Critical evaluation

After completing this tutorial, you will be able to:

- Assess the validity and usefulness of clinical trials.
- Describe the meaning of some common terms used to quantify the benefits and harms of a medicine.
- Outline how evidence from clinical trials is evaluated and applied by NHS prescribing committees.

Why this subject matters…

As a new hospital pharmacist you probably won't be undertaking detailed critical evaluation of trials or systemic reviews every day. However you will be weighing up evidence from all sorts of different sources such as the BNF, journals, national guidelines or online discussion forums to help you to make decisions. You need the best quality evidence to help you care for individual patients, update your practice, or write a guideline. So you must able to assess the validity and usefulness of different types of data with confidence.

Introduction

What do you think we mean by the term critical evaluation (sometimes known as critical appraisal)? One definition is that it is a process of carefully and systematically examining research to judge whether the data are **valid** and **useful**.

We assess **validity** by considering how robust the data are: is the clinical trial scientifically sound?

We can evaluate **usefulness** by thinking about what the results mean for our patients in the real world.

Most of this tutorial is devoted to examining validity and usefulness in more detail.

Endorsed by
ROYAL PHARMACEUTICAL SOCIETY

[Last updated: 24 December 2018]

Validity and bias

To evaluate validity we need to consider whether the research was done properly. All possible measures should be used to reduce the risk of **bias** in a clinical trial. The methods section of a paper should explain exactly what steps have been taken and the results should be fully and unambiguously reported. The CONSORT (CONsolidated Standards of Reporting Trials) 2010 checklist is a useful tool for assessing the validity of a trial.

For simplicity, the following points describe a trial of placebo (control) versus drug treatment, but they also apply to trials that compare drug treatments (e.g. drug A versus drug B).

- The number of participants **(sample size)** needs to be planned carefully. There are usually restrictions on numbers because of ethical, cost and time considerations. However the trial should be large enough to be adequately 'powered'. The **power** of a trial is the likelihood that it will detect a difference between 2 groups when one genuinely exists. The ideal power for a trial is at least 80%, so that if the trial was repeated 100 times a statistically significant treatment effect would be seen in 80 of them. The power of a study increases with sample size. Ideally authors should describe their power calculation. This should specify the primary endpoint (the main result being measured at the end of the trial) and take into account the expected outcome. If the difference between treatment and placebo is expected to be small, a larger sample size will be needed compared to a study looking for a bigger difference.
  The calculation will also need include an allowance for participants **dropping out**. For example: "To detect a reduction in hospital stay of 3 days with a power of 80%, we calculated a sample size of 75 patients per group was needed, given an anticipated dropout rate of 10%".

  Trials that are underpowered or have a small sample size don't have to be discounted, as they can still be useful, but larger trials or a meta-analysis may be needed to be more confident in the results. We also need to watch out for trials that are 'overpowered'. If the sample size is too big, small effects of little clinical importance can be reported as statistically significant (see p-values below).

- **Selection** of volunteers from a patient population should be **random**. This stops the researcher from choosing their preferred patient population, so affecting the outcome of their trial favourably. For example, a researcher may approach every third patient that comes to a clinic to ask them participate rather, than choosing who they think will do well.

Courtesy of Simon Wills

- **Allocation** of volunteers to placebo and treatment groups should be **concealed** from the researchers. This is a different concept to 'blinding'. It is recommended for all trials, including unblinded (open-label) trials. It prevents selection bias by ensuring the researchers do not influence which patients get the study treatment. For example, the early studies of diphtheria vaccine showed that more patients in the vaccine group died compared to placebo. This was because the sickest patients were chosen to receive the vaccine and the healthier patients were given a placebo. The best way of ensuring allocation concealment is to use a **centralised** service, where randomisation is carried out independently at a site away from the trial location (e.g. hospital pharmacy).

- Ideally as many people as possible involved in the trial should be **'blind'** (or masked) to whether volunteers are receiving placebo or treatment. The opposite is **'open-label'**, when everyone knows what the volunteer is receiving. **'Double-blind'** usually means the investigators and the volunteers do not know which arm of the study each volunteer is in, and **'triple-blind'**means the committee monitoring the data also do not know.

  However blinding is not always possible – for example with drugs that cause distinct side effects (e.g. peppermint oil capsules cause rectal burning) or if the treatment has a complicated dosage regime (e.g. warfarin dosed according to INR results). One way around this is to use something called a **'PROBE'** design: prospective, randomised, open-label, blinded endpoint evaluation where the people doing the evaluation of the endpoints do not know which group the volunteers have been assigned to.

Courtesy of the conversation.com

- The **baseline characteristics** of the groups under study should be as similar as possible. This helps to ensure that any effect seen in the treatment group is due to the treatment and not to pre-existing differences between the groups. The demographics of the groups should be described in the paper. If the baseline characteristics of the groups are very similar this can also be used as an indicator that allocation to groups was truly random.

- Apart from the treatment or placebo, patients should be **treated identically** during the trial; they should receive the same number of blood tests, X-rays, and clinic appointments.

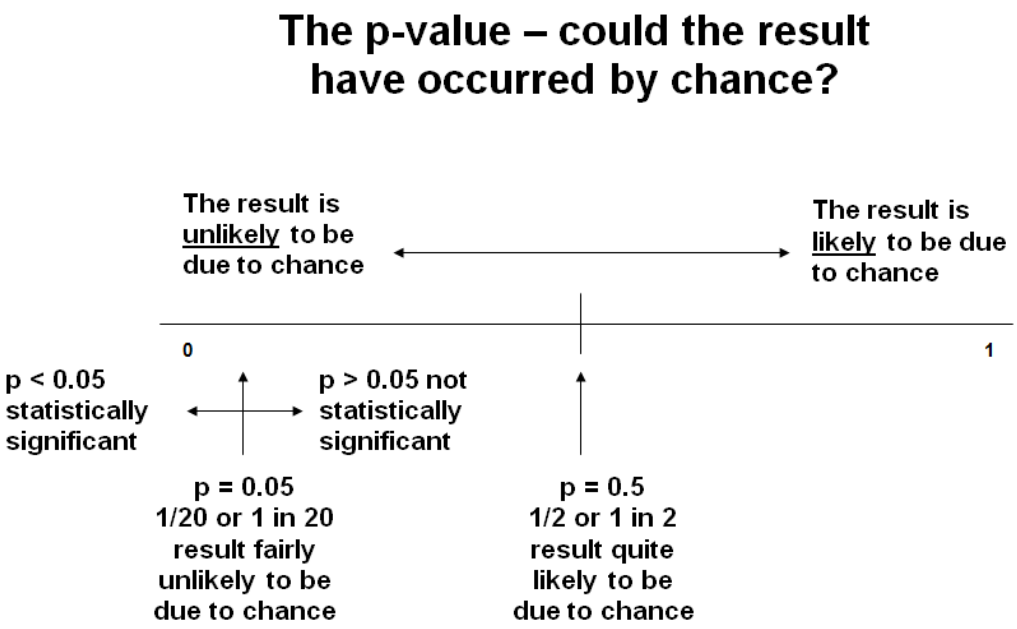
- **Participant flow** should be clearly reported, showing whether and why volunteers did not receive the treatment allocated, or were lost to follow-up or excluded after the trial had started. If this leads to imbalances between the groups it is known as attrition bias. It is important to know which and how many trial participants were included in the final analysis. If only those available for follow-up are included it is known as 'on-treatment' or 'per protocol' analysis. **'Intention-to-treat'** analysis includes all participants who underwent randomisation in their originally allocated groups, no matter what happened during the trial. This is generally favoured because it reduces bias and is more like real life, where people change their minds, or change or stop treatments.

Validity and chance

In critically evaluating a paper you also need to ask whether the results of a trial are valid or whether they occurred by chance. Statistical tests are used to assess this. The most commonly encountered terms are p-values and confidence intervals.

A **p- value** is the probability that a difference will be seen between two interventions in a trial when, in fact, there is no actual difference between the two interventions. In other words, it's an indication of whether the result occurred by chance.

Probability is measured on a scale of 0 to 1 where an impossible event is given 0 and an event that is certain to happen is given 1. In drug trials, by convention, $p<0.05$ is regarded as being statistically significant. It means that there is a less than 1 in 20 chance that you have observed a difference between your study drug and placebo when there is no actual difference between them.



Adapted from original courtesy of The Critical Appraisal Skills Programme (CASP) www.casp-uk.net

When evaluating trial data, it's important not to rely solely on P-values, but to consider whether the results are important. For instance, a trial might show that an antihypertensive drug improved blood pressure readings in 2 people per year, but would this be clinically important, even if it was statistically significant?

p-values are easily misinterpreted, and can be overtrusted and misused. The threshold of 0.05 to claim statistical significance is questionable, and many experts would advocate use of a lower threshold, e.g. 0.005.

It's also important to realise that p-values depend on the sample size and don't consider the size of an effect or its clinical relevance. So the effect maybe small and clinically unimportant, the p-value can still be "significant" if the sample size is large. On the other hand, an effect can be large, but fail to meet the $p<0.05$ criterion if the sample size is small.

We also need to consider that p-values are based only on data from a sample of people, and the results you get for that sample may not be the results you would get with a different sample.

Because of these limitations we should look at other statistical values.

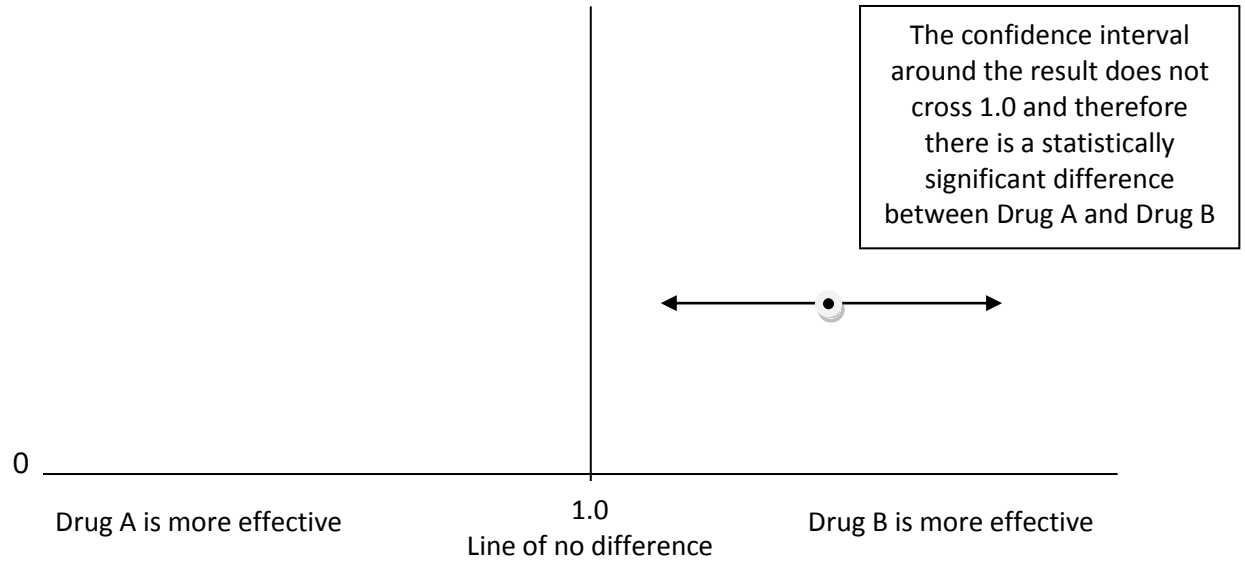
Confidence intervals can give us a measure of the certainty of a result. They are used to describe how sure we are the result we have obtained from studying our small sample of research participants would still hold true if we were able to study the whole population. They are expressed as a range of possible results, within which we expect the actual result to lie - the narrower the range, the more reliable the results.

By convention, 95% confidence intervals are normally used in drug trials, but you may also encounter 90 or 99%. A confidence interval at 95% means that you can be 95% sure that the true result lies within the range quoted, or, expressed another way, that there is a 1 in 20 chance (i.e. 5%) that the true value lies outside the range quoted.

Confidence intervals also show if the difference between interventions is statistically significant or not. When dealing with results which are expressed as ratios (e.g. relative risk, hazard ratio, odds ratio), if the confidence intervals do not contain 1.0 then the result is statistically significant.

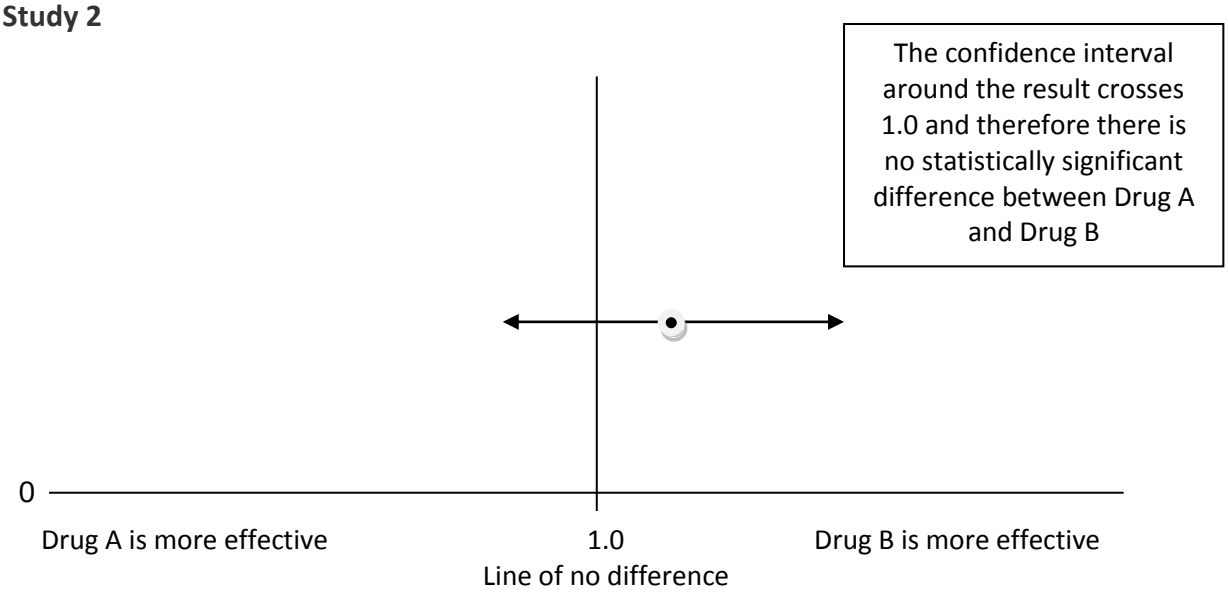
For example consider the following results of 2 studies comparing Drug A and Drug B in reducing the risk of stroke. In the first study, the odds ratio is reported as 1.25 (95% confidence interval 1.05 to 1.45) in favour of drug A.

Study 1



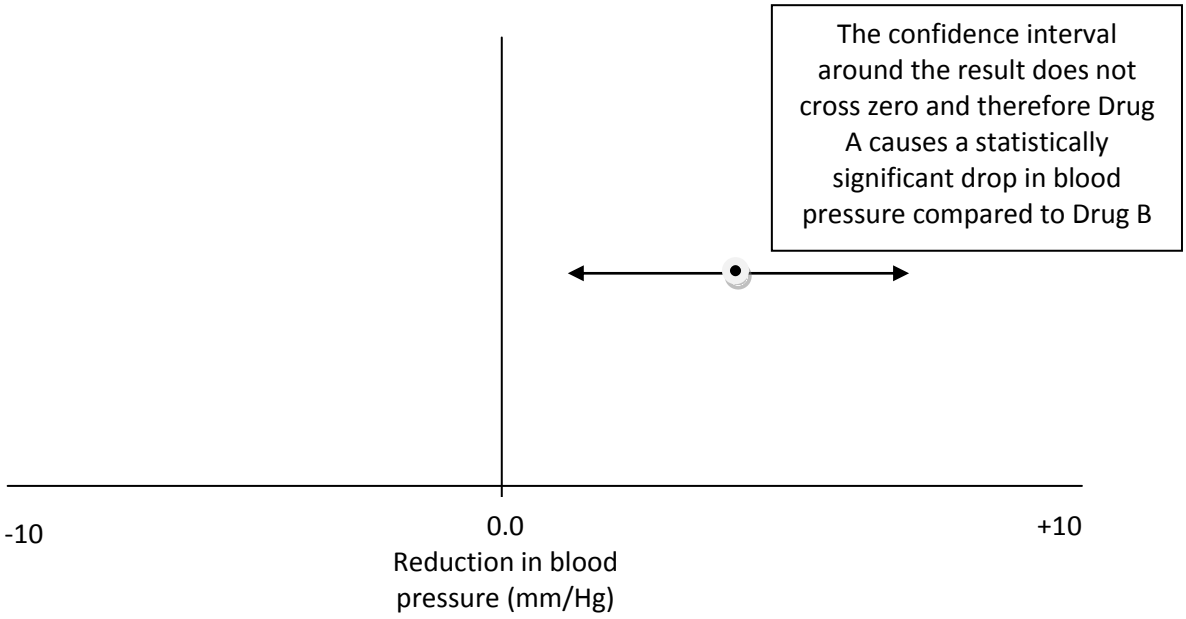
In the second study the odds ratio is reported as 1.10 (95% confidence interval 0.90 to1.30).

Study 2



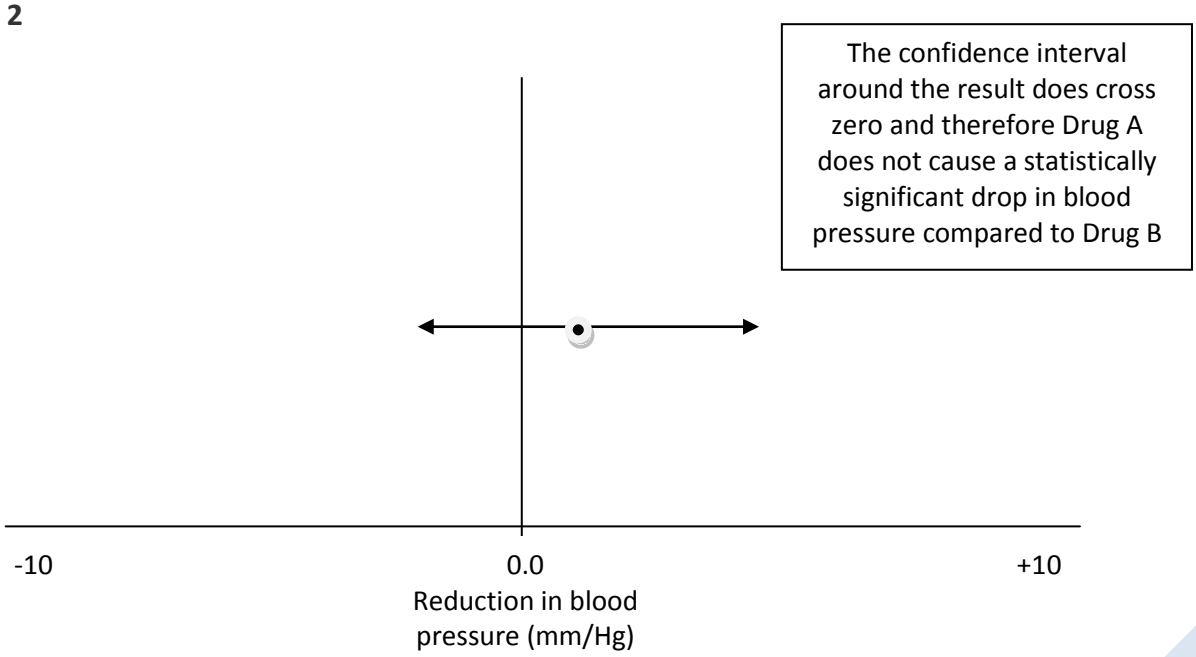
If you have a result not expressed as a ratio such as an absolute difference in blood pressure, then if the confidence intervals do not contain zero the result is statistically significant. For example consider the following results of 2 studies investigating Drug A for hypertension versus Drug B. In study 1 Drug A produced a mean drop in blood pressure of 5mmHg (95% confidence interval +1 to +7mmHg) more than Drug B.

Study 1

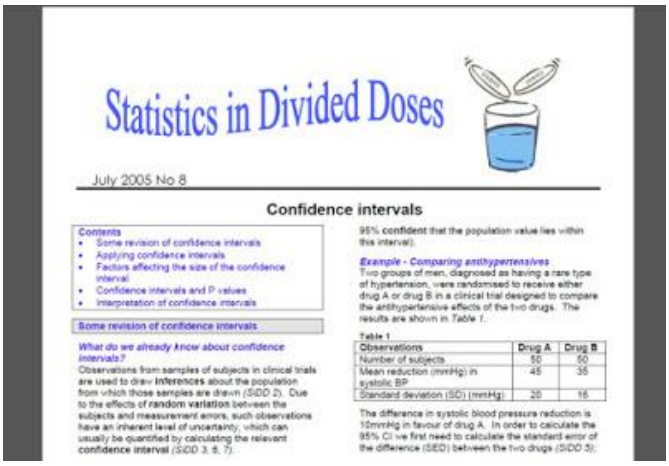


In study 2 of Drug A and Drug B, Drug A caused a mean drop in blood pressure of 1mmHg (95% confidence interval -2 to +4 mmHg) compared to Drug B.

Study 2



Information on how to calculate confidence intervals can be found in the bulletin 'Statistics in Divided Doses' number 3 and number 8.


Statistics in Divided Doses

July 2005 No 8

Confidence intervals

Contents
- Some revision of confidence intervals
- Applying confidence intervals
- Factors affecting the size of the confidence interval
- Confidence intervals and P values
- Interpretation of confidence intervals

Some revision of confidence intervals

What do we already know about confidence intervals?

Observations from samples of subjects in clinical trials are used to draw inferences about the population from which those samples are drawn (SIDD 2). Due to the effects of random variation between the subjects and measurement errors, such observations have an inherent level of uncertainty, which can usually be quantified by calculating the relevant confidence interval (SIDD 3, 6, 7).

95% confident that the population value lies within this interval).

Example - Comparing antihypertensives

Two groups of men, diagnosed as having a rare type of hypertension, were randomised to receive either drug A or drug B in a clinical trial designed to compare the antihypertensive effects of the two drugs. The results are shown in Table 1.

Table 1

Observations	Drug A	Drug B
Number of subjects	50	50
Mean reduction (mmHg) in systolic BP	45	35
Standard deviation (SD) (mmHg)	20	16

The difference in systolic blood pressure reduction is 10mmHg in favour of drug A. In order to calculate the 95% CI we first need to calculate the standard error of the difference (SED) between the two drugs (SIDD 5).

Usefulness and risk

In assessing the usefulness of a trial there are two important considerations: are the results of the trial clinically important and what is the size of the benefit (and any harm)?

It is essential to judge the **clinical importance** of a result. We can do this by checking if the outcome measured was disease-oriented or patient-oriented. For example, if a new osteoporosis drug increases bone mineral density we would call this a **disease-oriented outcome** (DOO). If the drug reduces the risk of fractures we call this a **patient-oriented outcome** (POO). So an increase in bone mineral density of 1%, for example, may be statistically significant but if the rate of fractures is not improved is this clinically important?

To quantify the size of the benefit and harms of a new treatment, some of the key terms you'll see are:

- Absolute risk reduction (ARR)
- Relative risk (RR) and hazard ratio (HR)
- Relative risk reduction (RRR)
- Number needed to treat (NNT) and Number needed to harm (NNH)

We're going to use an example to help you understand what these terms mean and how to calculate them. Consider a fictitious randomised, double-blind trial of 'anotheraban' versus placebo and the prevention of stroke over 2 years. Each group contains 2,000 volunteers. At the end of the study, the number of patients suffering a stroke in the anotheraban group is 120 and the number in the placebo group is 160.

The **absolute risk** (AR) of an event is simply the chance it will happen. To work out the absolute risk of a stroke in each group, divide the number of strokes by the number of volunteers:

AR placebo group = 160 ÷ 2000 = 0.08 or 8%

AR anotherban group = 120 ÷ 2000 = 0.06 or 6%

The **absolute risk reduction** (ARR) is the difference in risk of stroke between the two groups:

ARR = 0.08 – 0.06 = 0.02 or 2%

This means that anotheraban reduces the absolute risk of a stroke by 2%.

Relative risk (RR) tells us how many times more or less likely an event will occur in the treatment group relative to the placebo group. It is the risk of the outcome in the treatment group divided by the risk in the placebo group. From the above, AR anotheraban is 0.06 and AR placebo is 0.08, so:

RR = 0.06 ÷ 0.08 = 0.75

As this result is less than 1.0, anotheraban has made the risk of stroke less likely compared to placebo.

Some studies may use the term **hazard ratio** (HR) instead of relative risk – the two terms are broadly equivalent but hazard ratios are useful when the risk is not constant over time. It is weighted for the number of patients in the trial at different time points.

The **relative risk reduction** (RRR) is an alternative way of expressing the difference in risk; it is the reduction in risk of an event in the treatment group relative to the risk in the placebo group. Looking at the figures above you can see that the risk in the placebo group is 8% and the risk in the anotheraban group is 6%. Anotheraban has reduced the risk of a stroke by a quarter, from 8% to 6%. This is the relative risk reduction. Mathematically it is calculated like this:

RRR = (AR placebo group – AR anotherban group) ÷ (AR placebo group)

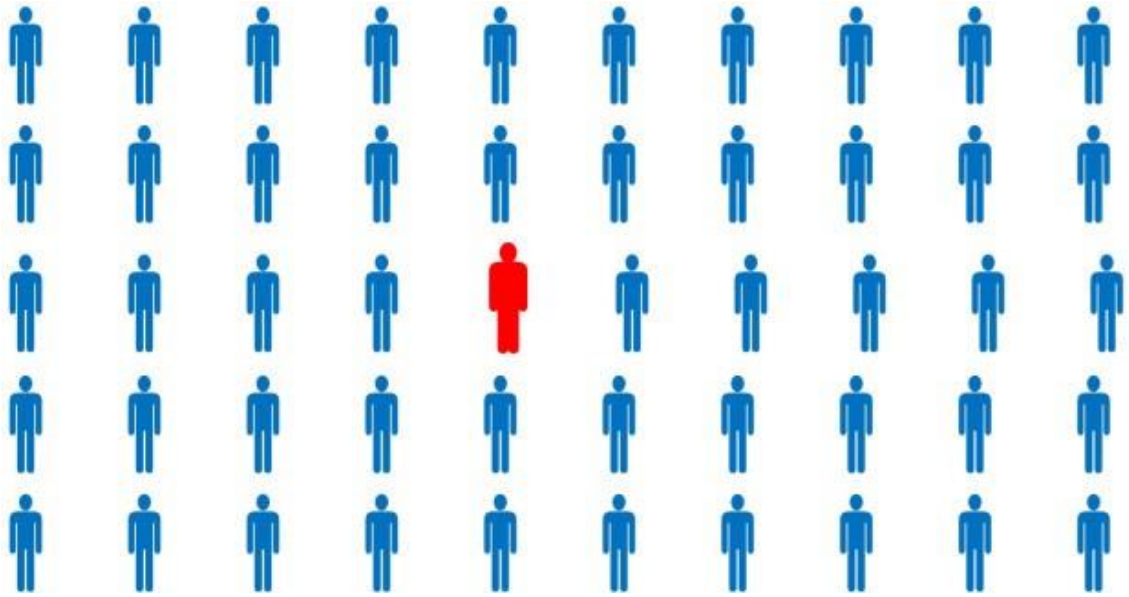
RRR = (0.08 – 0.06) ÷ (0.08) = 0.25 or 25%

Usefulness and NNT/NNH

The **number needed to treat** (NNT) is an expression that literally describes the number of patients that we would need to treat with anotheraban for 2 years to prevent one stroke. It is calculated as shown below, remembering that if you've been using percentages throughout the calculation then use 100 as the numerator to make the maths work. NNTs are normally rounded up to whole numbers:

NNT = 1 ÷ ARR = 1 ÷ 0.02 or 100% ÷ 2% = 50

This means that 50 patients need to be treated with anotheraban for 2 years to prevent one stroke.



So what does this mean in terms of usefulness? In general the smaller the NNT the better the treatment, with 1 being the ideal NNT. But this doesn't mean we should necessarily reject a drug with a high NNT. There are no rules about what an acceptable maximum NNT would be, and it will depend on several factors, such as the severity of the condition being treated, costs, side effects and individual values and preferences. We also need to realise that NNTs are open to interpretation. So some doctors may see an NNT of 50 over 2 years as a considerable benefit, whereas others may see the benefit as small.

Comparing NNTs can help us choose between drug treatments. So, if anotheraban has an NNT of 50 while yetanotheraban has an NNT of 35, we might choose yetanotheraban as it seems more effective.

However it is important to balance effectiveness or benefits of treatments against safety or potential harms. For this we can look at the **number needed to harm** (NNH).

In the same trial, 2 patients in the placebo group and 82 in the anotheraban suffer from life-threatening bleeding. We can describe how many patients we would need to treat for one to suffer from major bleeding (harm) using the NNH. We first need to work out the **absolute risk increase** (ARI) of major bleeding:

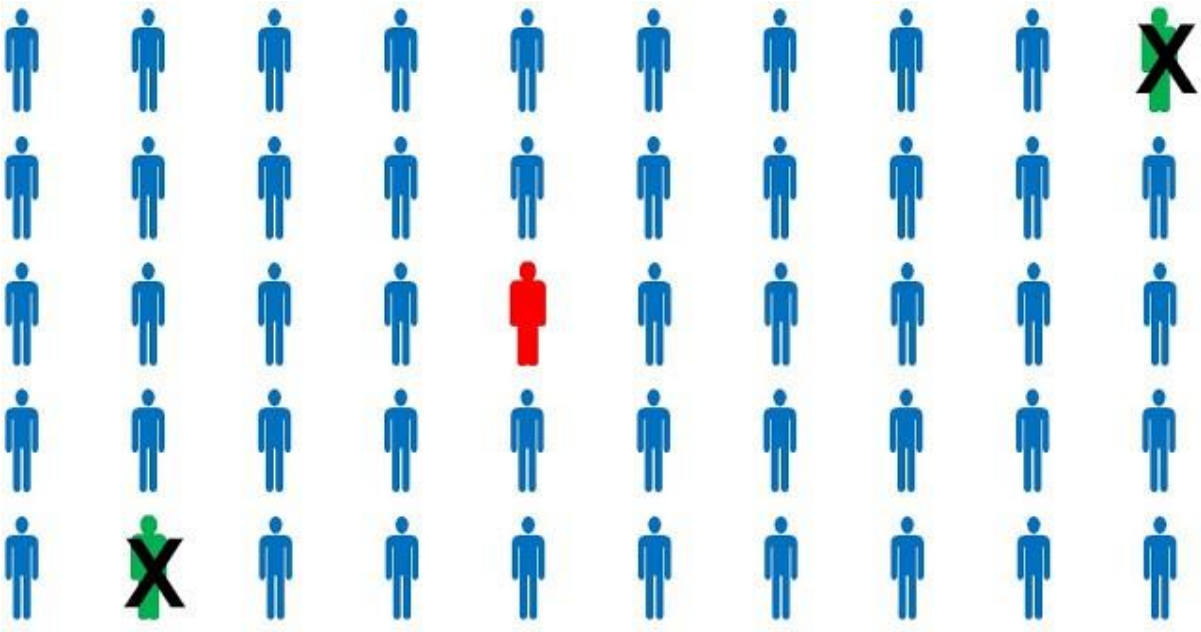
ARI = AR anotheraban group – AR placebo group

ARI= (82 ÷ 2000) – (2 ÷ 2000) = 0.04 or 4%

Then the calculation is similar to that used for NNTs:

NNH = 1 ÷ ARI = 1 ÷ 0.04 or 100% ÷ 4% = 25

NNHs are normally rounded down. A higher NNH is generally more favourable, but as with NNT, there are no set rules about what an acceptable value might be.



Commissioners, or the people buying a service in which anotheraban was to be used, would need to consider these calculations carefully. Is it worth treating 50 patients for 2 years to prevent one stroke when for every 25 patients treated, one of them will have a major bleed that might kill them?

Making the decision

So once we've assessed the validity and usefulness of a clinical trial paper, how do we then put that knowledge into practice and make a decision about the use of the medicine? An effective, agreed process is needed to review the drug, taking into consideration more than just cost. Most areas of the UK are covered by Area Prescribing Committees (APCs) or Medicines Management Committees (MMCs), and their remit includes balancing the potential benefits and harms of different medicines.

One way of deciding between different therapies is known by the acronym 'STEPS':

Safety: What are the risks for patients? Look at adverse events from trials. Are there any groups of patients who were excluded from the trials who might be likely to receive the drug in practice?

Tolerability: Do patients remain on therapy? Examine the withdrawal rates from trials.

Effectiveness: How effective was the drug in trials? Is it clinically significant?

Price: Not just acquisition cost, but cost of administration equipment or blood tests, for example.

Simplicity: Is the device complicated? Is one drug given orally and a comparator by infusion?

It is important that this decision making process is based upon clear criteria and is documented appropriately. This is especially important now that, in England at least, the NHS Constitution promises patients 'the right to expect local decisions on funding of other drugs and treatments to be made rationally following a proper consideration of the evidence'. If the local NHS decides not to fund a drug or treatment that the patient and their doctor think would be beneficial for them, then the patient also has the right to have that decision explained to them.

Being practical with critical evaluation

Critical evaluation skills are important for all healthcare professionals to support the application of evidence-based medicine. It's very likely that you'll need to use these skills to some extent whatever path your career takes, to influence healthcare professionals and patients to make the best decisions, and you may not always realise that this is what you're doing!

When you're doing critical evaluation more formally it's often in response to a request for a medicine to be added to a local formulary. Often these requests are linked to a newly identified need, the potentially inappropriate use of a medicine, a safety issue, or funding. With this in mind, here are some tips on how to save time and make your work a cut above the rest:

1. **See what's already available**

   You could spend several days retrieving papers and evaluating them, only to find that someone else has already done the work. For example, UKMi publishes a list of new product evaluations that are freely available to NHS staff. The list is updated monthly and can be found on the Specialist Pharmacy Service website (type "new product evaluations" in the search box and find the latest version). Databases such as Medline and Embase will help you find reviews in journals. NICE Evidence and the Cochrane Library can also be useful. Using someone else's work does not mean that you have to use it as it stands, but it gives you another point of view and at least helps you with your literature search.

2. **Find out what the issues are**

   The best evaluations pick out the real issues right from the word 'go'. You could do a wonderful assessment of the evidence to support the efficacy of a new antidepressant only to find that the key issue was its improved safety, and so your work wouldn't be very relevant. Identifying the issues can save you time too.

3. **Involve clinical experts**

   By an extension to the above point, it is very clear that some evaluations are not done with the 'inside clinical knowledge' that can make them really sharp and relevant. Work with clinical pharmacists and specialist doctors by asking them to comment on your work. Remember that with critical evaluation you are aiming to assess 'usefulness' and hoping to change or affirm practice: you're more likely to do this well if you invite experts to work with you.

4. **Evaluate, don't summarise**

   Regrettably it can happen quite often that evaluations are little more than summaries. What's the difference? Evaluations look at validity and usefulness: they place the evidence in a clinical context and point out any limitations. They help the reader use the evidence.

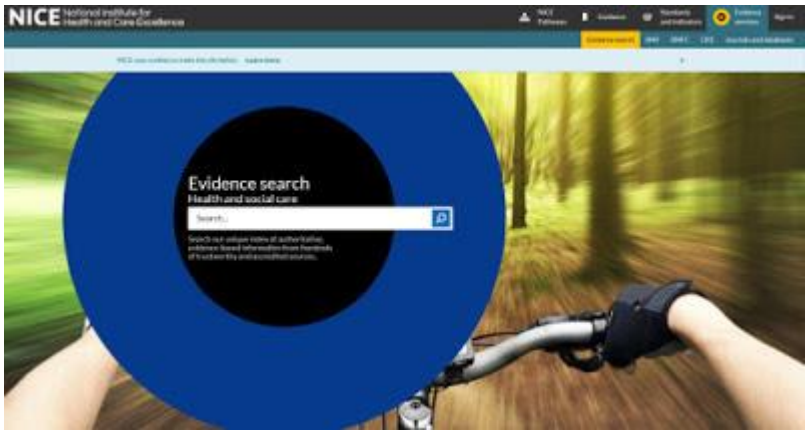
5. **Think 'big picture'**

   Once you understand the issues in clinical practice, you may find that you need to expand your remit. Maybe you have been invited to appraise one drug, but is it the whole drug class that really needs looking at? You may have been asked to evaluate one particular paper, but perhaps you need an audit of current local practice first? Is it an evaluation that's needed or a clinical guideline, or both?

Information sources

The UK Critical Appraisal Skills Programme (CASP) offers a series of practical checklists ('critical appraisal tools') to help you systematically evaluate clinical evidence including RCTs. Explore their website and look at the resources they have available, but note that you might need to update your browser or use a different browser to access the site properly. There are some free e-learning courses on the CASP website too.

The NICE Medicines and Prescribing Centre facilitates the promotion of high quality, cost-effective prescribing. Visit their website for evidence-based summaries about new medicines or unlicensed medicines. You might like to compare the evidence-based reviews of medicines (technology appraisals) and clinical guidelines on the NICE website, with those available in Scotland and Wales. Try the Scottish Intercollegiate Guidelines Network (SIGN), the Scottish Medicines Consortium (SMC) and the All Wales Medicines Strategy Group (AWMSG) for information on the effectiveness of new and existing medicines.



The Cochrane Library also has a collection of evidence-based reviews that are regularly updated, including The Cochrane Databases of Systematic Reviews. NICE Evidence Search is a gateway to many evidence-based resources including its own and those from professional bodies e.g. Royal Colleges.

Be careful about conducting a general internet search when looking for critically evaluated evidence. If you do, you may like to look at our brief guide to evaluating websites about medicines.