AIMS For a better birth...

Understanding quantitative research evidence

What is quantitative research?

There are two main types of research study: quantitative and qualitative (though some studies use a mixture of these methods).

Quantitative research deals with numbers and measurement and will usually use statistical analysis to draw conclusions. There are two main types of quantitative research: Randomised Controlled Trials (RCTs) and various kinds of trials using observational data (sometimes called population or cohort studies). The section below explains the differences between these two types of study, and their strengths and weaknesses.

The results of quantitative research are often quoted as the **Relative Risk** which means how much more common it is for a problem to occur in one group than another. In some types of research, particularly population studies (see below) you will see a similar measure called the Odds Ratio. However, for someone trying to decide about a treatment or medical intervention it is often more useful to look at the **Absolute Risk**. This is the chance of a problem occurring, usually expressed as one in 100, one in 1000 etc.

Qualitative research involves exploring people's opinions, experiences and preferences in depth, usually through interviews, group discussions or questionnaires.

This article is mainly about the use and limitations of different types of quantitative research study. For more about understanding the research relating to pregnancy and birth see our book AIMS Guide to Safety In Childbirth (principal author Gemma McKenzie)

What are the limitations of quantitative research?

Most of the evidence that is used to make recommendations about maternity care comes from quantitative research. This gives us evidence to compare the outcomes from different medical interventions or approaches to maternity care. Unfortunately, evidence is often not clear-cut or may be lacking altogether, and sometimes the research that has been done is of poor quality.

In thinking about what research evidence can tell us, it's also important to keep the following limitations in mind.

Definition of the study group

In any kind of research, grouping people together according to one characteristic, such as age, BMI or the

fact that they conceived through IVF, ignores the fact that there could be important differences between the individuals in a group. For example, pregnancy and birth outcomes might be very different for someone over 40 who has a healthy lifestyle, is completely well and has had a straightforward pregnancy, compared to one who has existing health problems, or who smokes or drinks heavily, yet recommendations may be made based purely on their age.

Publication and reporting bias

Selective reporting, in other words the failure to publish the results of some studies, may be due to Journals only accepting articles about research which have interesting findings or those which support the status quo.

Another issue is what authors choose to report from their research. They may decide not to publish at all, or only report the findings they wanted to prove, or only report some of the outcomes that they measured. Sometimes, if the study did not show the results the researchers were hoping it would, they will pick on some other finding to report, even if that wasn't an original aim of the study, or claim that there was "a trend towards" a finding even if it was not statistically significant (see below). It is often necessary to dig into the detail of a study to see whether the headline results are supported by the evidence or not.

Short term outcomes

Quantitative research usually only looks at outcomes that can be measured in the short term. Occasionally there will be follow-up studies that seek to understand the long-term consequences of a medical intervention, but these are the exception. This means that we often don't know about all the risks, as those that arise after the study ends will not have been recorded.

Focus on selected benefits and risks

Studies are only able to focus on a small number of outcomes, so may not be able to provide information about other risks and benefits which you would want to know about when making a decision. This is partly because of practical considerations, but also because if you include a very large number of outcomes in a study it makes the conclusions less reliable, as it much more likely that something will be found that is a chance finding, but not a real effect.

Lack of views of study participants

Only rarely do the researchers carrying out a clinical study ask how the people receiving the care felt about it. Even when they do, the information is usually limited to something that can be quantified, such as asking 'how satisfied' they were with the care.

What is a Randomised controlled trial?

The kind of research that is usually considered to be the 'gold standard' is a Randomised Controlled Trial

(RCT). This is where a group of people is randomly divided into two or more groups, each of which receives a different treatment or type of care. The fact that the allocation into groups is random helps to ensure that the groups contain a similar mix of people. That way any difference in outcomes is likely to be due to the treatment or care received, rather than to the groups including more or fewer people with certain characteristics.

RCTs can work very well if it is a case of comparing something like the effectiveness of two drugs but are more problematic when researching something as complex as pregnancy and labour. There are also limitations on how reliable the findings of any RCT can be, as discussed below.

As a result of these problems it is often the case that too few large and well-conducted RCTs have been done to allow any meaningful conclusions to be drawn about what care is best. In some cases, no good studies have been done at all.

The findings of even the best RCT will be limited to answering a specific question about a particular treatment for a particular group of people (and often in a particular healthcare setting such as a hospital). It won't be able to tell us everything that we might want to know about that treatment for other people or in other settings, or where additional factors are involved.

Blinding of participants and researchers

Each of the groups in an RCT receives a different treatment, so that the outcomes can be compared. Ideally, such trials would be 'blinded' which means that neither the person nor those conducting the trial know which group an individual is in.

The problem with RCTs looking at care in pregnancy and labour is that blinding isn't possible. Knowing which group a pregnant woman or person is in may affect both their level of anxiety (which itself can affect the outcomes) and the behaviour of their doctors and midwives, resulting in unconscious bias. For example, a doctor may believe that waiting for labour to start is riskier than inducing it after a set number of weeks of pregnancy. If they are caring for a mother in the group that waited for labour to begin, the doctor may feel they need to intervene if they notice signs that would not normally cause them concern. This could affect the frequency of unnecessary caesareans and assisted births. This and other issues are discussed in the article "Routine induction of labour at 41 weeks gestation: nonsensus consensus"[1]

Other issues of Bias

The results of an RCT can also be misleading if the study was carried out in a way that made it biased. For example, there might be important differences between the groups if

- the way in which people were allocated was not truly random
- a lot more people from one group dropped out during the study for some reason
- if there was a lot of cross-over.

Cross-over is where high numbers of people ended up having the opposite treatment or care to the one

they were intended to have. Some cross-over in RCT is expected. For example, in an RCT on planned caesareans it's very likely that some women allocated to a planned caesarean will go into labour before it can be done, and some who are in the planned vaginal birth group will decide to have a caesarean because a concern has arisen before their labour started. For the results to remain valid this cross-over needs to remain low.

In research on induction of labour it is quite common for a high proportion of those allocated to the expectant management (waiting for labour) group to have their labours induced because they have reached a pre-set deadline for the birth to take place or there is a concern over their or their baby's well-being. Similarly, some of those allocated to the induction group may go into labour before the induction is started. This can reduce the reliability of the findings.

Low recruitment rates

It is often not possible to recruit a large enough sample to be able to measure a difference in outcomes. For example, to detect a difference in very rare occurrences such as stillbirth it has been estimated that it would be necessary to include between 16,000 and 30,000 pregnancies in the trial, and this is not usually possible in practice.[2]

What is a meta-analysis?

One way around the problem of recruiting a large enough sample is a 'meta-analysis', a type of review which combines the data from multiple studies and uses statistical methods to analyse it. Effectively, a group of studies are analysed as if they were all part of one big study, but there are problems with this approach.

A meta-analysis can only be as good as the trials that go into it, and the results can vary according to which trials the authors choose to include. It is also difficult for a meta-analysis to compare the results of RCTs if they were carried out in different ways or using different methods, or if some important outcomes were not reported in all the studies.

The authors of such reviews will usually comment on the quality of the studies that they have selected for inclusion, and on the overall quality of the evidence available. The typical rating is High/Moderate/Low/Very Low where 'High' means that the authors are very confident of having detected a real effect and 'Very Low' means they are not at all confident.[3]

One of the best-known sources of meta-analyses for all kinds of medical questions is the global Cochrane network. Their approach is explained here www.cochrane.org/about-us .

What are observational or population studies?

These are studies which observe how the outcomes in real life situations differ between groups defined

by one or more characteristics or by a difference in the treatment that they receive. They are also sometimes called cohort studies. There is no random allocation of people to the different groups in this type of study, so they are usually considered to provide poorer quality evidence than an RCT. Nevertheless, they can provide useful information, especially on subjects where large, good-quality RCTs are lacking.

Mostly these are retrospective studies which look back at the records of a population, often over a period of years, and try to identify whether there were certain groups who were more likely than others to experience a given outcome. Alternatively, the investigators may look at how outcomes differed before and after a change in standard care procedures or else compare outcomes according to differences in characteristics such as age or the presence of a health condition.

There are also prospective studies which define the groups to be studied at the start of the research and then follow up what happens to those who fall within these groups, for example, those that do and don't have their labours induced.

The advantages of this type of study are that they can often involve larger samples than RCTs are able to recruit, and they are looking at what happened in real life. In some cases, especially where it would be impractical or unethical to do an RCT they may provide the only evidence we have.

The disadvantages are that the records that are used are often incomplete, and because there is no randomisation there could be important differences between the groups or in the ways in which their labours were managed which are often not identified but could have a major impact on the results.

What is "Statistical significance" and why does it matter?

Studies can report their findings in different ways, but the authors should always use some sort of statistical analysis to check the probability that their results reflect a real difference. If it is highly unlikely that the finding occurred by chance, this is usually described as being a "statistically significant" result. Note that this is a technical use of the word "significant". It is not saying anything about the relevance or importance of the finding.

Most medical studies use a significance level of 5%. This means that there's a 95% chance that the finding is real, but still a 5% chance that it is not real. Another way to look at this is that there is a one in 20 (5%) chance that a result found by a study is not a real effect, so a recommendation that is based on just one result in one study may not be reliable.

It is usual to report the **95% confidence intervals** for a research finding. This tells you the range of values within which there is a 95% chance that the true value lies. In other words, if an RCT finds that something is 2.5 times more common in one situation than in another, it might report this **Relative Risk** (RR) as RR 2.5 (95%CI 1.6 to 3.2). That means that the most likely value for the RR is 2.5, and there is a 95% chance that the true value lies somewhere between 1.6 and 3.2. This means we can be fairly confident that the

outcome being measured really is more common in one situation than in the other. We can also be fairly confident that it is between 1.6 and 3.2 times more common. However, there is a 5% chance that the true value is outside these limits, (meaning that there is a 2.5% chance (one in 40) that the RR is less than 1.6, and a 2.5% chance that it is greater than 3.2).

If both the upper and lower confidence intervals are less than one, then this also indicates that the result is statistically significant, but that there is a reduction rather than an increase in the risk.

For a result to be statistically significant both figures must be greater than one or both less than one. This indicates that it is likely that the study has identified a real effect. If the lower number is less than one (which implies that the effect is to reduce the risk) and the upper figure is greater than one (which implies that it increases the risk) the result is not statistically significant. The study does not show the whether the risk is reduced or increased in one situation compared to the other.

When the numbers in the confidence interval are very different (but either both greater than one or both less than one) this is referred to as having a wide confidence interval. This means we can be confident that there is an effect, but not very sure how big the effect is. We can still say what the most likely value is, but the true value could be very different.

Sometimes when a study did not produce the result that was expected (or hoped for) authors may make more of non-significant results than they should. The sort of phrases to watch out for are things like “there was a trend towards x” or “the findings were borderline significant” or “there was an increased/decreased chance of x, but this did not achieve significance”. What all these phrases mean is that we don’t know whether the effect is real or not, and there is a high chance that the effect was not a real, because the finding was *not* statistically significant.

Critiquing research

Research papers, especially if they are likely to have a major impact on clinical practice, will often be ‘critiqued’ by other experts in the field. This means reviewing the strengths and limitations of the research and deciding whether the conclusions can be relied on. A critique of an RCT will usually include analysis of things like:

- Whether the research question was clearly defined, and appropriate outcomes reported
- How good the random allocation of people to the different groups was, and whether it resulted groups with similar characteristics (e.g. age, education level etc.)
- Whether there were any differences in the care given to each group, apart from the treatment being investigated
- How much cross-over there was between group
- Whether a large proportion of those enrolled in the study dropped out along the way
- The size of effect, statistical significance and confidence intervals for each outcome reported

If you want to get a feel for how reliable a piece of research is, it is worth looking to see whether a critique has been published. Examples of this include "Routine induction of labour at 41 weeks: nonsensus consensus"[1] and "Parsing the ARRIVE Trial: Should First-Time Parents Be Routinely Induced at 39 Weeks?"[4]

An example of tools to use if you want to try critiquing different types of research yourself can be found here casp-uk.net/casp-tools-checklists

References

1. Menticoglou S.M. and Hall S.F. "Routine induction of labour at 41 weeks: nonsensus consensus" BJOG 109 (5) May 2002 pp485-491 obgyn. onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-0528.2002.01004.x
2. Mandruzzato G. et al "Guidelines for the management of postterm pregnancy." J Perinat Med. 2010 Mar;38(2):111-9 www.ncbi.nlm.nih.gov/pubmed/20156009/
3. Siemieniuk R. and Guyatt G. "What is GRADE?" BMJ Best Practice bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/
4. Goer H. "Parsing the ARRIVE Trial: Should First-Time Parents Be Routinely Induced at 39 Weeks?" 2018 www.lamaze.org/Connecting-the-Dots/parsing-the-arrive-trial-should-first-time-parents-be-routinely-induced-at-39-weeks