South Afr J Anaesth Analg. 2022;28(5 Suppl 1):S69-71
https://doi.org/10.36303/SAJAA.2022.28.5.2913

South Afr J Anaesth Analg
ISSN 2220-1181 EISSN 2220-1173

FCA REFRESHER COURSE

Clinical vs statistical significance

B Biccard

Department of Anaesthesia and Perioperative Medicine, Groote Schuur Hospital, University of Cape Town, South Africa
Corresponding author, email: *bruce.biccard@uct.ac.za*

Keywords: clinical significance, statistical significance

Background

In 1991 a prospective observational study of 48 500 women from the Nurses' Health Study was published.[1] It documented nearly 340 000 person-years of patient follow-up, and showed that oestrogen supplementation in postmenopausal women was associated with an adjusted relative risk (RR) of 0.56 (95% confidence interval [CI] 0.4–0.8) for major coronary disease and similar risk reduction for fatal cardiovascular disease. The following year a quasi-meta-analysis of 35 epidemiological studies showed a reduction in coronary heart disease, stroke and hip fracture, but a significant increase in breast cancer.[2] This scare led to a randomised controlled trial (RCT) where patients were randomised to oestrogen and progesterone or placebo, with outcomes of coronary artery disease and invasive breast cancer.[3] This trial showed that hormone replacement therapy was associated with an increase in coronary heart disease events, strokes and breast cancer and a reduction in hip fractures. The recommendation was that hormone replacement therapy should not be used as a primary prevention.[3] In a 10-year period, there was a turnaround in findings from protection to harm with hormonal replacement therapy in postmenopausal women. How did this happen?

Clinical research and finding the truth

The introductory story is a result of the *'ill-founded strategy of claiming conclusive research findings solely on… statistical significance, p-value < 0.05'.*[4] Indeed, what we should be doing is considering our ability to predict the 'truth' from our clinical research.[4] John Ioannidis famously stated that 'most published research findings are false.'[4] To find the truth we need to understand: i) the principles of hypothesis testing, and ii) the 'positive predictive value' of a study.

Hypothesis testing

Hypothesis testing is based on the dichotomous decision to accept the null hypothesis, or reject the null hypothesis in favour of the alternative hypothesis.[5] The null hypothesis significance testing is based on the *p*-value. A significance threshold is known as the α and it is usually set at 0.05.[5] This 'statistically significant' threshold merely describes the probability of the observed difference. At 0.05, there is a 5% chance of the observed difference occurring by chance.

There are three misconceptions associated with the assumptions of the null hypothesis and significance testing.[5]

Misconception 1

A nonsignificant result demonstrates that there is no effect. This is not true. Rather it simply indicates that there is insufficient evidence against the null hypothesis needed to accept the alternative hypothesis.[5]

Misconception 2

Lower *p*-values increase the 'significance' of the findings. This is not true. The importance is that the threshold was set *a priori* commonly for a *p*-value of 0.05 for a binary (yes/no) cut-point. Only the threshold *p*-value is what matters, as it is a binary decision. It is important to realise that larger sample sizes result in smaller *p*-values for the same observed effect. Therefore, there is no such term as 'highly significant'. This is illustrated in Table I.

Table I: The effect of the sample size on the *p*-value and the 95% confidence interval

Control group risk	Experimental group risk	RRR	*p*-value	RRR 95% CI
2/4	1/4	50%	1.00	-174–92%
10/20	5/20	50%	0.19	-14–80%
20/40	10/40	50%	0.04	10–73%
50/100	25/100	50%	0.0004	27–66%

RRR – relative risk reduction

Misconception 3

The null hypothesis is true. This is also a misconception. The *p*-value is calculated assuming that the null hypothesis is true, and it is therefore not the probability that the null hypothesis is true.

Finding the truth

The goal of our research should be to determine the 'truth'. The problem however is that we do know what the 'truth' is. John Ioannidis suggests that we need to consider the positive

predictive value (PPV) of a study, in order to understand how likely it is to predict the 'truth'. This is explained by the classic 2x2 contingency table shown in Table II.

Table II: A contingency table to describe the positive predictive value for the 'truth'

	The 'truth'	**'Not true'**
Research 'positive'	True positive (TP) = power X pretest probability of the 'truth'	False positive (FP) = α/ ratio of 'truth' to 'not true'
Research 'negative'		

PPV – TP/FP

True positive (TP) = power x pretest probability of the truth

(The power of a test is the probability of obtaining a significant result, e.g. detecting a real difference, if it exists. A 'type II error' or beta (ß) occurs when the null hypothesis is incorrectly accepted. It is concluded that there is no statistically significant difference between groups when a statistically significant difference does exist. The probability of avoiding a type II error is referred to as the power of the study. Type II errors occur when the sample size is too small for a clinically important difference to reach statistical significance.)

False positive (FP) = α/ratio of 'truth' to 'not true'

$$PPV = \frac{power\ x\ pretest\ probability\ of\ the\ 'truth'\ x\ ratio\ of\ 'truth'\ to\ 'not\ true'}{\alpha}$$

How are these components of the PPV reflected in studies?[4]

1. Bias. This is reflected by the ratio of 'truth' to 'not true' results. The greater the flexibility in designs, definitions, outcomes, and analytical models in a scientific field, the less likely the research findings are to be true. Bias distorts the relationship between variables. The estimated effect size therefore does not necessarily reflect the true effect in the population.
2. The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true. This often leads to publication bias due to studies publishing positive findings.
3. The 'hotter' a scientific field (with more scientific teams involved), the less likely the research findings are to be true. This often leads to publication bias due to studies publishing positive findings.
4. The greater the number of tested relationships in a scientific field, the less likely the research findings are to be true. It is important to remember that each time a relationship is tested, there is a 1 in 20 chance of a false positive statistical finding.
5. Studies of a smaller size with a positive finding, are less likely to be true.

To improve the likelihood of finding the truth we need to:[4]

1. Increase the power of the study. This will decrease the risk of a false negative statistical finding.
2. Decrease the bias. This is why we register studies, e.g. on clinicaltrials.gov, why we produce statistical analysis plans prior to analysis, and use the EQUATOR guidelines when reporting.
3. Only test major concepts. Major concepts are where the physiology, pathology and pharmacology support a consistent theory of potential intervention efficacy.
4. Conduct trials where the pre-study probability of success is already high.
5. Don't emphasise statistically significant findings by a single team.

If we follow these principles, then we will produce studies with results which are closer to the truth.

Clinical significance

Only once we are delivering good studies, and trials can we start to consider if a study result is of any clinical significance. This is because hypothesis testing (and statistical significance) does not tell us about: i) The magnitude of the effect, nor ii) the precision of the estimated magnitude of that effect.[5] These two factors are the foundations of potential clinical significance.

The magnitude of the effect

The RR tells us about the magnitude of the effect as shown in Table III.

Table III: Relative risk or magnitude of effect

Exposure	**Outcome**	
	Yes	**No**
Yes	a	b
No	c	d

Risk with exposure = a/(a+b)

Risk without exposure = c/(c+d)

$$\text{Relative risk} = \frac{a/(a+b)}{c/(c+d)}$$

An RR of < 1 for an adverse outcome in treated patients suggests potential benefit. Yet, it is the magnitude of this effect, and its precision, which will determine if it is clinically important. Conversely, an RR of > 1 for an adverse outcome in treated patients suggests potential harm associated with the therapy. Yet, it is the magnitude of this effect, and its precision of this effect which will determine if it is indeed clinically important.

As clinicians we decide on the magnitude of the effect that we consider clinically important. In cardiovascular outcomes, we often expect an RR reduction of approximately 25% before considering a therapy for secondary prevention effective. If we are in primary prevention, we would expect a vaccine to have an RR reduction in the order of 90% before we consider it to be clinically important.

The point estimate of the clinical effect is the result from the study. The estimated effect size (or point estimate) however does not necessarily reflect the true effect in the population. However, the true effect lies across the spectrum of the 95% CI. This spans

from the greatest possible effect to the smallest possible effect of an intervention.

Precision of the magnitude of effect

The 95% CI presents the precision of the magnitude of effect. In clinical trials, we determine outcomes as superiority, equivalence, non-inferiority and harmful.

Superiority

When a test is statistically significant in the desired direction, then we can conclude (statistical) superiority. This is a two-sided superiority test statistically significant in the desired direction, and the CI that does not contain zero. However, if we get a statistically insignificant result, we cannot assume equivalence as the study may not have been powered to show the difference (false negative) or we may have a false positive.[6]

Equivalence

This is when the CI for the difference between groups falls within the a priori defined equivalency region. Equivalence is claimed only if the treatment difference is concluded to be both significantly above the lower limit and significantly below the upper limit.[6]

Non-inferiority

The goal is to show that the intervention is at least as effective as the standard (i.e. equivalent or superior to the standard treatment).[6] Here, equivalence is not expected and superiority not needed, non-inferiority is a one-sided equivalency design that tests the null hypothesis that the preferred treatment is worse than the comparator.

Inferiority (or harm)

This is 'statistically inferior' to the control, and the CI does not include zero.

These principles are illustrated graphically as follows (Figure 1). Note that we need to define a minimal clinically important effect, for either benefit or harm to determine clinical importance. For superiority, this would mean that the smallest potential risk reduction cannot cross this boundary, and for inferiority the converse.

Importantly, when the CI contains a clinically important effect, a clinically significant effect cannot be ruled out, irrespective of the statistical significance. This can be seen both on the side

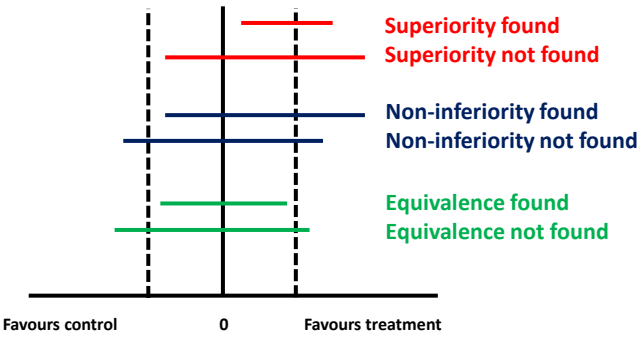

Figure 1: Interpreting the 95% confidence interval with respect to superiority, equivalence and non-inferiority

of treatment efficacy and harm. If the CI does not contain an important clinical effect, it is unlikely to have clinical significance, as is considered equivalent.

Conclusion

We need to strive to conduct research (and evaluate research) according to the PPV. Where PPV is low (usually evidence from observational studies, or small, biased RCTs), we then provide our best clinical practice based upon our understanding of potential theoretical benefits, based upon our understanding of risk-benefit relationships. In this scenario, we cannot be unreasonably dogmatic about our practice, as the evidence base is flimsy. Should there be a clinical trial in this area, we should participate. We need to create a culture of participation in large clinical trials (where if the trial is of limited bias, then the PPV will be high), as this is the way in which we will build evidence with a high PPV and get us closer to the 'truth'.

ORCID

B Biccard https://orcid.org/0000-0001-5872-8369

References

1. Stampfer MJ, Colditz GA, Willett WC, et al. Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the nurses' health study. N Engl J Med 1991;325(11):756-62. https://doi.org/10.1056/NEJM199109123251102.
2. Grady D, Rubin SM, Petitti DB, Fox CS, Black D. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med 1992;117(12):1016-37. https://doi.org/10.7326/0003-4819-117-12-1016.
3. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 2002;288(3):321-33. https://doi.org/10.1001/jama.288.3.321
4. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.
5. Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: What do *P* values and confidence intervals really represent? Anesth Analg. 2018;126(3):1068-72. https://doi.org/10.1213/ANE.0000000000002798.
6. Mascha EJ. Equivalence and noninferiority testing in anesthesiology research. Anesthesiology 2010;113(4):779-81. https://doi.org/10.1097/ALN.0b013e3181ec6212.