CRITICAL APPRAISAL FOR EMERGENCY MEDICINE TRAINEES
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 evaluation. Applying rigorous research methods to address these issues is challenging and it may be difficult to find robust data. Nevertheless this should not be used as an excuse for basing organisational decisions upon hunches or anecdote, rather than scientific evidence.

Appraisal of studies evaluating changes to service delivery have not traditionally been covered in great detail in texts on evidence-based medicine. However, their importance in emergency care means that it is worthwhile for us to be familiar with the key issues. If we don't consider these studies as a separate group then there is a risk that we may either attempt to inappropriately apply appraisal methods used to assess clinical trials, or accept at face value claims made on the basis of very weak methods, such as simple before-and-after intervention comparisons.

Randomised methods

The advantages of randomisation described in the previous article in this series also apply to evaluations of service organisation and delivery, although the practicalities of using randomisation are much more challenging. 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 receive one service or anther provides powerful protection against bias. However, this requires us to provide two services simultaneously, which is often not feasible. Furthermore there are some interventions, such as triage methods, that are inevitably applied to groups of patients rather than individuals.

In these circumstances cluster randomisation may be used. Groups of patients are randomised instead of individual patients. For example, periods of time (such as days of the week), members of staff, or whole hospitals may be randomised to one service or another, with patients being randomised in groups according to time of attendance, treating clinician or hospital attended.

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. This may not be a problem in emergency care because people are unlikely to be able to choose when and where they have their emergency, although they can choose whether or not to enter an evaluation.
2. Standard statistical tests are not appropriate. Analysis requires specialist statistical tests to take clustering into account and statistical power may be substantially reduced.

"Unit of analysis error" is a common statistical flaw in cluster-randomised trials. Any study that randomises groups of patients instead of individual patients needs to take potential clustering into account. Clustering is the phenomenon whereby patients in the same group (or cluster) are more likely to share the same characteristics than patients in different groups. The practical implication of this is that statistical analysis needs to adjust for potential clustering of data. Standard statistical tests may underestimate the variance in outcome measures, leading to underestimates of the p value and over-estimate the confidence interval. This is may have important consequences for the interpretation of results. As with many complex statistical issues, the role of the non-statistician is to recognise the potential for error and seek expert statistical advice (or treat the conclusions with caution until reassurance has be sought).

It is also worth remembering that non-randomised studies may be subject to unit of analysis error. Any study that allocates patients in groups, rather than as individual patients, may be subject to clustering and should use appropriate statistical techniques.

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 in the same hospital (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 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, and influenced by the Hawthorne effect (see below). These issues are shown in example 2, along with some other potential limitations.

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 comparison allows control for 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.

Knock-on effects
Changes to the organisation and delivery of services can have unintended knock-on consequences. Interventions that reduce demand upon one service may increase demand elsewhere. Emergency physicians will be very familiar with well-intentioned changes to other parts of the health service that have then had important consequences for the emergency department.

Evaluations of service delivery and organisation will inevitably tend to primarily examine the most directly relevant processes and outcomes. Appraisal should therefore involve considering what the potential knock-on consequences for other services might have been.

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 have different approaches or work in a very different environment from those who will have to implement the new service elsewhere.

We should always question the generalisability of findings from a study of organisational change undertaken in only one centre. Almost by definition, a centre that pioneers a new service is likely to be atypical. Data from a single-centre study may provide evidence of a promising new development, but confirmation across a variety of settings in a multicentre study should be required before widespread implementation is recommended.

Summary
The organisation and delivery of emergency care should be guided by scientific evidence, although studies in this field will inevitably not be able to achieve the tight experimental control typical of clinical trials. Critical appraisal therefore involves thoughtful consideration of potential limitations rather than dogmatic application of appraisal checklists.

Example 1

A study of a new triage system randomised 20 consecutive days to either the new or the old system. Waiting time data were collected from all 4895 patients who attended during this time period. Analysis using a t-test showed that the mean waiting times for patients attending when the new system was in operation were shorter than when the old system was running (p=0.0395). The authors concluded that the new system significantly reduced mean waiting times.

This conclusion is probably wrong. Waiting time data are likely to be clustered. Patients attending on the same day are more likely to have a similar waiting time than patients attending on different days. Using a standard statistical test in this case (such as the t-test) will over-estimate the p value and under-estimate the confidence interval. If an appropriate analysis is undertaken and clustering taken into account it is very likely that the p value will exceed 0.05.

(If any smarty-pants thought that the use of a parametric test on skewed data was the problem here, then go and read the second article in this series, or read about the Central Limit Theorem.)

Example 2

A hospital reports that a new way of organising emergency department staff has significantly reduced waiting times. It involves appointing a Team Leader on every shift, who gets to wear a special hat, whilst every other member of staff is allocated a unique role, indicated by a prominent badge and armband. Waiting times were measured for two weeks after implementation of the new system and compared to waiting times in the two previous weeks. The results are so impressive that the Department of Health would like to see the system adopted across the country.

This study has a number of important flaws. As a before and after comparison it cannot control for concurrent changes over time, such as additional staff being employed. Staff behaviour may have been influenced by a Hawthorne effect, especially if they knew that the new system was being evaluated. The reduction in waiting times may have been achieved by staff enthusiasm that could prove difficult to sustain. Finally, the effect of the intervention may be difficult to generalise to other departments (particularly where staff are not easily impressed by headwear).