Evaluation of studies of treatment or prevention interventions. Part 2: applying the results of studies to your patients

In the previous article in this users’ guide series, we began to look at how a critical appraisal checklist could be used to help to decide whether a piece of research is sufficiently valid for its results to be applied to patients. This article continues the appraisal of the same study but focuses on its results to answer the questions:

What were the results?
- How large was the treatment effect?
- How precise is the estimate of treatment effect?

Will the results help me in caring for my patients?
- Are my patients so different from those in the study that the results don’t apply?
- Is the treatment feasible in our setting?
- Were all clinically important outcomes (harms as well as benefits) considered?

Review of the clinical scenario
You are a diabetes specialist nurse who, along with podiatrist colleagues, runs a foot care clinic for people with diabetes. A patient presents at the clinic with a full thickness plantar foot ulcer without any sign of arterial disease. The patient is enthusiastic to try an artificial skin replacement as she has read about the potential benefits. You are unfamiliar with this type of wound covering, and your search for the best available evidence identified no systematic review and one randomised controlled trial (RCT). You are now getting to grips with this RCT before the patient’s next visit.

What were the results?
The aim of this part of the appraisal is to help the reader to judge whether the results of an individual study are important. This decision takes into account the size of the treatment effect and whether the estimate of the treatment effect is precise.

How large was the treatment effect?
The effects of individual treatments are measured using one or more outcome measures. Previous EBN notebooks have described how outcome measures can be dichotomous (eg, yes or no, dead or alive, healed or not healed) or continuous (eg, length of stay, daily intake of fruits and vegetables), and how these measures are presented and analysed. By way of brief review, we can look again at the results of a trial of a nurse led, structured discharge package given to children with asthma on leaving hospital. At 6 months follow up, 15% of children in the intervention group had been readmitted to hospital (experimental event rate or EER) compared with 38% in the control group (control event rate or CER). Although the accompanying p value of 0.001 tells us that the difference between groups was statistically significant, the information it provides is of limited usefulness. There are, however, alternative ways of expressing the same data. The relative risk reduction (RRR) is the proportional reduction in rates of bad outcomes between experimental and control participants in a trial and is calculated as (CER − EER)/CER = (38 − 15)/38 = 23/38 = 0.60, meaning a 60% reduction in the relative risk of hospital readmission. The absolute difference in risk tells us how much of a difference is a result of the intervention itself. A third approach to presenting the same data is to report the number needed to treat (NNT). This gives the reader an impression of the effectiveness of the intervention by describing the number of people who must be treated with the given intervention in order to prevent 1 additional bad outcome (or to promote 1 additional good outcome). The NNT is simply calculated as the inverse of the ARR, rounded up to the nearest whole number; in the case of the asthma trial 1/23 = 5 (95% CI 3 to 12). Put into words, this means that 1 additional hospital readmission within 6 months of discharge would be prevented for every 5 children who receive the nurse led, structured discharge package, and we have 95% confidence that the true NNT value may be as low as 3 and as high as 12. When properly presented, reports of NNTs should incorporate a description of the follow up time, and also the 95% CI around the NNT estimate. The next issue of Evidence-Based Nursing will include a more detailed discussion of using NNTs in clinical practice.

When reading reports of statistically significant differences in treatment effects, it is always important to ask oneself whether the difference is clinically important. It is quite possible for a statistically significant difference to be unimportant, either because the outcome measure is unimportant or because the difference is too small to be noticed by the patient or to warrant a change in practice. For example, a systematic review of antibiotics for sore throat concluded that antibiotics shortened symptom duration by approximately 8 hours, which is probably clinically insignificant when compared with the problems of overuse of antibiotics.

Many published RCTs do not find a statistically significant difference between 2 treatments. These trials are just as informative as those with significant differences, if the studies were large enough to detect a significant difference if one existed. A review of 2000 trials of treatments for schizophrenia reported that the average number of participants in a schizophrenia trial was 65. The authors estimated that only 3% of these studies were large enough to detect a 20% improvement in mental state between groups (for which 150 patients in each arm of a trial would be needed). The true effect of a treatment can never really be known. Instead, we use the results of trials, which are estimates of effect. Each estimate is a neighbour of the true treatment effect—the
crux is the size of the neighbourhood! Confidence intervals (CIs) (often called confidence limits) are a statistical device used to communicate the magnitude of the uncertainty surrounding the size of a treatment effect; in other words, they represent the size of the neighbourhood. The 95% CI represents the range within which we are 95% certain the true value lies. If this range is wide, our estimate lacks precision, and we are unsure of the true treatment effect. Alternatively, if the range is narrow, precision is high, and we can be much more confident. The sample size used in a trial is an important determinant of the precision of the result; precision increases with larger sample sizes, and thereby reduces the width of the 95% CI. Small studies are likely to produce results with wide CIs.

Remember that if the 95% CI of an odds ratio or a relative risk includes 1, there is no statistically significant difference between treatments. Conversely if the CI of a risk or mean difference includes zero, the difference is not statistically significant. Readers of RCTs can look at the lower limit of the CI around an odds ratio or relative risk and, using that as the smallest possible effect size, ask if the effect of the intervention was as small as this; would it be worth using? If the outcome measures used in a study are continuous, readers can use the same approach, looking carefully at the CI for the estimate of the difference (often a difference in means), and judging whether the smallest difference (the lower end of the CI) would be clinically important.

**WILL THE RESULTS HELP ME IN CARING FOR MY PATIENTS? Are my patients so different from those in the study that the results don’t apply?**

In considering whether you can use the findings with your patients, look at the characteristics of the patients in the study and how similar they are (or are not) to your own. It makes most sense to look for compelling reasons as to why the results should not be applied, rather than looking for evidence that the study patients are almost exactly the same as yours. Clinical applicability is one of the main concepts addressed in the commentaries that accompany the abstracts in Evidence-Based Nursing.

**Is the treatment feasible in our setting?**

This is a judgment that depends on factors such as the cost of the intervention (and whether your healthcare system is prepared to pay for it), the skills and training required to deliver the intervention, and the cost and availability of special equipment.

**Were all clinically important outcomes (harm as well as benefits) considered?**

It is common for researchers to use various outcome measures to capture different elements of study participants’ responses to treatment. Typically these might include measures of quality of life and economics as well as direct measures of the ill health treated or prevented. The most important issue for readers of RCTs is that they should reassure themselves that the outcomes reported are likely to be important to the patients or communities targeted by the intervention. It is also important that indirect measures of outcomes are validated alternatives that have been shown to be directly related to the outcome of interest. Proxy, or surrogate, outcome measures are sometimes used by researchers for good reasons. For example, accurate self reports of smoking behaviour are notoriously difficult to obtain; however, salivary cotinine concentration has been shown to be a valid and reliable alternative because it relates directly to smoking behaviour.

Adverse events or side effects experienced by the trial participants should be clearly detailed in reports of RCTs; however, because such events are relatively rare and trials are usually quite small, larger observational studies are better suited to collecting this type of data.

Increasingly, health systems are placing great importance on the measurement of the cost effectiveness of interventions. Readers might therefore look for information relating to cost, and possibly cost effectiveness in a trial report. A future users’ guide will address how to critically appraise economic evaluations.

**Resolution of the scenario**

Returning to the study by Naughton et al on artificial skin, we see that the effect of the new dressing was measured in terms of the number of ulcers completely healed after 12 weeks of treatment. This outcome is highly objective, requires no complex measurement procedure, and is likely to be an outcome that matters to patients. The authors of this RCT did not report other important outcomes such as quality of life (2 treatments may have a differential effect on this), costs, or ulcer recurrence.

39% of patients who received the artificial skin dressing had healed ulcers at 12 weeks compared with 32% of patients who received traditional dressings. This difference was not statistically significant (p = 0.138). The authors then described how at an early point in the research they discovered that only 60% (76 of 126) of patients in the experimental group had received pieces of artificial skin that were “active”; 49% of the patients who received active artificial skin on at least their first treatment (87 of 76) had healed ulcers by 12 weeks compared with 32% of patients in the control group. This difference was statistically significant (p = 0.008). This result, however, should be treated with caution as although this subgroup analysis was planned at an early stage of the study, it is the opposite of intention to treat analysis, and subverts the randomisation (because a large proportion of patients were discarded from one of the groups).

You are not prepared to use this treatment on the basis of this subgroup analysis, although the result, if true, would equate to an ARR of 49% − 32% = 17%, and an NNT over 12 weeks follow up of 1/17 = 6 (95% CI 3 to 32). Instead, you describe to your patient the shortcomings of the current evidence and vow to watch for further evaluations of this new treatment.

**NICKY CULLUM, RN, PhD**

Centre for Evidence Based Nursing

Department of Health Studies

University of York

York, UK


