# Making sense of the p-value (part 2)

In a previous editorial, I discussed what is probably the most ubiquitous (and possibly also the most misunderstood) statistic in the history of academic publishing: the *p*-value. In brief, the *p*-value helps us to determine whether we should reject a null hypothesis (of "no effect") in favour of an alternative or research hypothesis, by quantifying the conditional probability of getting the result you actually got (or a more extreme result), under the assumption that the null hypothesis is true. A low *p*-value (the usual cut-off is 0.05) indicates that such an event is unlikely to have been observed if the null hypothesis really is true – suggesting that the alternative explanation of events may be more plausible.

Use of the *p*-value is quite controversial: one reason is its interpretation in the situation when multiple comparisons are being made. These can arise in many ways: for example, where several outcome measures are under investigation; where the key treatment variable has 3 or more levels; or where separate analyses are conducted on sub-groups of individuals.

Consider the situation where a researcher conducts, say, 3 comparisons within a study (none of which are pre-defined as a primary outcome), reporting a *p*-value for each of them. If, as usual, values below 0.05 are used to denote statistical significance, then there is a 5% probability of falsely claiming a non-existent effect for each individual test (known as a Type I error); hence a 95% probability of making a correct call on any one particular test.

So far so good. But just as it is harder to achieve full marks in an examination with 3 questions than in an examination with 1 question, it is harder to avoid an error if we conduct three tests rather than just a single one. The probability of making the correct call three times out of three is $0.95^3$=0.857, and hence the probability of making at least 1 Type I error (a false claim of significance) is 1–0.857=0.143 (14.3%). This is known as the familywise error rate, and increases quite rapidly as more tests are conducted. By the time the number of tests reaches 14, the probability of making a Type I error is over 50%.

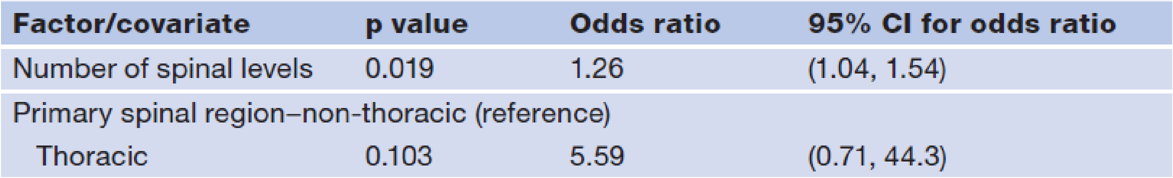
So you should be wary of claims made regarding isolated incidences of apparent significance from studies in which several comparisons are conducted. Many of these may be no more than chance findings, cherry-picked by study authors after results are known, with no pre-determined primary outcome.

A simple way of controlling error rates is the Bonferroni correction. This involves dividing the significance level by the number of tests to be conducted. So in a study involving, say, 10 comparisons, each one would be declared significant only if its *p*-value was below 0.05/10=0.005 (0.5%). This is a widely used method: in an evaluation of the functional properties of superabsorbent dressings and their effect on exudate management, Browning and White[1] used Bonferroni-corrected values to compare the absorbency of 5 dressings: pairwise comparisons of 5 dressings requires 10 individual comparisons. Babaei et al.[2] also applied the Bonferroni correction to their results in an investigation of the management of chronic diabetic foot ulcers in which patients were allocated to one of three groups, according to size of ulcer, and the significance of changes in time for wound healing were statistically assessed in each group. In both cases the authors' aim was to avoid inflating the likelihood of a Type I error that would have been associated with uncorrected comparisons.

However, the Bonferroni method has the disadvantage that it may result in a lot of true findings going unclaimed. This is particularly true when some of the tests being conducted are not independent of each other, or when additional speculative tests are conducted which the authors have no real reason to believe will lead to significant findings. Some authors prefer not to apply a formal correction to results but to instead present uncorrected results, and let readers make up their own minds about the implications. Bonferroni or other corrections are not generally applied to the significance levels of individual coefficients in multiple regression studies, or in the results of pilot or feasibility studies which are not designed to detect significant effects anyway.

In recent years the *p*-value appears to be losing ground to its close cousin, the confidence interval (CI). CIs are increasingly reported alongside *p*-values, and appear to many to be more informative and easier to understand. 95% CIs are the most commonly reported. These are usually (but not quite strictly correctly!) interpreted as the range of values within which we are 95% confident that a true population mean lies. CIs do not quantify the strength of evidence against the null hypothesis, as the *p*-value does, but instead give a measure of the precision of an estimate (for example, the difference between, or ratio of, the mean values in treatment groups).

There is an exact correspondence between CIs and the corresponding *p*-value: a 95% CI that excludes the key value 0 (for a difference between study groups) or 1 (for a ratio between study groups) corresponds to a *p*-value that is statistically significant at the 5% significance level (i.e. is less than 0.05). Conversely, a 95% CI that includes a key value corresponds to a *p*-value that is statistically non-significant at the 5% significance level (i.e. is 0.05 or greater). Atkinson et al[3] investigated the effect of various factors on risk of surgical site infection during spinal surgery, and in a typical presentation of tabulated results below, reported statistics from a model including both *p*-values and CIs. Note that the spinal levels factor, which is significant according to the *p*-value (0.019), is associated with a CI of 1.04 to 1.54, which excludes the key value (for a ratio) of 1; while the spinal region factor, which is non-significant according to the *p*-value (0.103), is associated with a CI of 0.71 to 44.3, which includes the key value. This table is also a good example of how an effect of relatively small magnitude (each additional spinal level is associated with a 26% increase in odds of infection) may be significant; whereas an effect of large magnitude (surgery performed in the thoracic, rather than non-thoracic region is associated with about a  fivefold increase in odds of infection) may be non-significant. While some professional bodies have gone as far as insisting that all *p*-values are replaced with CIs, my own opinion is that the two quantities are complementary, and both can add insight to study findings.



| **Factor/covariate** | **p value** | **Odds ratio** | **95% CI for odds ratio** |
|---|---|---|---|
| Number of spinal levels | 0.019 | 1.26 | (1.04, 1.54) |
| Primary spinal region–non-thoracic (reference) | | | |
| Thoracic | 0.103 | 5.59 | (0.71, 44.3) |

While *p*-values, used and interpreted correctly, can greatly aid the understanding of study findings, it is probably true to say that the mystique around them has led to over-estimation of their enlightening properties. They are often reported not merely without the corresponding CI, but also without the test statistic on which they are based; giving readers no clue about the practical or clinical importance of the treatment under investigation. You can provide me with virtually irrefutable evidence (*p*<0.001) that your expensive wound dressing is superior to my current product, but if that is evidence for only a 1% improvement in absorption properties, I may not be rushing to buy it!

1. Browning P, White RJ. Comparative evaluation of the functional properties of superabsorbent dressings and their effect on exudate management J Wound Care 2016; 25(8); 452-462

2. Babaei V, Afradi H, Gohardani H Z, Nasseri F, Azarafza M, Teimourian S.  Management of chronic diabetic foot ulcers using platelet-rich plasma J Wound Care 2016; 26(12); 784-787.

3. Atkinson R, Stephenson J, Jones A, Ousey K. An assessment of key risk factors for surgical site infection in patients undergoing surgery for spinal metastases J Wound Care 2016; 25(S9); S30-S34