

How distinguish between statistically significant results and clinically relevant results

21 December 2015

Derrick. A. Bennett, PhD,
Clinical Trial Service Unit and Epidemiological Studies Unit,
Nuffield Department of Population Health, Old Road Campus,
University of Oxford, Oxford, OX3 7LF, UK

Running head: Clinical versus statistical significance

Tables and figures: 2 tables and 2 figures

Word count excluding references, tables and figures: 5894

ABSTRACT

Background: A practising clinician will often be confronted with the results of a new clinical trial in their relevant field and will be faced with the dilemma of determining whether these results are clinically relevant to their own work. This chapter aims to describe the concepts of statistical significance in randomized control trials from a mainly classical statistical inference perspective. Approaches to assess clinical significance will be described as well as illustrating these approaches with examples from the contemporary neurological literature.

Results: There are several approaches that have been described in research literature to assess the clinical significance including the minimal important clinical difference, the fragility index, Bayesian approaches, and a graphical approach. Unfortunately none of these methods have been widely used in the neurological research literature. Examples are provided to illustrate how these methods can be applied to the contemporary neurological literature in order provide the clinician with some guidance on their use.

Conclusions: How the trial is designed can affect the external validity of the results and subsequently the clinical relevance of a randomized controlled trial. Large-scale streamlined clinical trials with inclusion criteria that are not too restrictive can improve the generalizability of trial results. Even highly statistically significant treatment effects can be unreliable if they are based on a small numbers of events. The approaches described in this chapter should provide the practicing clinician with a starting point in order to determine whether the statistically significant results are indeed clinically relevant.

Keywords: statistical significance, clinical significance, classical inference, minimally important clinical difference, fragility index.

Abstract: 244 words excluding keywords

INTRODUCTION

Statistical thinking is an integral part of the research process in many scientific disciplines. A good understanding of statistics is relevant for defining the research question, design of the study, data collection, processing and analysis through to interpretation and presentation of the results. In order to be as effective as possible, health professionals need to be able to read and evaluate the findings produced by research in their chosen field. Thus a good grasp of the key statistical concepts is crucial for critically appraising the evidence and being able to put that evidence into practice. This chapter is split into four sections. First, a brief overview of statistical inference and its role in medical research are described. Second, an outline of the role of the randomized controlled clinical trial (RCT), in medical research is discussed and why such study designs are the best way to assess causality. Third, the issue of generalizability (external validity) of an RCT is discussed and potential design issues that could make it difficult to apply the results of an RCT to clinical practice are described. Fourth, some approaches from the literature that aim to assess the clinical significance of RCT results are discussed with examples of trials from the neurological literature. The ultimate goal is to provide the readers with the basic tools to improve their ability to assess the clinical relevance of published neurological trials in relation to their patients and to public health in general.

1. Classical statistical inference

Classical statistical inference, involves drawing conclusions about a population on the basis of sample of the population. The two most common forms are hypothesis testing and estimation (confidence intervals).

1.1 Hypothesis testing

A hypothesis is some testable belief or opinion, and hypothesis testing is the formal process by which statistical methods are used to assess the plausibility of this belief or opinion.

Hypothesis testing attempts to measure the strength of evidence based on the sample of

data concerning the research question of interest. The *null hypothesis* (denoted H_0) is often the negation of the research question that generated the data. In medical research, where comparisons are being made between treatments or different groups of patients, the null hypothesis is that the “true effect” of interest in the population is often zero or unity. The *alternative hypothesis* is usually the opposite of the null hypothesis e.g. that the “true effect” of interest in the population is not zero. The *alternative hypothesis* (denoted H_1 or H_a) can on occasion specify that the “true effect” in the population falls in one direction e.g. that the true effect is greater than zero or unity or that the true effect is less than zero or unity.

A *one-sided* alternative hypothesis is used when the *direction of effect is specified*. A *two-sided* test is used when *no direction of effect is specified* (i.e. the “true effect” could be greater or less than the hypothesized value).⁽¹⁾ There are only four possible results when we test a given hypothesis.

- a. We accept a true hypothesis - a correct decision.
- b. We reject a true null hypothesis - **an incorrect decision (Type I error)**.
- c. We reject a false hypothesis - a correct decision.
- d. We accept a false null hypothesis - **an incorrect decision (Type II error)**.

It is not possible to make a correct decision with a 100% certainty when a hypothesis is tested via sampling and there is always the possibility of Type I or Type II error. In the statistical literature the probability of a Type I error is denoted by α (alpha) and the probability of a Type II error is denoted β (beta).

1.2 Significance levels

Statistical significance measures how likely is it that any apparent differences in the observed sample data (for example in outcome between treatment and control groups) are real and not due to chance. The comparison of the test statistic from a statistical test with the

appropriate distribution will return a *p-value*, which will indicate the probability of obtaining a value as large as or more extreme than the test statistic when the null hypothesis is true.

The p-value is usually compared to some selected cut-off value known as a significance level. The cut-off value for statistical significance in most medical studies is conventionally taken as 0.05 (5% significance level). The hypothesis test aims to assess whether the difference, between the hypothesis and the sample data, can be attributed to random (chance) factors or not. If the hypothesis test indicates that the effect is probably not due to chance factors then the null hypothesis can be rejected and the result is said to be statistically significant (p-value <0.05). Although the 5% significance level is conventionally used in the medical literature depending on the nature of the study question the p-value may need to be much smaller than 0.05 before the study results can be considered to provide strong evidence against the null hypothesis in many situations. (2)

1.3 Confidence intervals

Confidence intervals estimate the range within which the real results would fall if the study was conducted many times. Specifically, the 95% confidence interval (CI) of the difference in treatment outcomes between two groups would indicate the range which the differences between the two treatments would fall on 95% of the occasions, if the study was carried out many times. (3) So, if the study was conducted on twenty occasions then in 1 out of the 20 occasions the 95% CI would not contain the “true effect” size just by chance. Hypothesis testing and confidence interval approaches to classical statistical inference are complementary as shown in **Figure 1**. If the hypothesis test concludes that there is no difference between the two groups the 95% CI would contain the “no difference” value and the p-value would be greater than 0.05. Conversely, if the hypothesis test concludes that there “is a difference” between the two groups the 95% CI would exclude the “no difference” value and the p-value would be less than 0.05.

FIGURE 1 HERE

2. Bayesian statistical inference

The Bayesian paradigm differs from the classical statistical inference paradigm in that the uncertainty about an unknown parameter (e.g. a treatment effect) is expressed through an entire distribution called the prior distribution.⁽⁴⁾ The prior distribution for the treatment effect, expresses the prior uncertainty about the size of the treatment effect before collecting the data for the study. The basis of the Bayesian statistical inference is to revise the estimate of the treatment effect based on the prior information to obtain a posterior distribution of the treatment effect.^(4, 5)

A major consideration of Bayesian statistical inference is the choice of the prior distribution. If there is not much prior information available then a non-informative (sometimes called a vague, flat or reference prior) can be used and this will have a minimal impact on the overall Bayesian analysis.⁽⁶⁾ Information regarding the likely treatment effect and the uncertainty in this prior information may be obtained from the medical literature, pilot studies, or elicited from recognized experts in the relevant clinical area of interest. However, it is crucial that a wide range of experts are consulted to elicit prior information in order to encapsulate or represent a range of points of view, so that the uncertainty in their estimates is fully appreciated.⁽⁷⁾ The posterior estimate contains the information from both the prior estimate and the findings of the new study. The posterior estimate enables the assessment of a range for the treatment effect but there is no arbitrary cut-off whereby the study results are deemed positive or negative. This contrasts with classical hypothesis testing, in which a p-value of 0.049 may be considered positive (i.e. statistically significant at the conventional 5% level), whereas a p-value of 0.051 may be considered negative (i.e. statistically non-significant at the conventional 5% level). ⁽⁷⁾ Several authors have proposed that more widespread use of Bayesian statistics would prevent the mistaken interpretation of $P < 0.05$ as showing that the null hypothesis is unlikely to be true and suggest that the adoption of

these approaches would greatly improve the quality of medical research. (8, 9) We have already mentioned in the previous section that confidence intervals may be calculated in a classical analysis but confidence intervals derived from a Bayesian analyses have a slightly different interpretation to the classical confidence interval and are called credible intervals. A 95% credible interval is one that has a 95% chance of containing the true population parameter of interest (e.g. treatment effect). (2, 4)

3. Introduction to randomized controlled trials (RCTs)

Randomized controlled trials are considered to be the most robust study design for assessing causality in medical research.(10) For example, in the simplest parallel group design to test the efficacy of a novel drug in the treatment of a specific disease, some patients with the disease are given the new drug and some are given the best existing drug. The two groups are then compared prospectively for incidence of disease and other important outcomes (also known as endpoints) of interest. Briefly, a high quality RCT needs to utilize the following principles (11) :

1. Proper randomization (allocation to the particular drug is done by a probabilistic random process that ensures balance for known and unknown factors).
2. Minimize bias (patients, clinicians and outcome evaluators should be unaware of the treatment allocation to the particular drug, and an adequate comparator group should be used).
3. Intention to treat analyses (the treatment efficacy is based treatment allocated rather than treatment received - that is the statistical analyses are performed according to the randomized groups regardless of adherence).
4. Minimize the number of data-driven subgroup analyses (analyzing many subgroups can produce chance results)

3.1 RCT useful metrics

Relative risks (RR) are independent of the prevalence of the disease and can be applied to populations with different prevalence of the disease. The RR is the ratio of the risks in the treatment group to the event rate in the control group. The absolute risk reduction (ARR) is the difference between the treatment group event rate and the control group event rate. The ARR and the derived metric $1/ARR$ known as the number needed to treat (NNT) vary with the prevalence of the disease. NNT is the number of patients needed to treat to prevent one adverse event, and the converse is known as the number needed to harm. (12)

When reading the results of a study the reader will need to consider the point estimate and confidence interval reported in the manuscript. **Figure 2** shows the hypothetical results of three intervention studies. Study A suggests that the intervention has no effect (i.e. the true relative risk is 1) and is very precise (i.e., the confidence interval is narrow). You can be confident that it is not missing an important difference. Study B suggests that the intervention has no effect (i.e., the true RR is 1) but is very imprecise (i.e., the confidence interval is wide). This study may be missing an important difference. An investigator should be worried about type II error, but this study is just as likely to be missing an important harmful effect as an important beneficial one. Study C suggests that the intervention has a potentially clinically important beneficial effect (i.e., the true RR is much less than 1) and is also very imprecise. A large part of the confidence interval includes potentially clinically important beneficial effects. As a consequence of this an investigator might be concerned that a type II error is very likely. This is a study that should be repeated using a larger sample.

FIGURE 2 HERE

Care should be taken when interpreting non-statistically significant results such as those of Study B in Figure 2. It is quite common for investigators to confuse “absence of evidence of effectiveness” with “evidence of absence of effectiveness”.(13) For example, suppose a small trial is conducted to investigate the effect of a drug on death or dependency after

stroke. Nine of 20 patients treated with drug are dead or dependent at follow-up compared with 10 of 20 untreated patients, giving a p-value of 0.8. The absolute treatment effect is 5% (10/20 minus 9/20) with a (95% CI: -24%, 33%). Based on this it is plausible that the drug could cause either substantial harm or substantial benefit, so regardless of the p-value the conclusion should be that there is still uncertainty about whether the drug works and a much larger trial is required in order to answer this research question reliably.(14)

3.2 Size of an RCT

Given the side effects, costs and inconveniences of a treatment, the minimal clinically important difference (MCID) is the smallest treatment efficacy that would lead to a change in a patient's management. The MCID is crucial in both planning of clinical trials and interpretation of their results. Sample size of a study is extremely important because if a study is too small then it is unlikely that the investigators will be able to detect a modest, but clinically important, difference. (10) Conversely, if a study is too large then time and resources will be wasted. Prior to the commencement of a study the investigators should therefore spend time carefully deciding on a feasible sample size required in order to meet the goals of the study. The required sample size for a study depends on four criteria:

1. The MCID that you wish to detect for the endpoint of interest.
2. The required power of the study (usually set at 80% or 90%).
3. The required significance level (usually set at 5%).
4. The variance of the endpoint or outcome of interest.

The power of a test is $1 - \text{probability of a Type II error (i.e. } 1 - \beta)$. There are many different formulae for computing the required sample size which depend on the design of the study and the nature of the endpoint under study (e.g. time to event or dead or alive). (15) For both standard errors and confidence intervals it is possible to increase precision by collecting a larger sample. Larger studies provide more precise estimates of effect size than small trials, and they may allow a few sensible and predefined subgroup analyses. Small studies, with wide confidence intervals around the effect size estimate are likely to be clinically uninformative (although they may add to a meta-analysis of all similar studies or generate

enthusiasm for further investigation in a larger study). (14) In a systematic review of the literature Hislop et al.(16) found that there were a variety of methods available for specifying the MCID target difference in an RCT sample size calculation. The choice of approach was dependent on the aim (e.g., specifying an important difference versus a realistic difference), context (e.g., research question and availability of relevant data), and underlying framework adopted for the statistical analysis (e.g., classical statistical approach versus Bayesian methods). They concluded that no single method provided an ideal solution for all contexts.(16)

4. Generalizability or external validity of an RCT

The practicing clinician would like to know if the results of an RCT are applicable to their patients. The results are likely to be “clinically relevant” if the patients included in the trial were similar to those who would be treated in practice. Thus if a treatment has only been tested in men aged under 40 then it is generally impossible to know if it will be of benefit to a woman aged over 65. (17) Usually, trials with broad inclusion criteria are more generalizable than those with very strict criteria. It should be possible by examining the inclusion and exclusion criteria and the baseline characteristics in the published trial report whether the trial sample is representative of the people a clinician wishes to treat in their own work.(18) However, it is important that the main emphasis of a trial results should be placed on the overall effect not on particular subgroups in the trial, especially if these subgroups have not been pre-specified beforehand.(10, 14, 19)

4.1 Threats to the generalizability of an RCT

There is often concern about the generalizability of trials done in secondary or tertiary care to practice in primary care. Differences in the healthcare system can affect external validity. Even if the healthcare systems are similar other national differences can still affect generalizability. (20) For example with cerebrovascular disease, there are many important differences between countries in methods of diagnosis and management, as well as

important racial differences in susceptibility to disease and natural history of the disease, all of which could affect the external validity of the trial results. (17, 20) Selection of participating centres from secondary care as opposed to primary care has obvious implications for external validity, but RCTs of interventions that are confined to secondary care may also be undermined if they are restricted to specialist units. RCTs that have been conducted in a particular country will usually be generalizable to others, but this generalizability should not be taken for granted.(17, 20)

4.2 Outcome measures and follow-up

The external validity of an RCT also depends on whether the outcomes are clinically relevant. This can depend on who actually did the measurements, but is mostly dependent on what was measured and when. (20) RCTs sometimes use questionnaires, scales or indexes that are often a mixture of symptoms and clinical signs. It is thus important to use validated and reliable questionnaires, scales or indexes (with good repeatability and reproducibility) with clear definitions of what can be learned from absolute changes in the scale.(21) Many RCTs use composite outcome measures that can sometimes combine events of different severity and treatment effects can be driven by the least important outcome, which are often the most frequent events. (20) This is usually the situation in trials of stroke that include transient ischaemic attacks in a composite outcome. Inadequate duration of treatment and/or follow-up can also have an effect on the external validity of a trial.(11) For example patients with Alzheimer's disease may require treatment for many years; most RCTs of new drugs in this area follow up the effects of treatment only for a few months or years. (22)

4.3 Stopping a clinical trial early

Some trials are planned to be much larger than the size of the study that is reported in the final published manuscript. Trials may end up smaller because they are stopped early due to an apparently substantial treatment benefit (stopping for efficacy). The results of such

studies may need to be treated with a degree of scepticism because if they had been allowed to continue the final estimated treatment effect may well have been much smaller, so the reported estimate may be an overestimate of the treatment effect.(23, 24) In the very early stages of a trial, the treatment effect tends to fluctuate between extreme values before becoming more settled as more data accrues. Thus trials that stop early may have just stopped on a random high or low. (14) Obviously, there can be strong ethical reasons why a trial may stop early such as unexpected harmful effects (or no apparent emerging benefits [stopping for futility]). However, the trial data monitoring committee needs to think carefully about the possibility that early harm may be outweighed by longer term benefits later as more information becomes available. (14, 25-27)

5. Approaches to assess clinical significance

In classical statistical inference statistical significance measures how likely is it that the observed differences in outcome between treatment and control groups are real and not due to the play of chance. Clinical significance measures how large the differences in effect size (e.g. relative risk, mean difference or some other metric of interest) are in clinical practice. There have been several attempts in the literature to try and combine statistical and clinical significance in such a way the trial results can be interpreted more easily. These approaches are now described in more detail.

5.1 Using minimal clinical important difference (MCID) in order to assess clinical significance

Man-Son-Hing et al.(28) have suggested that clinical importance can take four forms, depending on the relationship of the minimal clinical important difference (MCID) of the intervention to the point estimate (the best single value of the efficacy of the intervention that has been derived from the study results) and the 95% CI surrounding it:

- a. ***Definite***: when the MCID is smaller than the lower limit of the 95% CI;

- b. **Probable:** when the MCID is greater than the lower limit of the 95% CI, but smaller than the point estimate of the efficacy of the intervention;
- c. **Possible:** when the MCID is less than the upper limit of the 95% CI, but greater than the point estimate of the efficacy of the intervention; and
- d. **Definitely not:** when the MCID is greater than the upper limit of the 95% CI.

These concepts are now illustrated using some examples from the contemporary neurological literature.

5.1.1 Study result of definite clinical importance.

A study will have results of *definite clinical importance* if the lower limit of the 95% CI is greater than the MCID. Given that the 95% CI is greater than the MCID estimate studies with definite clinical importance will always be statistically significant ($p < 0.05$). An example is the FREEDOMS II trial (29) that investigated the efficacy and safety of fingolimod (an oral sphingosine-1-phosphate receptor modulator, that has shown reductions in clinical and MRI disease activity in patients with multiple sclerosis). The primary objective was to assess whether fingolimod 0.5mg per day was superior to placebo in reducing annualized relapse rates in patients with relapsing-remitting multiple sclerosis treated for up to 24 months. The authors report that the MCID was to detect a relative annualized relapse rate reduction of 40% (i.e. they could detect a rate ratio of 0.60). Three hundred and fifty-eight patients were randomized to fingolimod and 355 to receive placebo. Mean annualized relapse rate was 0.40 (95% CI 0.34, 0.48) in patients given placebo and 0.21 (95% CI: 0.17, 0.25) in patients given fingolimod 0.5mg. This corresponded to a rate ratio of 0.52 (95% CI: 0.40, 0.66) and a p-value of <0.0001 for fingolimod versus placebo.

Another example of such a result that was statistically significant is from the MEMREHAB trial, a double-blind placebo-controlled trial of 86 patients with definite multiple sclerosis (MS).(30) The objective was to examine the efficacy of the modified Story Memory Technique (mSMT), a 10-session behavioural intervention teaching context and imagery to

facilitate learning, to improve learning and memory in person with MS. Participants completed neuropsychological assessment at baseline, a repeat assessment immediately post-treatment and a long-term follow-up assessment at 6 months. After completion of the treatment phase, persons were assigned to receive or not receive booster sessions, in order to assess the efficacy of monthly booster sessions. The authors reported a MCID of 0.13 in the learning slope for the California Verbal Learning test between the treatment and control group. The treatment group showed a significantly improved learning slope difference when compared to the placebo group post-treatment of 0.54 (95% CI: 0.21, 0.87) with an associated p-value of 0.021.

5.1.2 Study result of probable clinical importance.

The results of a study that is of *probable clinical importance* will have 95% CIs that include the value of the MCID, with the point estimate being greater than the MCID estimate. The study results may or may not be statistically significant. An example from the neuroepidemiological literature is the trials of Amyotrophic Lateral Sclerosis (ALS), a progressive degenerative disease that is characterized by weakness in the limbs and bulbar muscles. The first trial in this area by Bensimon et al. (31) used a drug called riluzole in order to assess improved survival at 12 months in ALS compared with placebo. The authors reported a MCID of 30% improved survival at 12 months for patients with bulbar-onset disease for the basis of their sample size calculations. They found that 12 month survival was 35% in the placebo group versus 73% in the riluzole group (absolute reduction increase in survival of 38%), which was statistically significant with a p-value of 0.014. The same trial also investigated patients with limb-onset ALS and found no significant improvement in survival at 12 months ($p=0.17$). There was concern about the disproportionate benefit in the patients with bulbar-disease onset compared to lumbar disease-onset ALS. As this trial was relatively small (~ 150 patients) a much larger trial that varied the dosage of riluzole was carried out and this also found a small but statistically significant prolongation of survival in participants receiving the intermediate and high dose of riluzole. (32) Some patients were

ineligible for the large trial and as a consequence another trial of rituzole was conducted (in parallel with the larger study) with patients with either more advanced ALS or aged over 75 years as the patient population. This study was not statistically significant, but the study was not designed to assess efficacy, so it did not demonstrate a survival advantage of rituzole. However, although not statistically significant the result was of *probable clinical importance* as it did demonstrate that rituzole was well tolerated in this patient population, and that the adverse events were similar to those observed in the larger definitive trials. (33)

5.1.3 Study result of possible clinical importance.

Studies with results of *possible clinical importance* have 95% CIs that include the value of the MCID, and an MCID greater than the efficacy point estimate. The results may or may not be statistically significant. An example of a study result that is of *possible clinical importance that was **not** statistically significant* is the URICO-ICTUS study. (34) This study aimed to assess whether uric acid therapy would improve functional outcomes at 90 days in patients with acute ischaemic stroke. The primary outcome was patients with an excellent outcome [i.e., a modified Rankin (mRS) score of 0-1, or 2 if premorbid score was 2] at 90 days. The MCID was estimated to be a 14% difference between uric acid and placebo. Two-hundred and eleven patients received uric acid and 200 received placebo, of these 83 (39%) that received uric acid and 66 (33%) that received placebo had an excellent outcome. The absolute treatment difference observed was 6% with an adjusted risk ratio was 0.81 (95% CI: 0.64, 1.04); p=0.099.

An example of study result that is of *possible clinical importance that **was** statistically significant* is the TEAM-AD VA Cooperative randomized trial (35). This study aimed to investigate whether vitamin E (alpha tocopherol), memantine, or both slow progression of mild to moderate Alzheimers disease (AD) in patients taking an acetylcholinesterase inhibitor. The trial involved 613 patients with mild to moderate AD and 152 participants received alpha tocopherol , 155 received memantine, 154 the combination of memantine and alpha

tocopherol, and 152 placebo. The MCID was a 4 point mean treatment difference in the Alzheimer's disease Cooperative study Activities of Daily Living (ADCS-ADL) between either treatment given alone vs placebo. The MCID was also estimated as a 20% reduction in the annual rate of AD decline. Participants receiving alpha tocopherol had a mean change difference compared with placebo of 3.15 (95% CI: 0.92, 5.39); p-value after adjustment was 0.03. This change translates into a delay in AD progression of 19% per year compared with placebo.

5.1.4 Study results that are definitely not clinically important

Studies that produce results that are *definitely not of clinical importance* have the upper limit of the 95% CI that is below the MCID. Again the study results may or may not be statistically significant. An example of a study result that was *definitely not clinically important* in the neurological literature that was also **not statistically significant** is the ALIAS trial (36) that aimed to assess whether albumin given within 5 hours of the onset of acute ischaemic stroke increased the proportion of patients with a favorable outcome. The MCID reported by the authors was a 20% absolute difference between thrombolysis and non-thrombolysis strata. Four hundred and twenty-two patients were randomized to receive albumin and 419 to receive saline. The trial was stopped early for futility with 814 patients recruited into the study. The primary outcome did not differ between the patients in the albumin and saline groups [186 (44%) vs 185 (44%)], risk ratio 0.96 (95% CI: 0.84, 1.10).

An example of a statistically significant result that was *definitely not of clinical importance* was the GISSI-Prevenzione trial that assessed the effects of dietary supplementation with polyunsaturated fats on death, non-fatal myocardial infarction and non-fatal stroke. (37) The MCID was considered to be a 4% absolute reduction in deaths for the composite endpoints between the groups over a 3.5 year period. The study actually observed a 1.3% (95% CI: 0.1%, 2.6%) absolute reduction in the primary endpoint.

5.2 A potential new metric to assess clinical significance

As has already been described, p-values and 95% CI can be used to help determine how likely observed effects are on the basis of chance. Walsh et al. (38) have proposed that in the case of RCTs with dichotomous outcomes a shift of only a few events in one group could change typical hypothesis tests above the usual thresholds considered statistically significant (i.e. $p\text{-value} < 0.05$). They suggest a new approach that they feel can better communicate the limitations of p-value thresholds as this new metric demonstrates how easily significance based on a threshold p-value may be exceeded. They refer to this metric as a fragility index (FI). The FI helps to identify the number of events required to change statistically significant results to non-significant results.

We will illustrate the use of this metric using the GISSI-Prevenzione trial results. The study aimed to compare the effects of polyunsaturated fats (n-3-PUFA) versus placebo on the composite endpoint of death, non-fatal myocardial infarction and non-fatal stroke in 11,324 patients. (37) The number of events in the n-3-PUFA group was 715 (21.6%) versus 785 (13.9%) in the placebo group with a RR of 0.90 (95% CI: 0.82, 0.99) and a corresponding value of 0.05. The results of the trial are presented as a two-by-two table with the number of events over the entire follow-up being used to construct the table (**Table 1a**).

TABLE 1A HERE

The FI is calculated by adding an event from the group with the smaller number of events (and subtracting the nonevent from the same group to keep the total number of patients constant) and recalculating the two-sided p-value for Fisher's exact test.(38) **Table 1b** shows the RR after adding 3 events to the n-3-PUFA event group (and subtracting 3 events from the No Event group to keep patient numbers constant). The RR is still 0.90 but the Fisher's exact p-value is 0.06. So it only took 3 events to change this result from significant to non-statistically significant at the conventional 5% level of significance. The importance of

the FI can be appreciated if consideration is given to RCTs where initial reports showed statistically significant effects that were later shown to be either substantially less than previously reported or ineffective.(39)

TABLE 1B HERE

5.3 A Bayesian approach to clinical significance

Burton et al.(40) describe a Bayesian approach to clinical significance. They report that their approach is most useful in situations in which a conventional classical statistical approach to the analysis may be difficult or misleading. They suggest that their approach includes circumstances in which: (a) a statistically non-significant result is large enough to be clinically relevant (small sample size); (b) a statistically significant result is too small to be of clinical relevance (very large sample size); or (c) it is desired that conclusions are drawn about the probable similarity of two outcomes without concluding that non-significant means that there is no difference. Although this Bayesian approach was first described almost two decades(40) ago it has not received widespread use in the neurological literature, (probably due to the difficulty in eliciting prior information), where classical statistical approaches to statistical inference dominate in published reports.

5.4 A graphical approach to clinical significance

The level of treatment effect regarded as clinically significant also depends on the severity of the disease and any potential side effects of the treatment. As described earlier in this chapter, a common strategy to assess combined statistical and clinical significance is to report an appropriate metric (for example, RR, ARR, or NNT) with its associated 95% CI. The levels of treatment effects regarded as clinically worthwhile are likely to differ among clinicians and settings. If one clinician considers that introduction of the new treatment is worthwhile only if the actual risk is reduced by 15%, say, it is important to know how likely the clinical trial observations would have arisen by chance with the null hypothesis that the new treatment has an ARR of less than 0.15 and not with the null hypothesis that the two

treatments are equally effective. This leads to the Bayesian approach by Burton et al. (40) reported in the previous section. As an alternative, Leung (41) proposed a plot of p value-ARR or p value-NNT which he believes would be useful to the clinicians who practise evidence based medicine. The Leung approach assumes that the null hypothesis is that the ARR for the new treatment is less than some value x . p-values are calculated for a range of values of x . These p-values are plotted on the vertical axis and the ARR on the horizontal axis. Hence, for a range of values of ARR, the plot will show the corresponding probability that the clinical trial observations would have arisen by chance if the real ARR were less than the given values. (41) This approach was described over a decade ago and as with the Bayesian approach (40) it has not been used at all in the neurological literature where classical statistical approaches still prevail in the interpretation of trial results.

CONCLUSION

This chapter has given a general overview of the differences between statistical and clinical significance has been outlined in the context of RCTs of neurological disorders. A description of the practical design issues that can affect the external validity and subsequently the clinical relevance of an RCT (such as the choice of the study population, outcome measures used, length of follow-up, sample size, and the stopping a trial early) have also been provided. In addition, approaches that use the minimally important clinical difference with examples from the contemporary neurological literature have also been discussed in order to facilitate interpretation. Modest treatment effects on mortality or major morbidity are generally more plausible than large effects. It has been often the case that large and often striking effects from small-scale randomized trials (and other types of designs including non-randomized studies) will often be refuted. (10, 42) This implies that even highly statistically significant (e.g., 2-sided p-values of 0.001), differences that are based on only relatively small numbers of events in selected studies may provide untrustworthy evidence of the existence of any real clinically important difference. Finally,

when there is a lack of good evidence for any effect on major outcomes, estimates of the NNT to prevent such outcomes are of little or no value, and it is particularly important to provide confidence intervals around the NNT(12) in order that the range uncertainty can be ascertained. For this reason, based on classical statistical inference, statistically significant results with claims of large effects based on small randomized trials should be treated with caution by practicing clinicians until the clinical relevance can be reliably assessed.

Table 1a: Results of the GISSI-Prevenzione Trial Results

	Event	No Event
n-3-PUFA	715	4951
Control	785	4883

RR =0.90 (0.82, 1.00) Fisher's exact test =0.05

Table 1b: Calculated Fragility of the GISSI-Prevenzione Trial Results

	Event	No Event
n-3-PUFA	718	4948
Control	785	4883

RR =0.90 (0.81, 1.01) Fisher's exact test =0.06

Figure 1: Relationship between P-value and 95% confidence interval

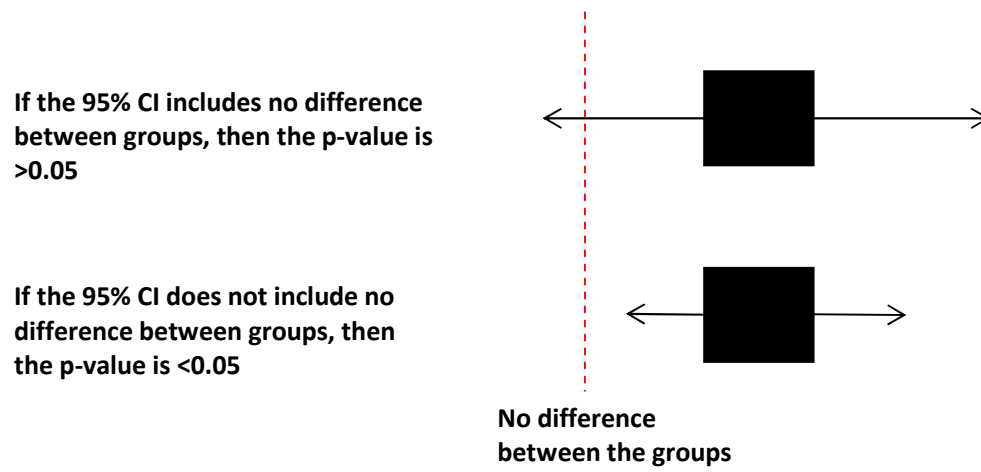


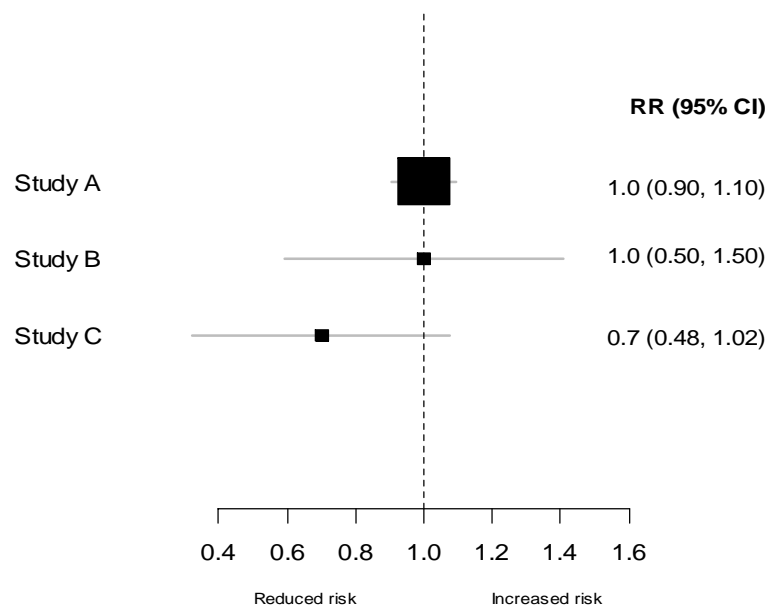
Figure 2: Role of 95% CIs in assessing type II errors

Figure Legends

Figure 1

Redrawn and adapted from “Primer on 95% confidence intervals”, Effective Clinical Practice, September/October 2001; 4: 229-231.

Figure 2

Redrawn and adapted from “Primer on Type 1 and Type II errors”, Effective Clinical Practice, November/December 2001; 4: 284-285.

REFERENCES

1. Bland JM, Bland DG. Statistics Notes: One and two sided tests of significance. *BMJ*. 1994;309(6949):248.
2. Sterne JAC, Cox DR, Smith GD. Sifting the evidence—what's wrong with significance tests? Another comment on the role of statistical methods. *BMJ*. 2001;322(7280):226-31.
3. Gardener M, Altman D. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ*. 1986;292(6522):746 - 50.
4. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clinical Trials*. 2005;2(4):295-300.
5. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to bayesian methods in health technology assessment. *BMJ*. 1999;319(7208):508-12.
6. Ibrahim JG, Chen M-H, Chu H. Bayesian methods in clinical trials: a Bayesian analysis of ECOG trials E1684 and E1690. *BMC Medical Research Methodology*. 2012;12(1):1-12.
7. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. *Annals of Emergency Medicine*. 22(8):1328-36.
8. Lilford RJ, Braunholtz D. For Debate: The statistical basis of public policy: a paradigm shift is overdue. *BMJ*. 1996;313(7057):603-7.
9. Carlin BP, Louis TA. *Bayesian methods for data analysis*. Boca Raton, FL: Hall/CRC; 2008.
10. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. *The Lancet*. 2001;357(9253):373-80.
11. C.Baigent, R.Peto, R.Gray, S.Parish, R.Collins. Large-scale randomized evidence: trials and meta-analyses of trials. Oxford, UK: 'Oxford University Press'.
12. Altman DG. Confidence intervals for the number needed to treat. *BMJ*. 1998;317(7168):1309-12.
13. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485. Epub 1995/08/19.
14. Lewis SC, Warlow CP. How to spot bias and other potential problems in randomised controlled trials. *Journal of Neurology, Neurosurgery & Psychiatry*. 2004;75(2):181-7.
15. Wittes J. Sample Size Calculations for Randomized Controlled Trials. *Epidemiologic Reviews*. 2002;24(1):39-53.
16. Hislop J, Adewuyi TE, Vale LD, Harrild K, Fraser C, Gurung T, et al. Methods for Specifying the Target Difference in a Randomised Controlled Trial: The Difference ELicitation in TriAls (DELTA) Systematic Review. *PLoS Med*. 2014;11(5):e1001645.
17. Altman DG, Bland JM. Generalisation and extrapolation. *BMJ*. 1998;317(7155):409-10.
18. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet*. 2002;359(9308):781-5. Epub 2002/03/13.
19. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-86. Epub 2005/01/11.
20. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*. 2005;365(9453):82-93. Epub 2005/01/11.

21. Bland JM, Altman DG. Validating scales and indexes. *BMJ*. 2002;324(7337):606-7.
22. Kryscio RJ. Secondary prevention trials in alzheimer disease: The challenge of identifying a meaningful end point. *JAMA Neurology*. 2014.
23. Guyatt GH, Briel M, Glasziou P, Bassler D, Montori VM. Problems of stopping trials early. *BMJ*. 2012;344:e3863. Epub 2012/06/19.
24. Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA : the journal of the American Medical Association*. 2010;303(12):1180-7. Epub 2010/03/25.
25. Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, et al. Issues in data monitoring and interim analysis of trials. *Health Technol Assess*. 2005;9(7):1-238, iii-iv. Epub 2005/03/15.
26. Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of clinical epidemiology*. 2008;61(3):241-6. Epub 2008/01/30.
27. Chalmers I, Altman DG, McHaffie H, Owens N, Cooke RW. Data sharing among data monitoring committees and responsibilities to patients and science. *Trials*. 2013;14:102. Epub 2013/06/21.
28. Man-Son-Hing M, Laupacis A, O'Rourke K, Molnar F, Mahon J, Chan KY, et al. Determination of the clinical importance of study results. *J GEN INTERN MED*. 2002;17(6):469-76.
29. Calabresi PA, Radue E-W, Goodin D, Jeffery D, Rammohan KW, Reder AT, et al. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. *The Lancet Neurology*. 2014;13(6):545-56.
30. Chiaravalloti ND, Moore NB, Nikelshpur OM, DeLuca J. An RCT to treat learning impairment in multiple sclerosis: The MEMREHAB trial. *Neurology*. 2013;81(24):2066-72.
31. Bensimon G, Lacomblez L, Meininger V. A Controlled Trial of Riluzole in Amyotrophic Lateral Sclerosis. *New England Journal of Medicine*. 1994;330(9):585-91.
32. Amyotrophic Lateral Sclerosis/Riluzole Study G, II, Lacomblez L, Bensimon G, Meininger V, Leigh PN, Guillet P. Dose-ranging study of riluzole in amyotrophic lateral sclerosis. *The Lancet*. 1996;347(9013):1425-31.
33. Bensimon G, Lacomblez L, Delumeau JC, Bejuit R, Truffinet P, Meininger V. A study of riluzole in the treatment of advanced stage or elderly patients with amyotrophic lateral sclerosis. *Journal of neurology*. 2002;249(5):609-15.
34. Chamorro Á, Amaro S, Castellanos M, Segura T, Arenillas J, Martí-Fàbregas J, et al. Safety and efficacy of uric acid in patients with acute stroke (URICO-ICTUS): a randomised, double-blind phase 2b/3 trial. *The Lancet Neurology*. 2014;13(5):453-60.
35. Dysken MW, Sano M, Asthana S, et al. Effect of vitamin e and memantine on functional decline in alzheimer disease: The team-ad va cooperative randomized trial. *JAMA : the journal of the American Medical Association*. 2014;311(1):33-44.
36. Ginsberg MD, Palesch YY, Hill MD, Martin RH, Moy CS, Barsan WG, et al. High-dose albumin treatment for acute ischaemic stroke (ALIAS) part 2: a randomised, double-blind, phase 3, placebo-controlled trial. *The Lancet Neurology*. 2013;12(11):1049-58.
37. Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. *The Lancet*. 1999;354(9177):447-55.

38. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *Journal of clinical epidemiology*. 2014;67(6):622-8.
39. Ioannidis JA. COntradicted and initially stronger effects in highly cited clinical research. *JAMA : the journal of the American Medical Association*. 2005;294(2):218-28.
40. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. *Journal of Epidemiology and Community Health*. 1998;52(5):318-23.
41. Leung WC. Balancing statistical and clinical significance in evaluating treatment effects. *Postgraduate medical journal*. 2001;77(905):201-4. Epub 2001/02/27.
42. Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med*. 2005;2(8):e124.