# INTERPRETING THE CLINICAL LITERATURE

*David Webb*

Many treatment decisions are guided by the results of published clinical trials. However, without a basic knowledge of how to assess the quality of a trial, the validity of the results and conclusions, and how to interpret the data in a way that is relevant to one's own clinical practice and individual patients, it is easy to be misled by poor trial design, biased reporting and irrelevant data. Furthermore, it is common to find trials that have a similar design, but directly opposite results. In that case, how does one decide which might be more relevant?

The following is a brief summary of some guiding points to bear in mind when reading published clinical studies.

## GENERAL PRINCIPLES

1. Unless you read the whole article, and very importantly thoroughly *understand the methodology and inclusion and exclusion criteria*, it is impossible to interpret the results for yourself.

   THE RESULTS OF A CLINICAL TRIAL APPLY TO THE TYPE OF PATIENT INCLUDED IN THE TRIAL, TREATED WITH THAT SPECIFIC INTERVENTION, AT THAT DOSE FOR THAT DURATION OF TIME UNDER THOSE SPECIFIC CONDITIONS.

   The results will not necessarily, and are often unlikely to be applicable to people with different characteristics, treated with different doses of drug and/or for different treatment durations, or treated with different drugs. Often in order to reduce the chance of non-intervention characteristics confounding the interpretation of the effects of the study intervention (e.g., a medication), patients with the profile of those who might be more common in everyday clinical practice are excluded from the study. For example, the results of studies done in men do not necessarily apply to women; those in patients without specific comorbidities may not apply to those with comorbidities. Results cannot be extrapolated to patients with different severities of disease (e.g., mild vs. severe depression) or risk of disease (e.g., high vs. low cardiovascular risk); and results from studies in adults cannot be extrapolated to people of different ages and especially not to children.

*David Webb*

2. The results of a study tell you what is expected to happen *in a population* of people with the characteristics of the inclusion and exclusion criteria and they give you an idea of probability.

   THEY DO NOT TELL YOU WHAT WILL HAPPEN TO AN INDIVIDUAL PATIENT IN CLINICAL PRACTICE.

3. Statistically significant results only indicate the probability of repeatability. They are not necessarily clinically relevant. The clinician will need to make his/her own assessment of what clinical significance is. Including a larger number of subjects in a study can make smaller

differences in outcomes statistically significant, which might not have been statistically significant in a smaller, or a differently designed study (or with application of a different statistical test!).

THE P VALUE IS A RUDIMENTARY AND CLINICALLY UNHELPFUL INDICATION OF PROBABILITY OF REPEATABILITY. ON ITS OWN IT PROVIDES NO INFORMATION ABOUT WHICH INTERVENTION IS BETTER, THE MAGNITUDE OF TREATMENT EFFECT OR CLINICAL RELEVANCE OF TREATMENT EFFECTS.

4. Relative risk (RR) or relative changes in risk after treatment can only be interpreted if you know the baseline risk (i.e., the risk before intervention) and what the absolute changes in risk (AR) are. They *can never* be interpreted on their own. Despite this, especially in review articles, RRs and risk changes are often discussed in isolation without mention of AR or baseline risk and are, in that case, misleading and meaningless.

5. BE CRITICAL WHEN READING THE RESULTS. BE CAREFUL OF MISLEADING PRESENTATION OF RESULTS IN THE TEXT, GRAPHS AND TABLES. FOR EXAMPLE, IT IS EASY TO MANIPULATE THE SCALE OF A GRAPH TO MAKE SMALL DIFFERENCES IN OUTCOMES VISUALLY APPEAR LARGER THAN THEY ACTUALLY ARE.

   Consider the authors' interpretation of the results carefully and make your own assessment

## TRIAL DESIGN

When reading the trial design, pay special attention to the following aspects:

- Inclusion and exclusion criteria.
- Number of patients included in the study as a whole (N) and in each intervention group (n).
- Whether the study is randomised (i.e., patients are randomly assigned to the different interventions) and stratified (i.e., patients are separated out by common characteristics within randomised groups; e.g., diabetics versus non diabetics).

  STRATIFICATION IS USUALLY DONE ONLY AFTER RANDOMIZATION TO ENSURE THAT THE DIFFERENT TREATMENT GROUPS ARE AS SIMILAR AS POSSIBLE AND AVOID RANDOMIZATION BIAS.

- The specific interventions, comparators and controls, their doses, allowances for dose escalation or other dose changes, and durations of treatment.
- Parallel or crossover study design. In a crossover study patients may sequentially receive more than one of the interventions. Even with a washout period, there may be a risk of the first intervention affecting the outcome of subsequent interventions.
- Is the study blinded? Blinding should apply to the patient, clinician and others who are interpreting outcomes (e.g., radiologist, laboratory). Any gaps in blinding may predispose to bias.
- Concomitant medications may influence the results positively or negatively. For example, an intervention added to usual standard of care (SOC) tells you what will happen when you add the intervention to that defined SOC. The results then do not apply to using the intervention alone or without any of those elements of that defined SOC.
- OUTCOMES AND *HOW THEY ARE DEFINED*. THE METHODOLOGY SHOULD INDICATE WHAT TREATMENT EFFECT IS REGARDED AS SIGNIFICANT AND WHAT TREATMENT EFFECT IS CLINICALLY RELEVANT (AND WHO OBSERVED THEM).

Some pitfalls to look out for in trial design that might indicate risk of bias are listed in Box 1.

**Box 1. Potential sources of bias in trial design and reporting**

1. Source of funding.
2. Is the study objective balanced, clinically relevant and in context? Are the proposed outcomes clinically relevant?
3. Inadequate power, so that a conclusion that there is no difference between treatments may be false.
4. Subgroup analysis to obtain significant P value.
5. Claims of equivalence despite lack of pre-defined level of equivalence and adequate power to demonstrate it.
6. Atypical population enrolled vs. clinical practice.
7. Not properly randomised.
8. Small patient numbers.
9. Inadequate blinding or inadequate concealed allocation of treatment.
10. Inappropriate controls or no control group. In comparisons of drugs, is the comparison appropriate? Is the dose regimen of all drugs comparable and representative of what is feasible in clinical practice?
11. Distortion of results: what results are emphasised? Is anything left out? Are the conclusions congruent with the data?
12. Is statistical and clinical relevance discussed?
13. Are the results put into a balanced context?

## ENDPOINTS

- ***Primary endpoints:*** The trial has been designed specifically to investigate these endpoints and the requirements for the statistical analysis, including patient numbers, to test these endpoints have been included in the trial design.
- *Secondary endpoints:* Additional endpoints that are pre-specified before the trial starts. The trial may not be powered to show statistical significance for these.
- ***Post-hoc observations:*** These are interesting outcomes (e.g., apparent treatment effects either in the group as a whole or in specific subgroups of patients) observed once the results have been analysed.

THESE RESULTS AND THE STATISTICS ASSOCIATED WITH THEM NEED TO BE INTERPRETED WITH CAUTION. IF IT IS WARRANTED, A NEW TRIAL MAY BE DESIGNED TO SPECIFICALLY CONFIRM THAT THESE OUTCOMES ARE REAL AND NOT JUST DUE TO CHANCE.

## OUTCOMES

***Mean:*** The sum of the values divided by the number of observations (the" mathematical mean"). Means are often used for continuous outcomes - that is outcomes that are continuously variable, such as blood glucose or blood pressure. However, on their own they can be clinically unhelpful, because they do not indicate what the spread of the results might be.

FOR EXAMPLE, IF 50% OF PATIENTS RESPOND WITH VERY HIGH VALUES AND 50% RESPOND WITH VERY LOW VALUES, THE MEAN VALUE, LYING SOMEWHERE IN THE MIDDLE, MIGHT NOT BE REPRESENTATIVE OF ANY OF THE PATIENT OUTCOMES IN THE STUDY.

***Median:*** The middle value when all the entire spread of values is considered. The median tells you the value where 50% of individuals lie in terms of results. That is, 50% of the study population will fall below the median value and 50% will lie above the median value. Medians are used for survival data - time to reach a specific endpoint (e.g., HAM-D score ≤24; time to relapse after achieving remission; death) and where there is a large spread of data and means are inappropriate.

WHEN IT IS PRESENTED WITH UPPER AND LOWER LIMITS OF OBSERVATIONS, THE MEDIAN MAY BE MORE HELPFUL FOR PREDICTING HOW PATIENTS IN CLINICAL PRACTICE ARE LIKELY TO RESPOND.

***Mode:*** The most common value. The modal drug dose can be more useful than the mean drug dose, because it tells you the specific dose that was most commonly required to achieve a specific outcome. In contrast, mean drug doses may be clinically meaningless – none of the patients actually receive that dose and medications are rarely available in formulations equal to the mean dose.

## PREVALENCE AND INCIDENCE

Prevalence and incidence are different, but are often confused and used interchangeably. *Prevalence* is the proportion of patients meeting specific criteria at a particular point in time. For example, the percentage of South Africans who are HIV positive on 1st January 2020. In contrast, the *incidence* is the proportion of people in a specific population who will develop a condition during a defined time period. For example, the proportion of South Africans who became HIV positive between 1st January 2019 and 1st January 2020.

HIV IS A GOOD EXAMPLE WHERE TRENDS IN PREVALENCE AND INCIDENCE MAY BE VERY DIFFERENT. WITH BETTER ACCESS TO ANTIRETROVIRAL THERAPY AND EDUCATION, THE PREVALENCE OF HIV MAY INCREASE OVER TIME, EVEN THOUGH THE INCIDENCE MIGHT BE DECREASING.

## STATISTICAL MEASURES OF DIFFERENCE

- ***P value***

The P value gives an indication of the probability of whether an event (e.g., difference between treatments) is real or has merely happened by chance. P is the probability that there is no real difference (i.e., that any observed difference is a chance event). Statistical significance is defined by an arbitrary cut-off point of $P \leq 0.05$. If $P = 0.05$ there is a 5% probability that the result is due to chance and a 95% probability that the observed event is not due to chance.

IF WE REPEAT THE STUDY 100 TIMES WE WOULD EXPECT TO OBSERVE A DIFFERENCE IN 95 OF THESE STUDIES. $P \geq 0.05$ IS REGARDED AS NOT BEING STATISTICALLY SIGNIFICANT.

The P value is a measure of statistical significance only. It does not give any indication of magnitude of result, or in which direction the result occurs (i.e., which treatment is better). Most importantly, regardless of how small the P value is, it tells you nothing about whether an event is clinically significant or not. On the contrary, by manipulating trial design and including large patient numbers, statistical analysis of small and clinically insignificant differences can yield very small P values that indicate very high statistical significance.

- ***Confidence intervals***

The confidence interval is calculated from the data. It provides a range of values that we can confidently say contains the true mean value. Most commonly, studies report the 95% confidence interval (CI95%). This gives a range of values which we can be 95% confident contains the true mean. For example if one antihypertensive drug reduces blood pressure on average by 3 mmHg more compared to another and the CI95% is 2 to 5 mmHg, that tells us that we can be 95% confident that the true mean difference in efficacy lies between 2 and 5 mmHg.

NOTICE THAT THE MORE SUBJECTS FROM A POPULATION WE INCLUDE IN A STUDY, THE MORE CERTAIN WE CAN BE THAT THE CALCULATED MEAN REPRESENTS THE TRUE MEAN OF THAT POPULATION.

Therefore our CI95% will become smaller (more certain) as the number of subjects included in the study increases.

- ***Relative and absolute differences in results (risk)***

Measures of risk are used for survival outcomes. Although these are called survival outcomes, they do not necessarily refer to death, but to any endpoint that is either achieved or not within a specified time interval. Other examples include time to myocardial infarction, hospitalisation, relapse or achievement of remission.

Here is a theoretical example of a study investigating the ability of a drug to prevent stroke (Table 1). One thousand patients are assigned to the placebo arm (i.e., do not receive active drug) and 1000 patients are assigned to receive drug. After a certain time period (e.g., 48 months), there are 40 strokes among the subjects receiving placebo and 26 strokes among the subjects receiving drug.

As can be seen from Table 1, the risk of having a stroke if you receive placebo and fit the inclusion/exclusion criteria of this trial is 4%.

THIS IS VERY IMPORTANT, BECAUSE IT TELLS YOU WHAT THE BASELINE (ABSOLUTE) RISK OF THIS POPULATION IS. THAT MEANS THAT YOU CAN COMPARE WHAT EFFECT A DRUG TREATMENT HAS ON STROKE TO NO ACTIVE TREATMENT (I.E. PLACEBO).

The drug reduced the incidence of stroke from 4% (untreated) to 2.6% (on drug treatment), giving an ARR of 1.4%. This is the true reduction in risk. However, one can also calculate risk reduction relative to the baseline risk. The baseline risk is 4% and drug treatment reduced the risk to 2.6%. Therefore, drug therapy reduced the risk of stroke to 65% of what it would have been without drug treatment (RR = 2.6/4.0 = 0.65). In other words, you have reduced the risk of stroke by 35% (relative risk reduction, RRR = 100%-65% = 35%).

**Table 1. Hypothetical example of drug versus placebo for prevention of stroke**

| | Placebo | Drug |
|---|---|---|
| Number of patients included in study | 1000 | 1000 |
| Number of patients with a stroke after 48 months | 40 | 26 |
| Incidence of stroke during 12 months | 40/1000 = 0.04 (4%) | 26/1000 = 0.026 (2.6%) |
| Risk of having a stroke in 12 months | 4% | 2.6% |
| Absolute risk reduction (ARR) | | 4%-2.6% = 1.4% |
| Relative risk of stroke (RR)* | | 2.6/4.0 = 0.65 = 65% |
| Relative reduction in stroke risk (relative risk reduction, RRR) | | 100%-65%=35%, or 1.4/4.0 = 35% |
| | | |
| Number needed to treat (NNT) | | 100%/1.4% = 71 |

*Relative risk is sometimes referred to as 'risk ratio'.

Note that you cannot interpret relative changes in risk if you don't know what your baseline risk is. For example, a 50% reduction in risk of cancer means nothing if you don't know what the chance of you getting cancer is in the first place. Consider this example. If the baseline risk of cancer is 10% (that is 10 out of 100 people in this population will get cancer over a prespecified time), a RRR of 50% reduces the chance of cancer to 5% (5 people out of 100); treatment has saved 5 people. However, if the baseline risk of cancer is 20% (that is 20 out of 100 people in this population will get cancer over a prespecified time), a RRR of 50% reduces the chance of cancer to 10% (10 people out of 100); treatment has saved 10 people.

MEASURES OF RISK ARE BASED ON POPULATION RESULTS – THEY TELL YOU WHAT WILL HAPPEN IN A POPULATION OF PEOPLE WITH PRESPECIFIED CHARACTERISTICS. THEY CANNOT TELL YOU WHAT WILL HAPPEN TO AN INDIVIDUAL.

RR can be plotted on a Forest plot, which gives a quick visual representation of the outcomes data (Figure 1). The RR Forest plot shows mean RR, represented by a dot, and the 95% confidence intervals for that RR. Note that if the treatment effect is the same for both interventions, the relative risk will be equal to 1. If the 95% confidence intervals do not cross this equivalence line, then it can be shown mathematically that the RR is statistically significant.

## NUMBER NEEDED TO TREAT

Number needed to treat (NNT) tells you how many patients (with these characteristics) would have to be treated to prevent one event. NNT = 1/ARR, or 100%/ARR%. In our example in Table 1, NNT = 1/0.014 or 100%/1.4% = 71. The NNT turns risk calculations into an easily understandable and clinically useful number. In this example, you would have to treat 71 patients (with similar characteristics to those included in the study) with the drug to prevent one stroke. Alternatively, out of every 71 patients treated, 70 would not benefit from treatment. Here, the NNT appears to be large, but in fact, depending on the population, it is not terribly different from that of many drugs used in prevention of disease.

THE PATIENT'S ESTIMATED BASELINE RISK, COST, COMPLIANCE AND THE POTENTIAL FOR ADVERSE EFFECTS WOULD NEED TO BE TAKEN INTO CONSIDERATION WITH THE NNT WHEN DECIDING ON WHETHER TO PRESCRIBE OR NOT.

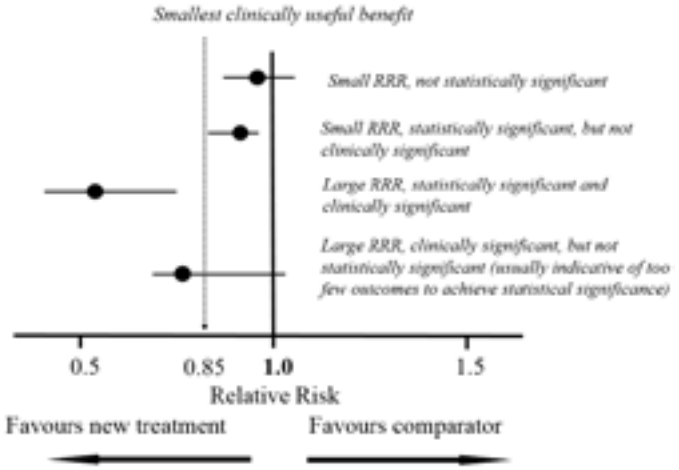


**Figure 1. Forest plot showing relative risk**
The black dot is the mean relative risk (RR) and the horizontal line running through it indicates the 95% confidence intervals. The vertical line indicates a RR of 1, meaning that the risk on both treatments is the same (absolute risk change on treatment = 0).

The Forest plot indicates only statistical significance. Therefore, the clinician will need to decide for themself the magnitude of risk reduction they consider to be clinically significant (i.e., that would justify use of the new treatment over the comparator). In this example, an arbitrary risk reduction of 15% is considered clinically significant and is indicated by the dotted vertical line.

RRR: Relative risk reduction

## ANALYSIS OF DATA

- ***Intention-to-treat analysis***

Intention to treat (ITT) analysis includes the results for everybody who was randomised in the study, regardless of whether they received intervention. It includes drop-outs and those with missing data. ITT is the most unbiased method by which to analyse data from a randomised controlled trial.

IF A LARGE NUMBER OF SUBJECTS IN THE STUDY DROP OUT OR DO NOT RECEIVE INTERVENTION, ITT ANALYSIS CAN UNDERESTIMATE OR OVERESTIMATE THE EFFECT OF INTERVENTION. FURTHERMORE, WITHOUT KNOWING THE REASON WHY PATIENTS DROPPED OUT (E.G., DUE TO LACK OF EFFICACY, POOR COMPLIANCE OR ADHERENCE, SIDE EFFECTS, DEATH) IT IS DIFFICULT TO FORM ANY CONCLUSIONS FROM THE AVAILABLE RESULTS.

Therefore, when there are substantial numbers of subjects without a full set of data, results and conclusions should be regarded with caution.

In drug trials, a modified ITT (mITT) analysis usually includes only patients who received at least one dose of active drug or comparator in each respective group.

- ***Per protocol analysis***

Per protocol (PP) analysis of results excludes patients in a trial who did not follow the study protocol. It excludes non-compliant patients, patients who used medications not allowed in the study or who violated inclusion or exclusion criteria, and those for whom data were missing.

## CONCLUSION

Conclusions presented in a clinical trial should never be interpreted at face value.The authors' conclusions are their own interpretation of the results obtained in the context of their own objectives, a trial design, a specific (and often unique and clinically unusual) patient population, specific intervention and the statistics that have been used to interpret the data.

HAVING A RUDIMENTARY KNOWLEDGE OF HOW TO INTERPRET TRIAL DESIGN AND BIOSTATISTICS CAN HELP TO ALERT TO POTENTIAL BIAS AND SORT OUT CLINICALLY RELEVANT DATA FROM THAT WHICH IS MORE DIFFICULT TO ASSESS.

By carefully reading through the whole article it is possible to gain better perspective and decide for yourself the relevance to your own practice.

***David Webb*** is a medical doctor, and medical writer. He offers a half-day workshop on interpreting clinical trials based on practical examples taken from the literature. More information and a descriptive brochure can be found at www.poetryofaddiction.com
**Correspondence: dawebb@mweb.co.za**