Ophthalmic statistics note 6: effect sizes matter

Jonathan A Cook,[1] Catey Bunce,[2] Caroline J Doré,[3] Nick Freemantle,[4] on behalf of the Ophthalmic Statistics Group

WHY STATISTICAL SIGNIFICANCE ALONE IS NOT A SUFFICIENT BASIS TO INTERPRET THE FINDINGS OF A STATISTICAL ANALYSIS

Medical statistics plays a key role in clinical research by helping to avoid errors of interpretation due to the play of chance. It is, however, critical to understand the limits of what statistical analysis provides and interpret the findings accordingly. Statistical analysis can summarise the statistical evidence but it cannot tell us whether a difference is important per se, as clinical judgement is needed. Suppose we have a clinical trial which compares two drugs for reducing the intraocular pressure (IOP) in the eye and there is evidence of a statistically significant difference in favour of drug A. Does this mean A is superior to B? Perhaps, if the average difference is 5 mm Hg but what if it is only 0.1 mmg Hg? We would be very unlikely to conclude clinical superiority under these circumstances, or at least it would be necessary to take account of differences in other key outcomes such as adverse events before reaching such a conclusion. It might be thought that the p value from a statistical hypothesis test can serve this purpose but this is not the case. Solely considering whether a p value is 'statistically significant' is not a sufficient basis to interpret the findings of a statistical analysis. There is a number of reasons why this is the case.[1 2]

- First, when a statistical hypothesis test is performed we are usually implicitly testing for evidence of a difference of *any* magnitude. To equate this difference to clinical importance is to view all statistically significant differences as clinically important and of equal value (at least in a practical sense). Clearly, as in the IOP example above, this may not be the case.
- Second, the p value is not a direct measure of the strength of evidence for the null hypothesis. Instead, it is the probability of obtaining a result as, or more extreme than, the one observed assuming the null hypothesis is true. It is therefore, at best, a partial and indirect measure of the evidence regarding the truth of the null hypothesis. In particular, it is not the probability that the null hypothesis is true given the data.
- Third, the criterion used for statistical significance, typically a p value of less than 0.05 (ie, 5% significance level), is arbitrary. It is based on convention rather than statistical theory. It guards against making one type of error too often (rejecting the null hypothesis when it is true), but gives no protection against making the other type of error (not rejecting the null hypothesis when it is false).
- Fourth, even once we meet our criteria we must take into consideration the magnitude of an effect. It's all fine and well saying we have found a significant difference, but it is crucial to consider how big this difference is and whether we view it as clinically important. We should design our study such that the difference which would be statistically significant would also be considered to be clinically important. It is also true, as covered in an earlier BJO statistics note,[3] that when we don't find a statistically significant difference, we must consider how precise our estimate is (the width of our CI) and what magnitude of effect would be important to us if it did truly exist. A p value neither quantifies the size of an observed effect nor its importance.

EFFECT SIZE MATTERS

Wherever possible, a statistical method should be used that can estimate the magnitude of the observed effect (an effect size) along with the associated uncertainty around this estimate, rather than only carrying out a statistical (hypothesis) test and presenting the corresponding p value. In the commonest situation of comparing two independent groups, a t test and corresponding CI for the difference in means is preferred, provided its underlying assumptions hold, over a non-parametric alternative such as a Mann–Whitney U test (it should be noted that this non-parametric test does not compare the group means or medians but the two distributions). Effect measures differ according to the type of outcome often with several alternatives available. Common effect measures for binary outcomes include differences in proportions, risk ratios or ORs. A HR (which may be interpreted similarly to a relative risk) can be used for time-to-event data. These effect measures are often natural products of a particular statistical analysis (eg, a Cox regression analysis for time-to-event data estimates HRs, logistic regression for a dichotomous outcome estimates ORs). We note in passing that a Bayesian statistical approach offers a theoretically appealing alternative to conventional statistical testing which avoids the use of p values although not without difficulties both in principles and practice.[2 4]

The importance of considering the effect size is illustrated in table 1 where the results of four hypothetical trial scenarios are presented. The presence of any ocular adverse event (eg, eye irritation, dryness, eyelash growth) was compared between two IOP-lowering drugs in patients with open-angle glaucoma or ocular hypertension. Comparing drug A with drug B in scenarios 1, 2 and 3 using Fisher's exact test gives fairly similar p values (between 0.026 and 0.125). However, two of these p values are smaller than 0.05 and one is not. If we use the conventional cut-off of p=0.05 for statistical significance, we would conclude that there was a statistically significant difference between drugs A and B in scenarios 1 and 3, but no difference in scenario 2. If we look at the effect sizes, we see that scenarios 1 and 2 have an observed difference of over 20% with similar CI widths although the first does not contain a difference of zero whereas the second does. Clearly, the strength of evidence is similar for these scenarios, with the caveat of statistical significance not being shown in the second (if the 5% significance level is adopted) but to ignore the size of the effect would be foolish. On a practical level, it is dangerous to rely solely on a p value and censoring p values which are above a cut-off (eg, p=0.054 being replaced by 'NS') is uninformative and potentially misleading. Indeed

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, Centre for Statistics in Medicine, Oxford, UK; [2]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK; [3]UCL Comprehensive Clinical Trials Unit, University College London, London, UK; [4]Department of Primary Care and Population Health, University College London, London, UK

Correspondence to Dr Jonathan A Cook, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, Centre for Statistics in Medicine, Oxford OX3 7LD, UK; jonathan.cook@ndorms.ox.ac.uk

Table 1 Ocular adverse event rates from hypothetical randomised controlled trial scenarios comparing a 12-week course of two IOP-lowering drugs for open-angle glaucoma

Scenario	Drug A n/N (%)	Drug B n/N (%)	Difference in percentages (95% CI)	p Value
1	1/30 (3)	8/30 (27)	−23 (−41 to −5)	0.026
2	4/30 (13)	10/30 (33)	−20 (−40 to 2)	0.125
3	28/300 (9)	45/300 (15)	−6 (−11 to −0.4)	0.045
4	6/300 (2)	9/300 (3)	−1 (−4 to 2)	0.603

Calculations were carried out in Stata[6] using command csi and rdcii; the p value is from a two-sided Fisher's exact test and the CI for the percentage difference is from Newcombe's Method 10[7].
IOP, intraocular pressure; n, number of adverse events; N, number of patients.

p values pose many challenges to interpretation which are mitigated through the use of point estimates and CIs of effect sizes.[5]

The trials in scenarios 3 and 4 are much larger (10 times the size) and therefore have greater statistical precision and are able to detect substantially smaller effects. If we were only to consider the p value in scenario 1, we would conclude that the drugs are different, but this provides no information about the magnitude of the difference—which in this instance is considerable. If we compare scenarios 1 and 3, we see that the p values are similar but the effect sizes are very different (23% vs 6%). If we compare scenarios 3 and 4, we see that both have much tighter CIs with a difference in the rate of adverse events that would be viewed as 'clinically important' likely to be ruled out in scenario 4.

Interpreting the result of a statistical analysis relies upon more than statistical understanding. It is helpful to differentiate between 'statistical significance' and 'clinical importance'. If there is statistical evidence of an effect, we wish to know if it matters. Conversely, where there is no statistical evidence of an effect, we wish to know if a clinically important effect has been ruled out. Calculating an effect size and a corresponding 95% CI provides the information upon which such judgements can be made. It can be remarkably difficult to know what is clinically important in a particular context. How easily clinical importance can be assessed varies by the setting, perspective adopted and outcome analysed. Quality of life measures are particularly difficult to interpret, whereas mortality is more straightforward. A variety of approaches of varying complexity have been proposed (including minimal clinically important difference approaches) which seek to provide a transparent, reproducible and robust way to achieve this.[8 9]

SUMMARY

The magnitude of an observed effect and the uncertainty regarding it should always be considered when interpreting statistical evidence. Whenever possible, a statistical analysis which produces an effect size estimate with associated uncertainty (eg, difference in means and its CI) rather than solely a p value should be used and reported.

Acknowledgements The post of CB is partly funded by the NIHR Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Collaborators Valentina Cipriani, David Crabb, Phillippa Cumberland, Gabriela Czanner, Paul Donachie, Andrew Elders, Marta Garcia Finana, Neil O'Leary, Rachel Nash, Ana Quartilho, Luke Saunders, Selvaraj Sivasubramaniam, Chris Rogers, Simon Skene, Irene Stratton, Wen X.

Contributors JAC drafted the paper, CB reviewed and revised the paper, CJD and NF conducted an internal peer review and provided additional comment.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

OPEN ACCESS

To cite Cook JA, Bunce C, Doré CJ, *et al*. *Br J Ophthalmol* 2015;**99**:580–581.

Published Online First 26 February 2015

Br J Ophthalmol 2015;**99**:580–581.
doi:10.1136/bjophthalmol-2014-306303

REFERENCES

1 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;292:746–50.
2 Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med* 1999;130:995–1004.
3 Bunce C, Patel KV, Xing W, *et al*. On behalf of the Ophthalmic Statistics Group. Ophthalmic statistics note 2: absence of evidence is not evidence of absence. *BJO* 2014;98:703–5.
4 Bland JM, Altman DG. Bayesians and frequentists. *BMJ* 1998;317:1151.
5 Wood J, Freemantle N, King M, *et al*. The trap of trends to statistical significance: how likely it really is that a near significant P value becomes more significant with extra data. *BMJ* 2014;348:g2215.
6 StataCorp. *Stata: Release 12. Statistical Software*. College Station, TX: StataCorp LP., 2011.
7 Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med* 1998;17:873–90.
8 Wells G, Beaton D, Shea B, *et al*. Minimal clinically important differences: review of methods. *J Rheumatol* 2001;28:406–12.
9 Cook JA, Hislop J, Adewuyi TE, *et al*. Assessing methods to specify the target difference for a randomised controlled trial—DELTAr (Difference ELicitation in TriAls) review.