentally undermines the diagnostic validity possible with any test of lower reliability. Reliability of test scores is examined in detail in Chapter 5 of the current volume.

CRITICALLY APPRAISED TOPIC

With the Critically Appraised Topic (CAT) procedure, evidence for the validity of a diagnostic test, including the evidence noted above, is critically appraised to help a clinician answer a specific question (see Bowden et al., 2013; and Chapters 11 and 12 of this volume). Critical appraisal supports clinicians in translating the research evidence available to them into practical guidance regarding how to interpret a particular test score in a particular circumstance.

In summary, the CAT procedure is a systematic way of collating, appraising, and making use of the available research evidence to guide clinical practice.

The CAT procedure encourages clinicians to critically evaluate the quality of the scientific evidence available to them and use the calculations herein to help them determine the most appropriate way to interpret the results of testing undertaken with their patient. For further discussion, the reader is referred to Chapter 11 of this volume where a CAT of a performance validity test is presented.

REFERENCES

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., . . . Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute of Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dementia, 7*(3), 270–279.

Baldessarini, R. J., Finklestein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry, 40*, 569–573.

Bowden, S. C., Harrison, E. J., & Loring, D. W. (2013). Evaluating research for clinical significance: Using critically appraised topics to enhance evidence-based neuropsychology. *The Clinical Neuropsychologist, 28*(4), 653–668.

Faust, D. (2003). Alternatives to four clinical and research traditions in malingering detection. In P. W. Halligan, C. Bass, & D. A. Oakley (Eds.), *Malingering and Illness Deception* (pages 107–121). Oxford: Oxford University Press.

Fagan, T. J. (1975). Nomogram for Bayes's theorem. *New England Journal of Medicine, 293*(5), 257.

Frederick, R. I., & Bowden, S. C. (2009). The test validation summary. *Assessment, 16*(3), 215–236.

Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology, 19*, 475–487.

Grimes, D. A., & Schulz, K. F. (2002). An overview of clinical research: The lay of the land. *Lancet, 359*(9300), 57–61.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy and Law, 2*(2), 293–323.

Haynes, R. B., Devereaux, P. J., & Guyatt, G. H. (2002). Clinical expertise in the era of evidence-based medicine and patient choice. *Evidence Based Medicine, 7*, 36–38.

Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., & Hodges, J. R. (2006). The Addenbrooke's Cognitive Examination–Revised (ACE-R): A brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry, 21*, 1078–1085.

Ruscio, J. (2007). The clinician as subject. Practitioners are prone to the same judgment errors as everyone else. In S. O. Lilenfield & W. T. O'Donohue (Eds.), *The Great Ideas of Clinical Science: 17 Principles That Every Mental Health Professional Should Understand* (pages 29–48). New York: Routledge.

Schoenberg, M. R., & Scott, J. G. (Eds.). (2011). *The Little Black Book of Neuropsychology: A Syndrome-Based Approach*. New York: Springer.
Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable digit span: A systematic review and cross-validation study. *Assessment, 19*(1), 21–30.
Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199–223.
Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25.
Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-Based Medicine. How to Practice and Teach It* (4th ed.). Churchill Livingstone Elsevier.
Wechsler, D. (2010). *Wechsler Memory Scale–Fourth UK Edition. Administration and Scoring Manual.* London: Pearson Education, Pearson Assessment.

FURTHER READING

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-Based Medicine. How to Practice and Teach It* (4th ed.). Churchill Livingstone Elsevier.