CRITICAL APPRAISAL FOR EMERGENCY MEDICINE TRAINEES

MODULE 1. CONCEPTS AND DEFINITIONS

What is a hypothesis?

A hypothesis is a prediction. Having made a prediction, observation or experimentation is then used to determine whether the prediction is true.

Validity and generalisability?

Critical appraisal involves determining whether the findings of a research study are valid and generalisable. If the findings are likely to be true, then they are valid. If the findings are likely to apply to settings or situations outside the research study, then they are generalisable.

Validity = is this finding true?
Generalisability = is this finding applicable elsewhere?

There is little point trying to generalise a finding that is not valid. So validity should be considered before generalisability.

There is often a trade-off between validity and generalisability. Tight experimental control may produce valid results that are difficult to generalise. Broadening criteria to enhance generalisability can risk validity if experimental control is lost.

For example, a double-blind placebo-controlled trial in a centre of excellence, with patients who agree to (and attend) rigorous follow-up, is likely to produce valid findings, but they may not be generalisable. A multicentre observational study of unselected patients in a routine hospital setting will produce generalisable findings, but validity may be compromised.

Chance, bias and confounding

There are three reasons why the findings of a research study may not be valid:

1. The results may have been affected by chance (i.e. due to a random error)
2. The results may have been affected by bias (i.e. a systematic error)
3. The results may have been misinterpreted, and ascribed to one factor, when another factor (a confounder) was actually responsible.

<u>1. Chance</u>

Random errors (chance) reflect the observation that most systems, be they human bodies or emergency departments, are subject to variation. Some people are healthier than others and some emergency departments have better staffing. Any measurement of these systems may be influenced by the play of chance. For example, it may just be bad luck that an emergency department has long waiting times on the day that we measure them.

The probability of a random error is estimated using statistics (p values and confidence intervals). The impact of random error depends upon how much variation there is in the population studied and the number of observations used to estimate the measurement (the sample size).

Random error will determine the precision of the results. The greater the sample size, the less the overall estimate will be affected by random error, and the more precise the estimate will be.

2. Bias

Bias reflects a systematic error in the methods used in the research. For example in the way the study sample was selected or the measurements were made. Many forms of bias have been described: selection bias, measurement bias, analysis bias etc. The important thing is to understand how any bias may occur and how it may affect the results rather than being able to name or classify it.

If a measurement is subject to bias it will be inaccurate.

Chance = Random error, which leads to imprecision
Bias = Systematic error, which leads to inaccuracy

3. Confounding

Confounding is an error of interpretation. The results of the study may be precise and accurate, but they are misinterpreted and a false conclusion is drawn.

Confounding may happen when we look for an association between a factor and an outcome (for example, between drinking coffee and developing lung cancer). Confounding describes the situation where the apparent association between a factor and an outcome is actually mediated by another unmeasured factor (the confounder). For example, we may observe that people who drink coffee are more likely to develop lung cancer than those who don't. However, this apparent association would probably be confounded by smoking: smokers being more likely to drink coffee and to develop lung cancer.

If a confounder is known, it can be taken into account during analysis. Common confounders include age, gender, smoking, socio-economic status, and previous morbidity. These should always be considered in analysis of non-randomised data.

Unknown confounders cannot be taken into account during analysis. However, randomisation ensures known and unknown confounders are randomly distributed between groups in a study.

Accuracy and precision

Accuracy and precision both describe how close an estimate is to the true value. An inaccurate estimate will differ from the true value because bias has led to a systematic error in the estimate. An imprecise estimate will differ from the true value because random variation has led to a random error in the estimate.

Statistical techniques, such as confidence intervals, can give you an idea of the precision of an estimate. Wide confidence intervals indicate an imprecise estimate. Narrow confidence intervals indicate a precise estimate.

You can only determine whether an estimate is accurate or not by looking at the methods used in the study and deciding whether these methods may have led to bias.

Efficacy and effectiveness
Efficacy and effectiveness are not the same. A study of efficacy determines whether a treatment can work under ideal conditions. A study of effectiveness shows whether a treatment does work under normal conditions.

Pragmatic and explanatory research
When appraising a study it is important to identify what sort of research question is being asked. The methods need to be appropriate for the question. We can only determine whether the methods are appropriate if we know what sort of question is being asked. Research questions in emergency care can be broadly characterised as either pragmatic or explanatory.

Pragmatic research simply asks whether a treatment works, or how useful a test is, in routine practice. It does not attempt to determine whether the treatment could work under certain circumstances or to determine how or why a treatment works. It simply attempts to answer pragmatic questions, like "should we use this treatment?"

Pragmatic research should use routine staff and settings, unselected populations, research methods that do not interfere with clinical practice, and should measure outcomes that are directly relevant to patients, such as mortality or quality of life.

Explanatory research explores how or why a treatment works, or whether it works under specific (usually ideal) circumstances. Explanatory research may use specific staff or settings, selected populations, and may measure clinical outcomes, such as PFR, blood pressure or radiological appearance. The research methods may interfere with clinical care or produce care that is highly structured and protocol-driven.

Two apparently similar research questions may require different methods, depending upon whether they are pragmatic or explanatory. For example:

Pragmatic: Should non-invasive ventilation be routine treatment for patients presenting with acute cardiogenic pulmonary oedema?
This question would require a multicentre randomised trial involving a variety of hospitals. All patients who appear to have cardiogenic pulmonary oedema on the basis of routine testing should be recruited and randomised. Regular staff should provide the treatment according to simple protocols that allow plenty of scope for physician judgement. Some patients might not receive the "appropriate" treatment, but all would be analysed as if they had. Outcomes would be mortality and quality of life.

Explanatory: Can non-invasive ventilation improve outcomes for patients with acute cardiogenic pulmonary oedema?
This question could be performed in a single centre, perhaps a specialist hospital with an interest in this topic. Patients could be selected if they appear suitable for non-invasive ventilation. Patients may receive additional testing, such as echocardiography, to confirm the diagnosis before they are recruited. Specialist staff with additional training would provide the treatment according to strict protocols. Outcomes would include physiological measures, such as change arterial blood gases or oxygen saturation.

MODULE 2. STATISTICS

Statistics can be used to estimate the potential effect of random error upon the results of a study. There are broadly two approaches, depending upon the way the research question is asked:

1. Hypothesis testing – the P-value
2. Estimation – the confidence interval

There is no reason why both approaches cannot be used in the same analysis. Indeed they often complement each other. However, a number of journals, including the BMJ and Annals of Emergency Medicine, prefer confidence intervals to be reported rather than P-values.

Hypothesis testing (the P-value)

The research question is phrased in the form of a hypothesis and the data are collected to determine whether the hypothesis is true. For example, if we were evaluating the effectiveness of drug A compared to placebo we would ask: Is drug A more effective than placebo? The study would compare the effect of drug A to placebo and determine the probability that this effect is due to chance.

The objectives should state a hypothesis. The opposite of the hypothesis (i.e. that the stated hypothesis is false), is known as the null hypothesis. So the null hypothesis might be that there is no difference between active treatment and placebo, or that there is no difference between our sample and the rest of the population.

The p value can be expressed in two different ways, depending upon how precise/pedantic you wish to be with your definition:

1. Put simply, the p-value is the probability that the null hypothesis is true. This makes sense and is easy to understand. The smaller the p value is, the less likely the null hypothesis is to be true, therefore the more likely we are to reject the null hypothesis and accept out alternative hypothesis.
2. Strictly speaking the null hypothesis must be either true or false. It cannot be "probably true" and whether it is true or not cannot be changed by the findings of a research study. Therefore, a more precise definition is that the p value is the probability of observing results at least as extreme as those we have, given that the null hypothesis is true. The smaller the p value, the more unlikely it is that we would observe these results given the null hypothesis being true. Hence the more likely we are to reject the null hypothesis.

Don't worry if you don't understand the more complicated definition, it's perfectly possible to survive statistics without making any mistakes by using the simple definition.

Estimation (confidence intervals)

The research question is phrased as a measurement and the data are collected to provide an estimate of the measurement. For example, evaluating the effectiveness of drug A compared to placebo would involve asking: What is the effectiveness of drug A compared to placebo? The study would estimate the relative or absolute risk reduction and use confidence intervals to indicate uncertainty around this estimate.

The 95% confidence interval is a range of values around an estimate that have a 95% probability of encompassing the "true" value of that estimate. Put simply, the true value probably lies within the confidence interval. Confidence intervals will tell you how precise an estimate is. The wider the confidence interval, the less precise the estimate is.

Confidence intervals or p-values?
The hypothesis testing and estimation approaches both have advantages and disadvantages:

1. Hypothesis testing should require the researcher to decide before the study what difference they consider to be statistically significant (although this is often not the case). It therefore has the advantage of requiring a definition of "success" against which the treatment or test may be judged.
2. Both P-values and confidence intervals can tell you whether a result is statistically significant: the p-value if it is less than 0.05, the confidence interval if it does not cross the value for no effect (e.g. a relative risk reduction of 1).
3. Confidence intervals provide information about clinical significance, regardless of whether the result is statistically significant or not. P-values do not.
4. Confidence intervals can be used to estimate the likelihood of a type II statistical error (see below)
5. Too many p-values in an article suggest the possibility of multiple hypothesis testing and type I statistical errors (see below)

The list above suggests that confidence intervals have rather more advantages than p-values. At a very simplistic level it is reasonable to interpret this as "confidence intervals = good, p-values = bad".

Type I and type II statistical errors
The table below shows how type I and type II statistical errors are defined.

	Alternative hypothesis is true	Null hypothesis is true
Experiment shows significant result	True positive No error	False positive Type I error
Experiment shows no significant result	False negative Type II error	True negative No error

The probability of producing a type I or type II statistical error depends upon the sample size and the level at which statistical significance is set.

The level of statistical significance is the p value below which we consider the results to be so improbable, given the null hypothesis, that we will reject the null hypothesis and accept our alternative. By convention it is normally set at 0.05.

The level at which statistical significance is set is called alpha. The p value is the probability that a significant (i.e. positive) result is actually a false positive. The value at which we set alpha is therefore the maximum probability of a false positive that we are prepared to accept.

Having set alpha to determine what probability of a false positive result we are prepared to accept, the probability of a false negative result (defined as beta) is determined by the sample size. The larger the sample size, the smaller beta will be.

The power of a study is defined as 1-beta. By convention, a study should aim to recruit a sufficient sample size for the power to be 80 or 90%.

Study power is determined by- 1) The level at which alpha is set, 2) The sample size, 3) The variability of the outcome measure, as defined by its standard deviation, and 4) The minimum clinically significant difference we wish to detect.

Since 1) is typically set by convention and 3) is beyond our control, researchers should adjust the sample size to detect the minimum clinically significant difference. In the real world, they may be tempted to adjust the minimum clinically significant difference to fit the sample size they think they can achieve.

Type I (false-positive) errors

For any individual test, the probability of a false positive result is reflected in the p value. However, if more than one test has been performed this is no longer true. With alpha set at 0.05 the probability of obtaining a false positive result if there is no true difference is 1-(0.95 x 0.95) for two tests, 1-(0.95 x 0.95 x 0.95) for three tests, etc. If you do enough tests you will ultimately get a positive result even if there is no true difference.

Therefore, isolated positive results among a series of tests should be viewed with suspicion, particularly if there is no scientific rationale why that test should be positive and not the others, or why that test has been done, and not any others.

For example, if the authors report a significant result in blue-eyed women in the 30-40 age group, it is reasonable to assume that this was one of many (presumably non-significant) results in all other age groups and eye colours, and therefore likely to be a chance finding. Unless the authors can convince you that there is a scientific rationale for studying this subgroup in more detail.

Multiple hypothesis testing is a common flaw in poorly planned studies. If researchers collect their data without any clear objective and then analyse the data to look for any statistically significant results, they will almost certainly find some – and they will almost certainly be meaningless, false positive results.

Multiple hypothesis testing is easy to spot when all the authors report all the hypothesis tests they have performed. However, they will often only report the positive (significant) results. It is worth asking the following questions about any positive result, particularly from observational data, to identify the possibility of multiple hypothesis testing:

1. Is there a clear rationale for the hypothesis tests? This should be explained in the background.
2. Does the hypothesis test flow from the study objectives or does it only appear in the results?

3. Does the methods section include a plan for analysis that sets out what tests will be performed, does it include the positive test, and does it follow from the objectives?
4. Do the reported hypothesis tests suggest other, more obvious, associations that should be explored, but have not been reported? For example, it is slightly suspicious if a significant p-value has been reported for an association between educational status and outcome has been reported, but no association between age or gender and outcome have been explored.

Type II (false-negative) errors

If a study produces a negative result it is worth asking the following questions to identify whether it could be a type II (false negative) error:

1) Look for confidence intervals. If they are wide this suggests that estimates are imprecise and a false negative result more likely.
2) Look at the extreme ends of the confidence intervals. If important differences are possible within the confidence intervals then this study has not ruled out the possibility of an important effect, even though the study is negative.
3) Look at the power calculation. Were sensible values for alpha and beta used, and what was the minimum clinically significant difference. Is this really the smallest difference that would be worthwhile detecting? Has it been justified?

MODULE 3. EVALUATION OF A THERAPY

Evaluation of a therapy involves comparing a group of patients receiving the therapy to a group of patients who do not receive it (the control group). With a few rare exceptions (such as diseases that currently have 100% mortality) a control group is always required to demonstrate that any improvement observed after treatment is not simply due to the natural course of the illness.

There are a number of key elements of the design of these studies that will determine whether the findings are valid and generalisable.

Selection and allocation of study participants

Patients are selected to a trial by a process of recruitment that usually involves identification, assessment of eligibility, and then request for consent to participate. Selection processes can occur at any of these stages to influence the constitution of the study population.

Selection of patients for a trial will affect generalisability. If most eligible patients are identified and recruited then the results will be generalisable to the wider population. If recruitment is highly selective then findings may not be generalisable.

Once patients have been selected into a trial, they are then allocated to a treatment or to control. Bias may result if patients, carers or researchers can influence allocation. For example, patients may choose a treatment that they think will be beneficial. This will result in certain types of patient being allocated to certain treatments, leading to bias. The more that patients, carers and researchers can influence allocation to treatment group, the greater bias is likely to arise. This bias may be known as allocation bias or selection bias.

Randomisation

Randomisation is a technique used to ensure that allocation to treatment group is completely removed from any influence by carers, patients or researchers. Patients are allocated to treatment group by a random process, such as tossing a coin. By making allocation to treatment group a random process, those involved in the trial will not be able to predict allocation, and thereby control it.

However, simply using randomisation does not eliminate allocation bias. If those involved in the trial know the randomisation schedule in advance they can select patients with a more favourable prognosis to one treatment group by controlling recruitment into the trial (even though they do not control allocation to treatment group). For example, we could randomise patients by randomly allocating days of the week, so that on some (random) days they receive the intervention and on others they receive control. However, if patients, carers or researchers know which treatment is being provided on that day, then they could choose to participate in the trial only if the treatment they want is being offered.

Allocation concealment

Allocation concealment ensures that a randomised trial will genuinely remove any influence over the allocation process. All those involved in the trial should be unable

to predict the allocation of the next participant in the trial until that participant is irreversibly enrolled in the trial.

The ideal method to achieve allocation concealment is the telephone randomisation hotline. The randomisation sequence is held at a separate location that must be telephoned whenever a patient is recruited. The allocated treatment group is only revealed when all the patient's details have been recorded and they are irreversibly entered into the trial.

Consecutive, sealed, opaque envelopes can also be used to achieve allocation concealment, but all the envelopes must be accounted for at the end of the trial and regular checks must be made for tampering. It is surprising how far people will go to subvert the randomisation process!

Allocation concealment is the key to avoiding bias. Randomisation alone is not sufficient. In fact, if allocation concealment is in place then randomisation schedule does not have to be completely random. Block-randomisation, in which the randomisation sequence if split into blocks with equal (or fixed) numbers of treatments and controls, can be used to ensure equal numbers of treatment and controls. However it is important that the sequence should not be predictable, because this would mean that allocation would no longer be concealed.

Allocation concealment ensures that those involved in the trial are unaware of the allocated group until the patient is irreversibly entered into the trial. Blinding refers to subsequent concealment of the treatment group from those involved in the trial. A triple-blind, placebo-controlled trial will effectively achieve allocation concealment because (if blinding is effective) patients, carers and researchers will be unaware of group allocation throughout the trial.

Blinding

Blinding tackles a different form of bias from allocation concealment. It is concerned with ensuring that the measurement of outcomes is free from bias. If any of the following individuals are aware of the treatment received they may alter their interpretation of the outcomes measured: patients, carers providing the treatment studied, carers providing subsequent care, those measuring outcomes, and those analysing outcomes. "Double blind" does not really cover the issue! If a study is described as blinded you need to identify exactly who was blinded.

The most important person to be blinded is the person measuring the outcomes. If they are aware of treatment group then results will be subject to measurement bias. Blinding of patients and carers helps to combat the placebo effect (the beneficial effect of simply receiving treatment or just attention). Whether patients or carers should be blind depends upon the type of research question. For a pragmatic trial we may simply wish to know whether treatment makes people feel better, so we are not interested in whether it is due to a placebo effect or not. For an explanatory trial we will want to know how and why the treatment is effective, so we will want to eliminate any placebo effect.

In drug trials it may be possible to ensure that everyone concerned is blinded. In trials of surgery and other physical interventions this is clearly important. However, bias

may still be minimised by ensuring that those who can be blinded are blinded. In particular, those responsible for measuring outcomes should be blind, even if carers and patients are not.

Blinding and outcomes
The potential for lack of blinding to lead to bias will depend upon the outcome being measured. "Soft" outcomes, such as patient satisfaction, quality of life, range of movement, or pain, have a strong subjective element. This does not mean that they are not important, but it does make them susceptible to bias if blinding is inadequate. "Hard" outcomes, such as death, are less subject to bias due to lack of blinding.

Intention to treat analysis (analyse as you randomise)
The main analysis should always be done on an intention-to-treat basis, and the overall conclusion should be based on this analysis. Intention-to-treat analysis means that patients are analysed in the group to which they were originally randomised, regardless of whether they actually received the treatment they were allocated to. It ensures that the protection from bias created by allocation concealment is maintained.

If patients are allowed to leave the group to which they were analysed this will introduce bias. Patients who withdraw, do not attend follow up, fail to comply, or have to change treatment because of adverse events will be different from those who complete their treatment as allocated. Therefore all patients should be analysed in the group they were originally allocated to, regardless of the treatment they ultimately received.

Follow-up
Ideally all patients recruited into a trial should be followed-up and outcome data reported. However, this is often difficult, particularly if follow-up is prolonged and patients are mobile. If patients are lost to follow up then the researchers will have to make some sort of assumption (usually implicit) about whether those lost to follow-up are typical of the study population.

It is usually assumed that those who are lost to follow up are essentially similar to those followed up and do not differ between treatment groups. Therefore their loss from analysis can be accepted. If losses to follow-up are considerable (greater than 30%, for example) or differ markedly between treatment groups, this assumption is unlikely to hold, and significant bias is a reasonable possibility.

Outcome measures
There is no perfect outcome measure. "Hard" outcome measures (death, MI) may be resistant to bias from lack of blinding and considered "important", but are often rare and subject to type II (false negative) statistical errors. Clinical outcomes (blood pressure, PFR) may be sensitive to change and possible to record with adequate blinding, but lack relevance to the patient. Patient-centred outcomes (satisfaction, quality of life, pain) are important and relevant, but often subject to bias due to inadequate blinding. Therefore trials should ideally measure a range of different outcomes to address difference objectives.

Measures of effectiveness

Rather than simply report whether a treatment is effective (i.e. is the difference in outcome between the treatment groups statistically significant) the article should report how effective the treatment is and provide a confidence interval for this estimate. This allows the reader to decide whether the treatment effect is clinically important.

The relative risk reduction (RRR) is the difference between the intervention and control groups in the proportion of patients with the outcome (e.g. death) divided by the proportion with the outcome in the control group.

So if 20/100 patients die in the control group and 15/100 patients die in the intervention group the RRR = (0.2 – 0.15) / 0.2 = 0.25 (i.e. 25%)

The absolute risk reduction (ARR) is simply the difference between the intervention and control groups in the proportion of patients with the outcome.

So in the same example the ARR = 0.2 – 0.15 = 0.05 (i.e. 5%)

At a very simplistic level, reporting the RRR makes the treatment sound more impressive than reporting the ARR.

Both measures have their uses, but the ARR may be more useful for decision-making in the individual patient, particularly if it is used to calculate the number needed to treat (NNT).

The NNT is the number of patients who would need to receive the treatment to avoid one negative outcome, such as death. It is calculated as 1 / ARR.

So in the example above the NNT = 1/0.05 = 20. We would need to treat 20 patients to avoid one death.

MODULE 4. EVALUATION OF SERVICE ORGANISATION AND DELIVERY

Emergency physicians should use scientific evidence to decide how to organise emergency services. Triage, staffing changes, educational interventions and short stay facilities are examples of organisational interventions that require robust evaluation. Applying rigorous research methods is challenging, but this should not be used as an excuse for basing organisational decisions upon hunches or anecdote, rather than scientific evidence.

Randomised methods

The advantages of randomisation also apply to evaluations of service organisation and delivery. If patients, carers or researchers can select which service the patients receive in a comparison of two services, then the findings are very likely to be subject to bias. Randomising patients to the service they receive provides powerful protection against bias. However, it is often impossible to provide two services simultaneously to allow individual patients to be randomised to one or the other.

In these circumstances cluster randomisation may be used. Instead of randomising individual patients, groups of patients are randomised by, for example, randomising periods of time (such as days of the week), members of staff, or whole hospitals.

Cluster randomisation has some disadvantages compared to randomisation of individual patients:

1. Allocation concealment is not usually possible. Patients, carers and researcher know which service is being used when patients are asked to participate in the trial. They can therefore subvert randomisation by choosing not to participate in the study if the service they want is not being provided. However, this is often not a problem in emergency care because people are unlikely to be able to choose when and where they have their emergency.
2. Standard statistical tests are not appropriate. Analysis requires specialist statistical tests to take clustering into account. Also, statistical power may be substantially reduced.

Non-randomised methods

These offer a much simpler way of comparing services. Two different services may be compared as they run contemporaneously in two different hospitals. Alternatively, a new service may be compared to the previous service (historical controls). The latter option is very commonly used.

Although simple, these methods carry a high risk of bias. Contemporaneous comparisons will be biased if there are baseline differences in the type of patients who use the two different services. Historical comparisons will be biased by changes in service delivery occurring over time.

One potential solution to these shortcomings is to use both contemporaneous and historical controls when a new service is introduced. The contemporaneous comparison allows control for changes over time, while the historical comparisons allows control for baseline differences between patients using the two services.

The Hawthorne Effect
Studies that simply measure outcomes before and after an intervention, and then conclude that intervention caused the change in outcome may be subject to confounding by the Hawthorne Effect. Based on experiments undertaken at the Hawthorne works of the Western Electric Company in Chicago, this describes the observation that people change their behaviour when they think that you are watching them. Therefore any intervention, if subsequently monitored, will produce a recordable change in processes or outcomes, which is lost when monitoring ceases. The obvious solution, blinding staff and patients to the evaluation, is difficult to achieve.

Sustainability
Changes in service organisation and delivery need to be sustainable. During the period of evaluation it may be possible to provide a service for short period of time in a way that may not be sustainable in the long term. When appraising an evaluation of a change in service organisation it is worth examining what additional resources where required, what staffing arrangements were needed and what knock-on effects the change in organisation could have on other services.

Generalisability
It is particularly important to examine whether findings from service evaluation can be generalised between settings. Service delivery is very dependent upon the setting, staffing, patients and facilities. New services are often developed by enthusiasts who, by their very nature, may be very different individuals or work in a very different environment from those who will have to implement the new service elsewhere.

MODULE 5. EVALUATION OF A DIAGNOSTIC TEST

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 very accurate, e.g. contrast demography for deep vein thrombosis
- A combination of diagnostic tests that used appropriately will reliably rule-in and rule-out disease, e.g. VQ scanning for pulmonary embolus combined with pulmonary angiographies in equivocal cases
- Diagnostic testing with follow-up for negative cases to identify cases of disease that may have initially been misclassified as disease negative

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 involve weighing its' potential accuracy against the feasibility of any alternative approaches.

Independence of the reference standard

The same reference standard should be applied to all patients, regardless of the results of the diagnostic test under evaluation. If the same reference standard cannot be applied to all patients 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 reverence standard is used. For example, in an evaluation of Wells clinical score for pulmonary embolus patients with a high clinical score might receive a reference standard of CT pulmonary angiography while those with a low clinical score receive a reference standard of D-dimer testing. This is known as *work-up bias*.
2. The diagnostic test under evaluation forms part of the reference standard. For example, an evaluation of cardiac markers for diagnosing myocardial infarction will typically use a reference standard that include a cardiac marker result in the definition of myocardial infarction. This is known as *incorporation bias.*

Blinding

The person measuring or interpreting the diagnostic test under evaluation should be blinded to the results of the reference standard. If not 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.

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.

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 method is very prone to bias and over-estimation of sensitivity and specificity.

Interobserver error (reliability)
A diagnostic test or clinical finding is described as being unreliable if it gives different results when used 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 in question, i.e. the gold standard is positive.

Case negative
An individual without the disease in question, i.e. the gold 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.

How are the terms used?
Different terms have different implications for the diagnostic value of a test.

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

This table 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!

	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)

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

Typically in emergency medicine we are using diagnostic tests to rule out diseases that have a low prevalence, e.g. MI or PE in chest pain, fractures in ankle injury, SAH 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 CCU population of chest pain patients who have a prevalence of MI 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 was of presenting diagnostic data and can be applied to individual patients in a way that sensitivity and specificity cannot. 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 not diagnostic.
0.2 to 0.5	Moderately decreases likelihood of disease. Useful additional information, but not rule-out.
5 to 10	Markedly increases likelihood of disease. May be diagnostic 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

MODULE 6. SYSTEMATIC REVIEWS

What is a systematic review?

A systematic review is a scientific study. It follows the IMRD approach (introduction, methods, results, and discussion). The conclusion should represent an unbiased synthesis of available data relating to a specific question. It may not be very entertaining to read but, if undertaken properly, will provide and objective answer based upon the best scientific evidence.

A narrative review is not a scientific study. The authors present their opinions of a particular topic with reference to primary studies they have selected. A good narrative review should be interesting, entertaining or provocative, but it should not be considered to provide scientific evidence.

Systematic review	**Narrative review**
Focussed question	Broad question
Methodology described	No methodology described
Systematic and comprehensive literature search	Based on authors collected papers
Primary studies selected according to defined criteria	Primary studies selected at authors discretion
Quality of primary data assessed objectively according to predefined criteria	Quality of primary data assessed subjectively according to authors opinion
Synthesis of primary data may be attempted using statistical techniques	No formal statistical synthesis of primary data
Potential bias in selection of primary data may be assessed	Potential bias not considered
Conclusions result from a scientific study of the available data	Conclusions represent the authors opinions

Stages of a systematic review

Data collection for a systematic review involves three stages:

1) Literature searching and retrieval
2) Selection of appropriate papers
3) Quality assessment of selected papers

These three steps should each ideally be carried out by two independent assessors who are blind to each other's decisions. The review should report the total number of articles identified by the search, the number selected after scanning titles/abstracts, the number selected after assessment of the full article, and the number included in the review.

Literature searching

An inadequate literature search may miss important articles. A literature search may include:

- Electronic databases: Medline, Embase, Cinahl, Cochrane database etc.
- Hand search of journals
- Grey literature: reports (government or academic), conference proceedings, internet, libraries, professional societies, Kings Fund, Nuffield etc.
- Research registers: National Research Register, HTA database, Cochrane

- Retrieved articles: bibliographies, search authors names, citation threads
- Contact with researchers or "experts"
- Pharmaceutical industry

Publication bias
Publication bias occurs when the results of a study influence the likelihood that it will be written-up, submitted for publication or published, and thus the likelihood that it will be included in a systematic review. Positive studies (i.e. trials reporting s significant effect or diagnostic studies reporting high sensitivity/specificity) are more likely to be written-up, submitted and published, so are more likely to be included in a systematic review. This may lead to an over-estimate of treatment effect or diagnostic accuracy.

Publication bias can be minimised by undertaking a comprehensive search, but the possibility of publication bias can never be completely eliminated. Techniques such as the funnel plot can be used to search for publication bias, but these are often insensitive. Prospective registration of trials may offer the best solution to publication bias in the future.

Selection of retrieved articles for analysis
Literature searches will retrieve large numbers of articles, most of which are irrelevant. A systematic review must therefore define the method by which retrieved articles are selected for inclusion. This should be directly related to the research question. Often the inclusion criteria will relate to the "PICO" of the research question- the defined patients or population, the intervention, the comparison, or the outcome of interest.

The following criteria are sometimes used to exclude studies. They usually reflect convenience to the authors and their application may cause bias.

1) Small studies
2) English language only
3) Mainstream journals only
4) Insufficient data presented
5) Data presented in a form incompatible with planned analysis
6) Year of publication

Assessment of study quality
Ideally, all studies selected for inclusion should be assessed for quality. This will allow the authors to determine the overall quality of the available data and to explore the impact of excluding poor quality studies.

Quality assessment should be objective and based upon criteria that are known to influence study quality. The only factors proven to impair quality in trials are the use of historical controls, lack of blinding, inadequate follow-up, and failure to use intention-to-treat analysis. These factors are combined in a commonly used quality score, the Jadad score.

Heterogeneity
Studies of a similar intervention, using similar methodology, in a similar environment should give similar results. The only differences between results will be due to

random error. Heterogeneity is the term used to describe the amount of variation in the results of trials included in a systematic review.

The usual assumption behind a systematic review is that included studies are measuring the same result. This is particularly important if there is to be any attempt to combine results (meta-analysis). If there is substantial heterogeneity between results then studies may not be measuring the same thing and any conclusions based on assumptions of a common effect will be suspect.

It is therefore important to assess results for heterogeneity of effect. This can be done in several ways.

1. The Forrest Plot: Results of a systematic review are usually presented as a Forrest Plot. Individual study results, with 95% confidence intervals, are plotted alongside each other. Simply observing the overlap of confidence intervals gives a crude estimate of heterogeneity.
2. Statistical tests of heterogeneity: Various statistical methods can test the null hypothesis that all the studies come from the same population and are estimates of the same value. If the test is statistically significant this gives good evidence that studies are heterogeneous. However, a non-significant test does not rule-out potentially important heterogeneity.

Meta-analysis

This is the synthesis of data from various sources to provide an estimate of common effect. Meta-analysis should not consist of simply adding results together or calculating a mean effect. This does not take into account the size or variance of each individual study. Although meta-analysis software is available free on the internet, the involvement of someone with statistical expertise is usually required.

Meta-analysis assumes that all the individual studies are estimates of the same value. Combining results provides a more precise estimate and reduces the chances of a type II (false negative) statistical error (i.e. missing a potentially important treatment effect). This is the principal value of meta-analysis. It does not overcome bias in the original data. Combining biased data (such as the results of historically controlled trials) will just give a precise, but inaccurate, estimate.

Clearly meta-analysis is much more controversial if there is any evidence of heterogeneity of effect. Combining the results of fundamentally different studies simply does not make sense. Clinicians may feel intimidated by fancy statistical tests and discussion of “fixed effects” and “random effects” models. However, clinicians are often well placed to comment on heterogeneity and inappropriate combination of results.

Rather than trying to decipher the stats, have a look at the studies that have been combined. What were the patient inclusion/exclusion criteria? What was the setting? What exactly was the intervention? What was the control? Meta-analysis is sometimes described as the statistical equivalent of combining apples and oranges. However, it may become apparent that the statisticians, for all their fancy tests, are not just trying to combine apples and oranges. They are trying to combine apples, oranges, potatoes and cabbages, with the odd sock thrown in as well.