CRITICAL APPRAISAL FOR EMERGENCY MEDICINE
1. CONCEPTS AND DEFINITIONS

Critical appraisal of a scientific article involves using a number of terms and concepts that are taken from epidemiology and statistics. If we use these terms or concepts without having a clear understanding of their meaning then confusion is likely to ensue. This article will define terms and concepts used in critical appraisal and provide examples of how they apply to analysis of studies in emergency medicine.

What is a hypothesis?

Research articles often describe testing a hypothesis, and critical appraisal will often involve identifying what hypothesis has been tested. A hypothesis is a prediction. Having made a prediction, observation or experimentation is then used to determine whether the prediction is true. A hypothesis should be supported by previous work. In other words, there should be a clear explanation for why we might expect the hypothesis to be true. If a hypothesis seems to have been plucked from thin air, without any supporting rationale, then we should be suspicious that it may have arisen by chance during analysis of the data.

What are 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 obviously little point trying to generalise a finding that is not valid. So validity is usually considered before generalisability. However, many would argue that generalisability is equally important, because a finding that is only valid in one specific setting has as little practical use as a finding that is not valid. In practice, of course, validity and generalisability cannot be judged in simple yes/no terms, but as degrees of validity and 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 to typical patients in routine practice. On the other hand, a multicentre observational study of unselected patients in a routine hospital setting will produce generalisable findings, but validity may be compromised. Example 1 shows how validity and generalisability interact.

What are chance, bias and confounding?

There are broadly 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.

1. Chance (random error)

Random errors 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), which are explained in more detail in the next article in this series. 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). The greater the sample size, the less the overall estimate will be affected by random error, and the smaller the p value and confidence interval.

Random error will determine the precision of the results. The less random error, the more precise the results.

2. Bias (systematic error)

Bias reflects a systematic error in the methods used in the research, such as in the way the study sample was selected or the measurements were made. Unlike a random error, a systematic error will tend to produce results that are consistently wrong in the same direction (i.e. over-estimating or under-estimating the true value). Many forms of bias have been described, such as selection bias, measurement bias, and analysis bias. The important thing is to understand how any bias may occur, how it may affect the results and how it can be minimised rather than being able to name or classify it.

Statistical methods can be used to identify bias and adjust for it, but p values and confidence intervals do not reflect bias. In fact, the presence of important bias may make p values and confidence intervals misleading.

Systematic error (bias) will determine the accuracy of the results. The less systematic error, the more accurate the results.

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. It describes the situation where the apparent association is actually mediated by another unmeasured factor (the confounder). For example, we may observe that

people who attend the emergency department on a Monday are more likely to die than those attending on any other day of the week and conclude that emergency department organisation on Mondays leave a lot to be desired. However, the association between day or week and mortality may be confounded by another factor, such as illness severity. Patients who have waited over the weekend with a deteriorating condition are more likely to attend on a Monday. The association between day of week and mortality is true, but we have erroneously interpreted it as being a direct association, whereas it is actually confounded by illness severity.

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. Accuracy is usually assessed by looking at the methods used in the study and deciding whether these methods may have led to bias. This is demonstrated in example 2.

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 actually does work under normal conditions. Efficacy studies take place earlier in the development of an intervention, using selected patients, expert staff and highly controlled procedures. Effectiveness studies occur after efficacy has been demonstrated and evaluate the intervention in a wide spectrum of patients, using regular staff and routine working conditions.

Pragmatic and explanatory research

When appraising a study it is important to identify what sort of research question is being asked as it affects the method used.

We can only determine whether the methods are appropriate if we know what sort of question is being asked.

Research questions 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 tries to determine how or why a treatment works.

Pragmatic research should use routine staff and settings, unselected populations, research methods that do not interfere with clinical practice, and 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 can measure clinical outcomes, such as peak expiratory flow rate, blood pressure or radiological appearance. The research methods may interfere with clinical care or produce treatment that is highly structured and protocol-driven.

Two apparently similar research questions may require different methods, depending upon whether they are pragmatic or explanatory. Example 3 is a case in point.

Summary

Critical appraisal involves determining whether the findings of a scientific article are valid and generalisable. The three main threats to validity are chance (random error), bias (systematic error) and confounding (error of interpretation). Random error leads to imprecision, whereas bias leads to inaccuracy.

Research can broadly be defined as either explanatory or pragmatic. Explanatory research aims to determine how or why an intervention works (or doesn't work). Pragmatic research aims to determine whether an intervention is useful or not. Different methods are appropriate for explanatory and pragmatic research, so we need to ensure the correct ones were used in the study.

Example 1: Validity and generalisability

We are appraising two articles that both evaluate the performance of D-dimer for diagnosing deep vein thrombosis (DVT). Which is most likely to be valid and which is most likely to be generalisable to a typical emergency department population with suspected DVT?

Professor Clot has measured D-dimer in 500 patients presenting to his specialist vascular laboratory. Every patient had a reference standard test of contrast venography. The prevalence of DVT in the study population was 40%.

Mr Sprain has measured D-dimer in 500 patients presenting to his emergency department. High-risk patients or those with a positive D-dimer had a reference standard of compression ultrasonography, low risk patients with a negative D-dimer had telephone follow-up only. The prevalence of DVT in the study population was 15%.

Professor Clot's study used a rigorous, independent reference standard test, whereas Mr Sprain's used a flawed reference standard that could have missed cases of DVT. Professor Clot's study was therefore more likely to be valid. However, achieving this validity involved selecting a high prevalence population. Thus Professor Clot's study is likely to be less generalisable to the emergency department population than Mr Sprain's study, which appears to have recruited unselected emergency department patients with a relatively low prevalence of DVT.

Example 2: Accuracy and precision

We are appraising two articles that both aimed to measure the length of time that emergency physicians spend in direct patient contact. Which is likely to provide the most accurate estimate and which the most precise?

Dr Meticulous has observed 20 interactions between emergency physicians and patients and has estimated the mean duration of direct contact to be 12 minutes (95% confidence interval 5 to 19 minutes).

Dr Slapdash asked his colleagues to estimate the length of direct contact with each patient they saw over a two-week period. His data from 500 interactions show the estimated mean duration of contact to be 20 minutes (95% confidence interval 19 to 21 minutes).

Dr Slapdash has produced a much more precise estimate. The larger sample size has reduced random sampling error and produced a smaller confidence interval. However, his approach may be subject to bias if emergency physicians tend to over-estimate the time they spend with patients. Dr Meticulous has independently measured contact times by direct observation and may therefore have produced a more accurate (but imprecise) result.

Example 3: Pragmatic and explanatory studies

We are evaluating two studies of non-invasive ventilation for acute cardiogenic pulmonary oedema.

1. A multicentre randomised trial involving a variety of hospitals. All patients who appeared to have acute cardiogenic pulmonary oedema on the basis of routine testing were recruited and randomised. Regular staff provided the treatment according to simple protocols that allowed plenty of scope for physician judgement. Some patients did not receive the treatment they were randomised to but all were analysed as if they had. The outcomes were mortality and quality of life.

This trial addresses a pragmatic question: Does non-invasive ventilation work as a routine treatment for patients presenting with acute cardiogenic pulmonary oedema?

2. A single centre trial undertaken at a specialist hospital with an interest in this acute cardiac disease. Patients were selected if they appeared suitable for non-invasive ventilation. All patients underwent echocardiography to confirm the diagnosis before they were recruited. Specialist trial staff provided the treatment on a one-to-one basis according to strict protocols. The outcomes were physiological measures: change arterial blood gases or oxygen saturation.

This trial addresses an explanatory question: Can non-invasive ventilation improve outcomes for certain patients with acute cardiogenic pulmonary oedema?

Although the trials each used very different methods, they both used methods that were appropriate to the question they were addressing.