Editorial
The abuse of Power, P-values and the Misnomer of Statistical "Testing"

In estimating the sample size in the design of a randomised controlled trial (RCT) certain assumptions are necessary. Critical amongst these is an educated guess as to how much difference the intervention might be expected to make and whether this will be of clinical importance. Typically, investigators will estimate a likely value for the defined primary outcome in the control group, and then decide what a clinically meaningful difference would be in the intervention group as the "effect size". Roughly speaking, the larger the difference or effect size that the intervention is anticipated to make, the smaller the sample required for the study.

In this edition of SAJAA Nontshe and colleagues closely compared estimated Minimum Clinically Important Difference ("MCID") reported by authors in planning their sample sizes with the actual treatment effect found in the published study.[1] This was done for all parallel group inequality RCTs with a single primary outcome published in the top five anaesthetic journals in 2014. It is surprising that only 28 papers made the final cut (of ~ 500 manuscripts published across those prestigious specialty journals in that year). However, the key point highlighted is that in all but 20% of the trials the actual treatment effect was substantially smaller than the estimation used in the apriori sample size calculations. In almost half of the trials the authors had estimated that the likely treatment effect would be double what was eventually reported. Moreover, a handful of the studies used one-sided P-values which reduce the sample size further. The use of one-sided hypothesis testing for superiority RCTs should be avoided and is rarely appropriate.[2] In one of those one-sided hypothesis investigations the actual treatment effect was in the opposite direction to what was expected.

The implication is that many of these studies were doomed to be underpowered from the outset. The issue is that researchers manipulate the power calculation to get a manageable sample size so that the trials will be approved and funded. Whilst the process of power and sample size calculations does encourage and allow researchers to balance expectations, this should not be at the cost of performing meaningful research with realistic and clinically meaningful outcomes. At worst this suggests that many clinical investigators prioritise "getting research done" over producing sound scientific methodology. If true, then this is a major problem – accumulation of published evidence of variable quality actually hinders our understanding. It appears that a lot of patients may have been exposed to an intervention without anything useful being learned.

The inference however illustrates a common misconception in medical literature: a study that does not return a statistically significant result *is not a failed trial*. The purpose of a trial is to produce clinical data to reject the null hypothesis and accept the alternative of inequality (between the groups) in the vast majority of situations. On occasion, other alternative hypotheses for superiority, non-inferiority and equivalence are examined, the latter two being examples of the recommended use of one-sided hypothesis testing. The trial is not somehow to simply validate the authors` original estimate of the treatment effect. Put simply, the statistics are not the ultimate aim of the study, the data is.

Let us examine the power or sample size calculation in a bit more detail: power is "the probability of a study of a given sample size correctly identifying a true difference or effect in the populations studied". It is also the probability of rejecting a false null hypothesis.[3] To perform the sample size calculation for a trial, investigators must first define the primary endpoint. They should also define in their prospective statistical analysis plan whether the primary endpoint is to be characterised in proportions or as continuous or ordinal data. So, this might be a change in prevalence of a categorical outcome (say, incidence of a defined complication after surgery) or a change in the average (or median) value of a continuous endpoint such as peak serum creatinine or length of hospital stay. They then set the required power (typically 80 or 90%) and level of significance ("the probability of rejecting the null hypothesis in a statistical test when it is in fact true") usually 5%.[4] Next, they draw on previous published studies or pilot data of their own to estimate a likely event rate or average value (and the typical distribution around this value) for the control group.

Finally they need to make an estimate as to how much difference the intervention might make, that is, to define a Minimum Clinically Important Difference. The study then sets out to measure the likelihood that the intervention does in fact make *this much* difference.. Nontshe`s description of these steps in the paper somewhat blurs the distinction between the two different common sample size calculations (for continuous data and proportions, although both can be harmonised by framing them as "standardised differences") but their overall message is that there are two crucial elements:

1. Do the statistics support the notion that the intervention makes a difference?

2. *How much* difference?

We should all be asking 'How much of an effect is there?' rather than simply 'Is there an effect?'[5] This second element requires judgement on the part of the investigators and the reader as to whether this amount is plausible and actually important in clinical practice. It seems imperative that investigators give their sample size calculation, as Nontshe et al. put it, "thoughtful and realistic consideration."[1] Yet in 13 of the 28 trials detailed in their review, no justification for the estimated treatment effect was provided in the methods section. Only two studies explicitly cited pilot or observational data. Nontshe and colleagues do not describe how much effort was made to contact the original authors to find out more about the source of the MCID estimation. It is of course possible that some of the 13 in fact drew on pilot data, and that sufficient detail might be found in the protocols of these investigations, but this should be describedalso in the published study report.

Twelve (12) of the included RCTs cited previously published studies to justify their apriori assumption about effect size. Differences in inclusion criteria, intervention and endpoints might affect applicability of these to the authors' own cohort. Full disclosure helps the clinician reading the paper to put the data in context, rather than simply focus on the mathematics of the P-value or confidence interval.

The evidence based way that medical practice evolves is that new (or old) interventions are put to the test, usually against current practice (control) in a clinical trial. Statistics is a kind of mathematical language to help us understand probability, which can be applied to data to make

some sense of it. Indeed the terms 'statistical test' and 'test statistic' are actually misnomers. It is hypotheses that are tested not statistics. We use statistical methods to estimate probabilities and confidence intervals and then check if is there enough evidence to accept or reject the hypothesis under test. Indeed, these authors would prefer if we deleted the term 'test' from virtually all the usual statistical tests and use terminology such as analysis of variance (ANOVA), Kruskal-Walllis one-way analysis, Student t-distribution, correlation and regression, etc. The problem is that very few doctors understand the language. As it turns out, most authors of scientific papers in anaesthetic journals do not understand it either. Statisticians themselves appreciate that they are applying probabilistic models rather than measuring degrees of certainty, but these concepts seem difficult for clinicians to grasp. This is not a new insight. A comprehensive misunderstanding of statistics is rife in the medical literature and has prompted something of a print backlash from statistical experts and journal editors.[5-8]

Some journals have even placed an outright ban on P-values, which, unfortunately, is akin to 'shooting the messenger'. Fisher, the founder of the P-value intended it only as an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look.[5] Whilst reassuringly simple, it has instead somehow been misappropriated into a threshold which is applied rigidly to data – usually along the lines of "if the P-value is < 0.05 then you have found something significant (and if not, you have not). "This stems from what behavioural economists call "fast" thinking: an endeavour to rapidly weigh up available information and come up with a summary overall judgement which is usually rather black or white.[9] Such a decisive approach is often helpful in the practice of clinical medicine. Unfortunately, when applied to critical appraisal of scientific literature it is wholly misleading. The problem is not with the P-value, it is with the lack of understanding of the reader.

What is a P-value? It is the probability that the difference or effect size observed (or one more extreme) will occur under the null hypothesis. Consider an example from Nuzzo`s excellent 2014 essay in Nature on the slippery nature of statistics[5]: Data from a large study seemed to show that extremists quite literally saw the world in black and white. Political moderates saw shades of grey much more accurately than did people with extreme left or right wing views. The P-value was 0.01. Many scientists looking at that would interpret it as showing a highly significant finding, very unlikely to be due to random chance.

But they would be wrong. The P-value cannot say this. All it can do is summarise the data in relation to a specified null hypothesis. The P-value here says that if it is true there is no difference between how accurately extremist and moderates see colours, then there is a 1% chance of having observed this result. This is not at all the same as saying there`s only a 1% chance that it is a false alarm. If we were to repeat the whole experiment under similar conditions there is a substantial chance of getting a different result. Indeed the investigators in Nuzzo`s real life example decided to repeat the study and this time the P-value came out at 0.59, not even close to the conventional level of significance. This is an example of what statisticians call a "replication" problem. The crux is that the statistical value by itself cannot make statements about the underlying reality. What is missing is another piece of information: the probability that a real effect was there in the first place. In clinical studies this is a matter to be addressed by clinical experts with knowledge of the clinical area being addressed by the trial. An effect has got to be plausible, or else it does not matter how many zeros there are after the decimal point in the P-value.

Since statistical analyses are usually misinterpreted, one may wonder what if anything these analyses do for science.[8] Indeed, there is a ground swell of clinician investigators who suggest that because medicine is so complex, it is just about impossible to construct experimental models which control all circumstances that might affect outcome apart from the study intervention itself, and we should therefore abandon classic scientific trials as a means to advance knowledge.[10] In response, others explain that it is called evidence-based medicine for a reason.[11]

There are however some ways of reporting and reading clinical trials that might help. To avoid the trap of thinking about results as only significant or not significant, for example, if researchers always report effect sizes and confidence intervals then the reader can readily see the magnitude and relative importance of an effect.

It is useful also not to ignore all previous research. Bayesian analysis is a mathematical way of doing this, and is becoming more prominent. In essence, one decides by separate means, incorporating what is already known (and with a certain degree of subjectivity) what the probability is that the null hypothesis is true or false, and then considers how much the result of new evidence (in one direction or another) affects one's continuing belief in this likelihood.

Perhaps most useful is the "effect size ratio", a quick calculation to help us judge claims of inequality (semantically "superiority" is commonly but erroneously used – this is a one-sided hypothesis) based on a finding in a clinical trial.[12]

$$\textit{effect size ratio} = \frac{\textit{point estimate of observed effect}}{\textit{MCID specified in the a priori power calculation}}$$

Figure 1. Effect Size Ratio. MCID = minimum clinically important difference

If the effect size ratio is > 1, then you are seeing something where the effect is at least as large as the investigators estimated before the study. In contrast if < 1 then by the authors own criteria the observed effect is too small to be important. The effect size ratio is independent of the statistical analysis. It is also a quick way to check the methodological rigour of the study, as it can only be applied to primary outcomes (it relies on the MCID from the original sample size calculation). This therefore eliminates all secondary endpoints from consideration and so guards against data trawling. Further, since it is a ratio, it requires that numerator and denominator have the same units. Thus when reporting the primary outcome investigators have to stick with the same variable in the same metric as was specified in the study protocol. Only 10 (36 %) of the trials in Nontshe`s review[1] potentially have an effect size ratio > 1, although for some the calculation might still not be possible, if for example the primary outcome was reported as a proportion (e.g. number of patients with severe postoperative pain) when the protocol specified a value (e.g. average pain score).

The main limitation of the effect size ratio is that the authors may not have set a realistic MCID. If an artificially low value was used then an effect size ratio > 1 might not be meaningful, although, unless the sample size is very large, smaller MCIDs are less likely to produce statistically significant differences in the primary endpoints.

We should point out that an effect size ratio < 1 does not imply bad science. It is part and parcel of experimental research that little or no

effect may be found, despite expectations. This gets back to our original point that the whole purpose of conducting a trial is to measure the size of a previously unknown intervention effect, not to somehow help the investigators toward a significant result on a statistical analysis.

So then what is new in this article? Earlier commentaries have also highlighted miscalculation of power.[13] This well constructed paper by Nontshe and colleagues serves simply to show that it is still rife. It is really the job of journal editors and expert reviewers to make sure that when there are obvious discrepancies between sample size calculations, effect sizes and power, these are addressed during the review process and where relevant commented upon as limitations in the discussion section of the study report. It would of course be best if the authors paid more attention to MCID estimation in the first place. Ultimately readers themselves should look closely at the data and judge the apparent study findings in the light of their plausibility, how they sit with what was previously already known and whether this has implications for their own clinical practice.

Malachy O Columb,[1] Gary Minto [2]

[1] Consultant in Anaesthesia & Intensive Medicine, Acute Intensive Care Unit, Manchester University Hospitals NHS Foundation Trust, Wythenshawe, United Kingdom M23 9LT
[2] Consultant in Anaesthesia, University Hospitals Plymouth NHS Trust, & Honorary Associate Professor , Plymouth Univesity Peninsula Schools of Medicine & Dentristy, Plymouth, United Kingdom PL6 8DH

Correspondence to gary.minto@nhs.net

References

1. Nontshe M, Khan S, Mandebvu T, Merrifield B, Rodseth R. Sample Size Determination in randomised controlled trials published in anaesthetic journals. SAJAA. 2018; 24(2):40–46. https://doi.org/10.1080/22201181.2018.1439602
2. Columb MO, Polley LS. Potencies and probabilities: One-sided P values suggest a one-sided story. Anesthesia & Analgesia. 2001;92:278-9.
3. Columb MO, Stevens A. Power analysis and sample size calculations. Current Anaesthesia & Critical Care. 2008;19:12-4.
4. https://www.merriam-webster.com/dictionary/. Accessed on 28 Feb 2018
5. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506:150-2.
6. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2:e124.
7. Wasserstein RL, Lazar NA. The ASA's Statement on P-Values: Context, process, and purpose. American Statistician. 2016;70:129-33.
8. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337-50.
9. Kahneman D. Thinking Fast and Slow. London: Penguin, 2011.
10. Keane MJ, Berg C. Is science the answer? Br J Anaesth. 2017;119:1081-4
11. Moppett IK, Pearse RM. Evidence-based medicine: the clue is in the name. Br J Anaesth. 2017;119:1084-6.
12. Gibbs NM, Weightman WM. A forcing strategy to improve the evaluation of clinical superiority in anaesthesia trials. Br J Anaesth. 2016;117:281-3.
13. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005;365:1348-53.