

Statistical analyses which provide an effect size are to be preferred

Cook JA, Ranstam J.

Whenever possible a statistical analysis which provides a measure of the magnitude of the effect should be used. Unfortunately many commonly used statistical tests do not produce an estimate of the effect size. For example, χ^2 and Fisher's exact tests are often used to analyse a binary outcome (e.g. surgical complications). These tests assess how likely the observed results were to have occurred if there was no difference between the two groups; however they do this without estimating the magnitude of an effect. Fortunately alternative statistical methods are readily available which allow for more complex analyses, regression analyses (e.g. logistic or Poisson regression) or if an unadjusted difference is preferred calculation of the difference in the proportions between groups along with a corresponding 95% confidence interval. In the era of big data, it is more important than ever not to rely upon statistical significance alone as very large datasets allow small (sometimes tiny) effects to be statistically detected which are nonetheless clinically irrelevant. Different effect size metrics exist and can be used according to the outcome and the analysis method used (e.g. linear regression provides the mean difference, whereas Poisson regression can be used to estimate a risk ratio). It is important that the estimate of effect size along with its uncertainty (e.g. 95% confidence interval) is reported. Doing so also allows for an assessment of how precisely the quantity of interest has been measured.

Sometimes methods which do not produce an effect size estimate (e.g. Mann-Whitney) are used due to concerns regarding the assumption required for methods that do (e.g. an independent t-test may not be appropriate if data are not distributed normally). However while alternatives may be simpler to use and dependent upon a weaker set of assumptions, these analyses tend to be less informative; additionally they may also differ in what they test (e.g. Mann-Whitney U test quantifies evidence for a difference in the shape as well as the location). Even where normality cannot be plausibly assumed, statistical approaches which address this and enable a method which allows for an effect size estimate (e.g. bootstrapping) may still be possible.

Example: laparoscopic colectomy with and without natural-orifice specimen extraction

In a small randomised trial, 40 participants were randomised to receive laparoscopic colectomy with and without natural-orifice specimen extraction (ref Wolfsis BJS). The primary outcome of interest was use of additional analgesic (Piritramide) after surgery.

The original report provided the number of events in each group and reported the p-value from a Fisher's exact trial (1 of 20, 10 of 20 with/without NOSE respectively; p value of 0.003) showing statistical evidence of a difference in the use of additional analgesic between groups.

More informatively, a confidence interval for the difference in the proportions between treatment groups could have been calculated and reported: 0.45, 95% CI(0.18,0.66). This quantifies the observed effect size (0.45 difference in proportions i.e. 45% difference between groups) along with a plausible range of uncertainty (0.18 to 0.66 difference in proportions, or 18% to 66%). It can be seen from this analysis that the observed effect was very large and though there is substantial uncertainty about how large it is, nevertheless use of additional analgesic was estimated to be been at least 18% lower in this study.

Reference

Wolthuis et al, Randomized clinical trial of laparoscopic colectomy with or without natural-orifice specimen extraction *BJS* 2015; 102: 630–637

Word count 352

