When should readers care about p-values, power and confidence intervals?

Gordon S. Doig

G. S. Doig,

Northern Clinical School Intensive Care Research Unit, University of Sydney, Sydney, Australia, 2006.

Corresponding author:

Dr. Gordon S. Doig,

Mailing Address:

Royal North Shore Hospital,

Intensive Care Unit, Level 6,

Pacific Hwy,

St Leonards, NSW

Australia 2065

Telephone: +61 2 9463 2633

Facsimile: +61 2 9463 2057

Email: gdoig@med.usyd.edu.au

Word count: 1,398

Clinical Scenario: You are discussing one of your ICU patients with a colleague. This patient was admitted to your ICU after uneventful surgery for an abdominal aortic aneurism, and now they look like they may be developing sepsis. When the conversation turns to your plans for red cell transfusion, your colleague says "I don't follow the recommendations for a more conservative transfusion threshold (7 g/L Hb) because a recent clinical trial conducted in septic patients failed to show a significant difference between conservative and liberal transfusion thresholds.[1] Plus, a second trial reported a significant *increase* in mortality when a conservative transfusion threshold was used after cardiac surgery![2]" As a result of this conversation, you make a commitment to do some reading to find out which approach to practice is most appropriate for your patient.

Published clinical trials can be divided into three main categories based on their results: **1)** trials that demonstrate one therapy is superior to another; **2)** trials that demonstrate one therapy is equivalent to another or; **3)** trials that are unable to demonstrate either superiority or equivalency. Since the first two categories can inform clinical decisions, it is important to be able to identify these main types of trials. The purpose of this short review is to address *statistical issues* that can help differentiate between all three main categories.

1) *Superiority*

The essential first step towards demonstrating one therapy is superior to another is to establish that there is a *statistically significantly difference* between the treatment outcomes. Statistical significance is established using an appropriately calculated p-value.

A p-value can be thought of as the *probability* the treatment effects observed in the clinical trial arose due to random chance.[3] The smaller the p-value, the more likely that

"either an exceptionally rare chance has occurred or the theory [of no difference in treatment outcomes] is not true."[4]

Establishing the presence of a statistically significant difference in treatment outcomes is only the first step towards identifying a superiority trial that can inform practice decisions. In addition to statistical significance, we must also consider whether the *magnitude* of the difference in treatment outcomes is plausible and meaningful to the patient,[5] whether prior evidence (physiological or clinical) supports the hypothesised mechanism of action and whether all key risks and costs associated with the proposed new treatment are known and acceptable.[6,7]

2) *Equivalency*

Only the *magnitude of the difference* in treatment outcomes needs to be considered to establish *equivalency*. More specifically, the study needs to rule out a *minimally clinically important difference* (MCID) between the therapies under comparison.

The MCID is defined as the smallest difference in the outcome measure (systolic blood pressure, kidney function, risk of death etc) which would be beneficial to the patient and which would mandate, in the absence of concerning side effects or excessive costs, a change in the patient's management to the more favourable therapy.[5] Under this definition, any treatment effect *larger* than the MCID can be considered meaningful to the patient's health. The magnitude of an MCID for a specific outcome measure can be defined using research evidence, clinician consensus and/or consultation with patient groups. Consultation with patients becomes important when risk/benefit trade-offs are subtle, such as the values and preferences surrounding the increased risk of bleeding related stroke vs. reduced risk of ischemic stroke when considering antithrombotic therapy for patients with atrial

fibrillation.[8] MCIDs for some key outcomes from critical illness have been established by clinician consensus.

Based primarily on clinician consensus, the MCID for survival from critical illness has been defined as a three percent absolute improvement in the chance of being alive at follow-up after hospital discharge. In other words, if two different therapies are shown to have *less than* three percent difference in mortality, clinicians are likely to accept the therapies as being *equivalent* with regards to their impact on survival.[9] So, how can the results of a clinical trial rule out the presence of a meaningful treatment effect, *greater than* the MCID?

The *95% confidence interval* is a statistical construct that provides a reasonable and reliable estimate of the *minimum* and *maximum* expected treatment effects.[10] For example, whilst the 7,000 patient saline vs. albumin fluid evaluation (SAFE) study reported an *average* mortality difference between groups of -0.20% (20.9% vs. 21.1% mortality in patients receiving albumin vs. saline, respectively) the 95% confidence interval around this risk difference ranged from -2.1% to 1.8%. The lower confidence level rules out any benefit greater than 2.1% attributable to albumin and the upper confidence level rules out any benefit greater than 1.8% attributable to saline. Thus, because we used a 95% confidence interval, we are 95% certain that if any real differences between albumin and saline exist, the differences are not greater than the accepted MCID. As long as we accept the pre-defined threshold of three percent absolute difference in mortality to be a reasonable threshold for an MCID, we can consider albumin *equivalent* to saline.

3) Neither superiority nor equivalency

The majority of clinical trials fall into this category. These trials fail to demonstrate superiority *and* cannot rule out the possibility of a meaningful treatment effect because the

95% confidence interval is too wide. Although these types of trials cannot inform clinical decision making based on their own findings, they may be able to contribute useful information to a meta-analysis.[11]

So, what about power?

Power can be thought of as the *probability* a treatment effect of a pre-specified magnitude will be detected by a clinical trial with a set number of participants. Power is primarily useful *before* a clinical trial begins, to support planning. After a clinical trial is completed, power does not provide any useful clinical information that cannot be obtained from the combination of the p-value and 95% confidence interval.[10,12] Post hoc power calculations may be useful for planning future research, but they should not be used to inform clinical decisions.

Odds Ratios, Relative Risk, Hazard Ratios or Risk Difference?

The main results of a clinical trial, the outcome difference between two therapies, can be presented using a measure of *relative* effectiveness (odds ratio, risk ratio or hazard ratio) or *absolute* effectiveness (risk difference or number needed to treat). Although the math behind the calculation of p-values and confidence intervals to assess measures of *relative* effectiveness is well established, clinicians prefer *absolute* measures when making clinical decisions.[13] Given today's modern computing platforms, exact chi-square tests, exact logistic regression, extensions of classic regression and test-based approximations can always be used to calculate p-values and confidence intervals for absolute measures. Many simple web-based tools are also available to help calculate absolute measures of effect, and confidence intervals, if the published study only reports relative measures (See www.EvidenceBased.net/files for some simple Excel tools).

Back to our clinical scenario

Figure 1 reports the absolute risk differences and confidence intervals for the results of the transfusion in critical care trial,[14] the transfusion in sepsis trial,[1] and the transfusion after cardiac surgery trial [2]. Of these three clinical trials, only the transfusion after cardiac surgery trial reports a significant p-value for mortality (p=0.045 for 90 day mortality), thus establishing *superiority* of a liberal transfusion threshold in this patient population. Calculation of the 95% confidence interval around the absolute risk difference reveals it ranges from -3.3% to 0.0% in favour of liberal transfusion, thus whilst the effect is statistically significant, the benefit is only marginally greater than the MCID mortality threshold of 3%. Unfortunately, the results of the transfusion in sepsis trial do not help resolve this issue: The p-value is non-significant (p=0.44 for 90 day mortality) and the 95% confidence interval is wide (-4% to 8%). The information provided by these two new trials is best interpreted in the context of a meta-analysis. Fortunately a meta-analysis on this topic has been undertaken.[15]

Using a comprehensive search to identify clinical trials conducted in critically ill patient populations, the authors of this meta-analysis conclude there is no significant mortality effect (p=0.52). However we can use simple tools to calculate the 95% confidence interval around the absolute risk difference reported in the meta-analysis. Given a 0.7% absolute effect on mortality (586/4154 - 558/4167 patients), the conservative 95% confidence interval ranges from -1.3% to 2.8% and effectively rules out any difference greater than the 3% accepted MCID. Thus, we can be 95% certain that there is no meaningful mortality effect, and our treatment decision can be based on costs and patient preferences!

Conflict of interest:

None.

Reference List

1. Holst LB, Haase N, Wetterslev J, Wernerman J, Guttormsen AB, Karlsson S et al. (2014) Lower versus higher hemoglobin threshold for transfusion in septic shock. N Engl J Med 371:1381-1391

2. Murphy GJ, Pike K, Rogers CA, Wordsworth S, Stokes EA, Angelini GD et al. (2015) Liberal or restrictive transfusion after cardiac surgery. N Engl J Med 372:997-1008

3. Biau DJ, Jolles BM, Porcher R (2010) P-value and the theory of hypotheis testing: An explanation for new researchers. Clin Orthop Relat Res 468:885-892

4. Fisher RA (1959) Statistical methods and scientific inference. 2 ed. Oliver and Boyd, Edinburgh

5. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 10:407-415

6. Guyatt GH, Sackett DL, Cook DJ (1993) Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 270:2598-2601

7. Guyatt GH, Sackett DL, Cook DJ (1994) Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 271:59-63

8. Alonso-Coello P, Montori VM, Diaz MG, Devereaux PJ, Mas G, Diez AI et al. (2015) Values and preferences for oral antithrombotic therapy in patients with atrial fibrillation: physician and patient perspectives. Health Expect 18:2318-2327

9. Finfer S, Bellomo R, Boyce N, French J, Myburgh J, Norton R (2004) A comparison of albumin and saline for fluid resuscitation in the intensive care unit. N Engl J Med 350:2247-2256

10. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S (1995) Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. CMAJ 152:169-173

11. Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K et al. (2014) How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. JAMA 312:171-179

12. Goodman SN, Berlin JA (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 121:200-206

13. Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N (1995) Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. CMAJ 152:351-357

14. Finfer S, Chittock DR, Su SY, Blair D, Foster D, Dhingra V et al. (2009) Intensive versus conventional glucose control in critically ill patients. N Engl J Med 360:1283-1297

15. Holst LB, Petersen MW, Haase N, Perner A, Wetterslev J (2015) Restrictive versus liberal transfusion strategy for red blood cell transfusion: systematic review of randomised trials with meta-analysis and trial sequential analysis. BMJ 350:h1354. doi: 10.1136/bmj.h1354.:h1354

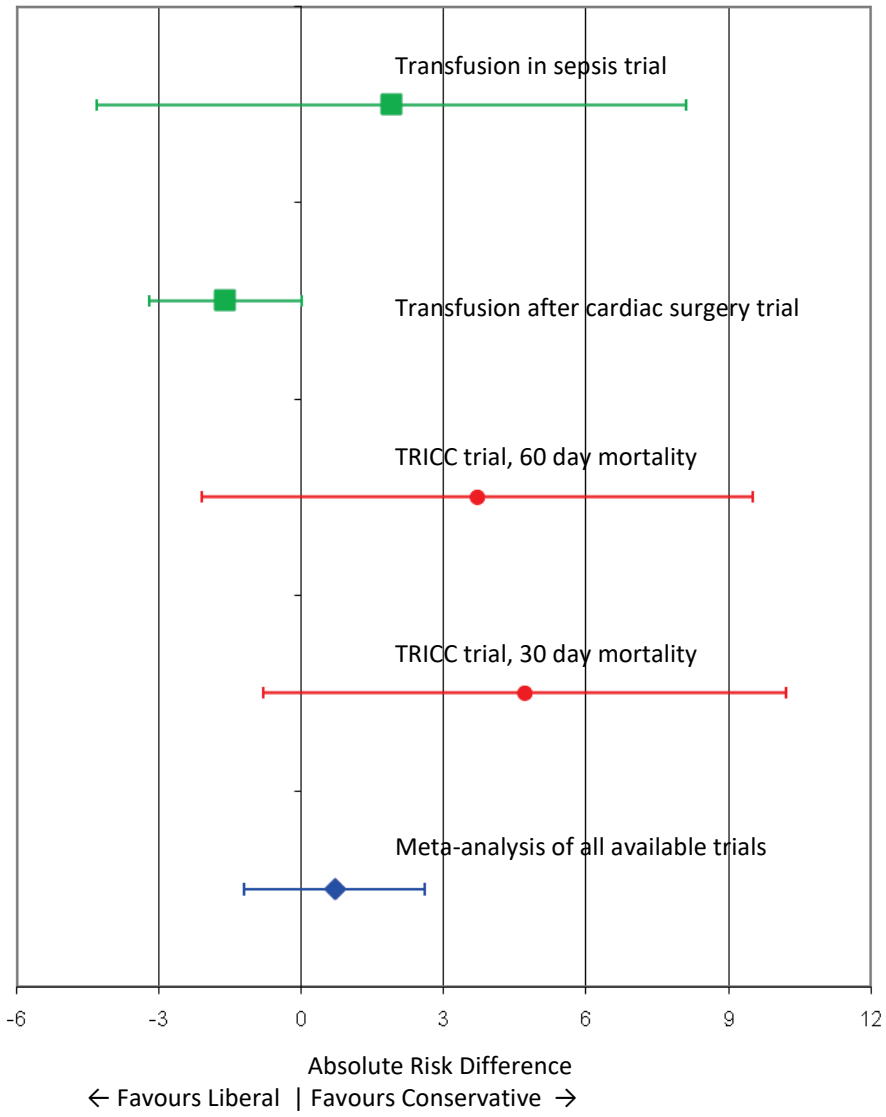
P-value: The P-value of a test of significance refers to the probability that the test will reject the Null Hypothesis (the assumption there is no difference between treatments) when the Null Hypothesis is *true*. Since the concept of a p-value is predicated on the assumption that a real treatment effect *does not* exist, the p-value is a conditional probability.

95% Confidence Interval: The use of 95% confidence intervals is derived from classical frequentist statistical theory, and is an extension of the p-value threshold of 0.05 proposed by RA Fisher. The confidence interval is an 'interval estimate', and attempts to estimate the lowest and highest expected treatment effect. Technically, the 95% confidence interval is interpreted as indicating that if the experiment were run 100 times, the true estimate of effect would fall within the bounds set by the confidence interval in 95 out of 100 of each theoretical repeats.

Power: The power of a test of significance refers to the probability that the test will reject the Null Hypothesis (the assumption there is no difference between treatments) when the Null Hypothesis is *false*. Since the concept of power is predicated on the assumption that a real treatment effect *does* exist, power is a conditional probability.

Text Box 1: Technically correct definitions for p-value, 95% confidence interval and power.

Figure 1. Mortality and 95% confidence intervals for transfusion trials.



Legend: Negative (< 0) Risk Difference, Favours *liberal* transfusion; **Positive (> 0) Risk Difference**, Favours *restrictive* transfusion; **Blue stippled zone,** *No clinically important mortality effect*. Defined using SAFE study *equivalency* sample size estimate: 90% power to detect a 3% absolute risk difference.