CRITICAL APPRAISAL FOR EMERGENCY MEDICINE TRAINEES
5. EVALUATION OF A DIAGNOSTIC TEST

Studies of diagnostic tests commonly inform practice in emergency medicine. They aim to determine whether the test under investigation accurately identifies patients with and without the disease, as defined by a reference standard test. Several key issues need to be considered in appraising whether a study is likely to yield a reliable estimate of the diagnostic parameters. It is also important to be able to interpret what the diagnostic parameters actually mean.

The reference standard (gold standard)

The reference standard is the criterion by which it is decided that the patient has, or does not have, the disease. Typical reference standards might be:

- A single diagnostic test that is known to be accurate, e.g. contrast venography for deep vein thrombosis
- A combination of diagnostic tests that will reliably rule-in and rule-out disease, e.g. lung perfusion scanning for pulmonary embolus combined with pulmonary angiography in equivocal cases
- Diagnostic testing with follow-up for negative cases to identify cases of disease that may have initially been misclassified as no disease

An ideal reference standard should correctly classify patients with and without disease. However, it should also be safe and simple to apply, because it would be unethical to ask patients to undergo dangerous or complex testing purely for research purposes. If an ideal reference standard does exist, then there is little need to evaluate new diagnostic tests! So we have something of a catch-22 situation. This is why judging whether a reference standard is acceptable involves weighing its' potential accuracy against the feasibility of any alternative approaches.

The choice of reference standard will often involve a trade-off between validity and generalisability. A study that uses a highly accurate reference standard in all patients is likely to be valid, but may struggle to recruit a wide spectrum of patients (see below). A study that uses a pragmatic reference standard, such as a combination of routinely available tests, may recruit a wide spectrum of patients, but may misclassify patients' disease status, leading to bias.

Independence of the reference standard

Ideally, the same reference standard should be applied to all patients, regardless of the results of the diagnostic test under evaluation. If this not possible then the diagnostic test under evaluation should not determine which reference standard is applied.

Two situations commonly occur in which lack of an independent reference standard leads to bias:

1. The diagnostic test under evaluation determines which reference standard is used. This is known as *work-up bias* and is illustrated in example 1.
2. The diagnostic test under evaluation forms part of the reference standard. This is known as *incorporation bias* and is illustrated in example 2.

These types of bias will tend to produce inflated estimates of diagnostic parameters. Work-up bias tends to over-estimate sensitivity because patients with the disease who

have a false negative result for the test under investigation may receive a “lesser” reference standard that fails to identify the disease. Incorporation bias on the other hand, tends to over-estimate specificity because patients without the disease who have a false positive result will be incorrectly classified as having the disease.

Blinding

The person measuring or interpreting the diagnostic test under evaluation should be blinded to the results of the reference standard. If they are not blinded they may be influenced in their measurement or interpretation by their knowledge of the reference standard result. Likewise, the person measuring or interpreting the reference standard should be blinded to the results of the diagnostic test under evaluation. Failure to institute blinding will tend to over-estimate sensitivity and specificity, as shown in example 3.

Studies may not report if blinding was undertaken, yet it may be possible to infer whether this is likely to be a source of significant bias. Firstly, the diagnostic process needs to involve a subjective element, such as X-ray interpretation, if it is to be influenced by lack of blinding. So quantitative measurement of a blood sample on a laboratory analyser is unlikely to be influenced by lack of blinding. Secondly, if the test under investigation is performed and interpreted before the reference standard then it is reasonable to infer that it was performed and interpreted blind to the reference standard.

Patient spectrum

The study population should be representative of the population who would receive the test in routine practice. If the population is highly selected then this will bias estimates of sensitivity and specificity. As described above, use of an invasive reference standard may lead to a selected population.

The disease prevalence provides a useful clue as to the degree of patient selection. A population with high prevalence are probably highly selected. However, it is often very difficult to achieve a low prevalence population. The research process typically involves asking patients to undergo multiple tests. Clinicians are reluctant to enrol and patients are reluctant to participate if there is a low probability of their having the disease.

Some diagnostic test evaluations assemble the study population by selecting patients on the basis of their reference standard test, i.e. selecting a group of patients with the disease and a group without. This is often known as a case-control design. It is very prone to bias and over-estimation of sensitivity and specificity. It can be useful in the early stages of evaluation to identify diagnostic tests for further investigation, but estimates of sensitivity and specificity derived from it should not be used in clinical practice.

Interobserver error (reliability)

A diagnostic test or clinical finding is unreliable if it gives different results when performed by different clinicians. For example, if two radiologists frequently produce conflicting reports from the same CT scan, then CT scanning (in this circumstance) is an unreliable test.

Evaluations of diagnostic tests should include some assessment of reliability, but they seldom do. Reliability cannot be estimated by simply measuring the percentage agreement between two observers because agreement may occur simply by chance. For example, if a test has only two possible results (positive and negative) then there is a 50% probability that two observers will agree in their interpretation purely by chance.

The most common method for estimating reliability is to measure the Kappa score. This calculates the agreement between observers beyond that expected due to chance. Values range from 0 (chance agreement only) to 1 (perfect agreement).

Terms used in reporting diagnostic test data
The following terms are often used to report diagnostic test data. It is well worth being absolutely sure that you know exactly what they mean. Specificity, in particular, is often confused with positive predictive value.

Case positive
An individual with the disease of interest, i.e. the reference standard is positive.

Case negative
An individual without the disease of interest, i.e. the reference standard is negative.

Test positive
An individual with a positive result for the diagnostic test under investigation.

Test negative
An individual with a negative result for the diagnostic test under investigation.

Prevalence
The proportion of the population with the condition of interest.

True positives
Patients correctly identified by the diagnostic test as having the disease.

True negatives
Patients correctly identified by the diagnostic test as not having the disease.

False positives
Patients without the disease who are incorrectly labelled by the diagnostic test as having the disease.

False negatives
Patients with the disease who are incorrectly labelled by the diagnostic test as not having the disease.

Sensitivity
The proportion of patients with the disease who are correctly identified by the test.

Specificity
The proportion of patients without the disease who are correctly identified by the test.

Positive Predictive Value (PPV)
The proportion of patients with a positive test who genuinely have the disease.

Negative Predictive Value (NPV)
The proportion of patients with a negative test who genuinely do not have the disease.

Using diagnostic parameters
Sensitivity and specificity are the most frequently reported diagnostic parameters. They can be used in the following ways to rule in or rule out the disease:

- Sensitivity is important if a negative test result is being used to rule out a disease. Sensitivity + Negative + Out = SnNOut
- Specificity is important if a positive test result is being used to rule a disease in. Specificity + Positive + In = SpPIn

The table below can help you to understand the relationship between prevalence and the test parameters. It will not help you to remember what sensitivity and specificity are. It could positively confuse you if you get it the wrong way round! For this reason it is probably better to make sure that you understand what each term means than worry about the maths.

	Case positive	**Case negative**
Test positive	A	B
Test negative	C	D

Sensitivity = A / (A+C)
Specificity = D / (B+D)
PPV = A / (A+B)
NPV = D / (C+D)
Prevalence = (A+C) / (A+B+C+D)

Sensitivity and specificity can be usefully applied at population level, but are difficult to use at an individual patient level. PPV and NPV are more useful at an individual patient level because they tell us the probability that the patient has the disease. However, as the table above shows, PPV and NPV will depend upon prevalence, where as sensitivity and specificity are constant.

If prevalence increases, A and C will increase, while B and D decrease. So both the numerator and the denominator for sensitivity will increase, and both the numerator and the denominator for specificity will decrease. Therefore, sensitivity and specificity remain constant. However, the numerator for PPV will increase, while the denominator remains (roughly) constant, so PPV increases as prevalence increases. Whereas the numerator for NPV will decrease, while the denominator remains (roughly) constant, so NPV decreases as prevalence increases.

Sensitivity and specificity are constant when the prevalence varies PPV increases with increasing prevalence NPV decreases with increasing prevalence

missing text - embolus in chest pain, fractures in ankle injury, subarachnoid

haemorrhage in headache. Hence NPV may appear superficially impressive, while PPV appears superficially poor.

Although sensitivity and specificity are mathematically constant as prevalence varies, they may have different values if the test is used in a different population. If prevalence is observed to vary between populations of interest, it is worth asking whether the populations are sufficiently similar to allow extrapolation of results from one population to another.

For example, sensitivity and specificity should remain constant, whether tested in a coronary care population of chest pain patients who have a prevalence of myocardial infarction of 30%, or tested in an emergency department population of chest pain patients who have a prevalence of 5%. However, these two populations may be so different that the test actually performs completely differently.

Likelihood ratios

Likelihood ratios provide a more useful way of presenting diagnostic data and can be applied to individual patients in a way that sensitivity and specificity cannot. Many people are put off using likelihood ratios by attempting to understand the statistics, particularly when the word "Bayesian" is used. This is entirely unnecessary. The great advantage of using likelihood ratios in a Bayesian approach to diagnosis is that it replicates the way in which clinicians intuitively use information to adjust the estimate of the probability of disease.

A likelihood ratio:

- Applies to a piece of diagnostic information, such as an observation, a clinical finding or a test result
- Tells you how useful that piece of information is when you are trying to make a diagnosis
- Is a number between zero and infinity
- If greater than one, indicates that the information increases the likelihood of the suspected diagnosis
- If less than one, indicates that the information decreases the likelihood of the suspected diagnosis

The table below shows how likelihood ratios indicate the value of a piece of diagnostic information.

Likelihood ratio	Value of additional information
1	None at all
0.5 to 2	Little clinical significance
2 to 5	Moderately increases likelihood of disease. Useful additional information, but does not rule-in
0.2 to 0.5	Moderately decreases likelihood of disease. Useful additional information, but does not rule-out
5 to 10	Markedly increases likelihood of disease. May rule-in if other information is supportive
0.1 to 0.2	Markedly decreases likelihood of disease. May rule-out if other information is supportive.
Over 10	Diagnostic. If this does not convince you that the patient has the

	disease then you probably shouldn't have done the test.
Less than 0.1	Rules out disease.

Studies evaluating diagnostic tests should present their results as likelihood ratios. If they do not, likelihood ratios for a simple dichotomous (positive or negative) test can be calculated from sensitivity and specificity as follows:

Likelihood ratio of positive test = sensitivity / (1-specificity)

Likelihood ratio of negative test = (1-sensitivity) / specificity

Using this simple piece of maths to estimate likelihood ratios from sensitivity and specificity can be very helpful in determining whether a diagnostic test will useful in practice.

Random error

As outlined in the second article in this series, all studies should address the potential influence of random error (chance) upon results. For diagnostic test studies this is best done by calculating a 95% confidence interval for each diagnostic parameter. It is worth then bearing in mind that the true value could lie anywhere within the confidence interval.

Summary

Different studies of the same diagnostic test often produce widely varying results. This variation is often due to differences in the patient population, the way the test was used or the people using the test, rather than methodological issues. Therefore appraising a diagnostic test study should rarely lead to our rejecting it as "fundamentally flawed" but should lead to us carefully questioning whether it applies to our patients and departments.

Example 1: Work-up bias

A clinical probability score for pulmonary embolus (PE) has been developed that dichotomises patients into PE likely and PE unlikely groups. To evaluate the score in practice patients classified as PE likely received a reference standard test of CT pulmonary angiography, while those classified as PE unlikely received a reference standard of D-dimer testing with CT pulmonary angiography only if the D-dimer is positive.

In this study, patients with PE who are classified as PE unlikely will have an (arguably) inadequate reference standard. Since D-dimer has only 90% sensitivity it will misclassify 10% of these false negative patients as true negatives. This will lead to an overestimate of the sensitivity of the clinical probability score for detecting PE.

Example 2: Incorporation bias

An evaluation of a new troponin assay for diagnosing myocardial infarction used the European Society of Cardiology / American Heart Association (ESC/AHA) definition of myocardial infarction as a reference standard. However, a positive troponin result is one of the diagnostic criteria used in the ESC/AHA definition. Even if a different troponin assay is used in the reference standard it will not be a truly independent test. Hence, the test under investigation is incorporated in the reference standard.

In this study, any patient who has a positive troponin but no myocardial infarction (for example, a troponin elevation due to renal failure) is likely to be misclassified as having a myocardial infarction. Hence the study will over-estimate the specificity of the new troponin assay.

Example 3

An evaluation of emergency physician ultrasonography to detect free intraperitoneal fluid in trauma compared ultrasound to a reference standard of CT scanning. The radiologists interpreting the CT scan insisted on being provided with full clinical details before they interpreted the CT scan.

The reference standard test (CT scan) was not interpreted blind to the test under investigation (ultrasound). The radiologists would have been aware of the findings of ultrasound and this may have altered their interpretation of the CT scan. For example, if the ultrasound reported free intraperitoneal fluid then the radiologist may have looked more carefully for fluid on the CT scan than if the ultrasound was negative. This bias could lead to over-estimates of both sensitivity and specificity.

The converse situation could also lead to bias. If the emergency physician performed the ultrasound already knowing what the CT scan showed then they might be influenced to either look very carefully for fluid if they knew the CT was positive, or ignore any features suggesting fluid if CT was negative.