assessing adequate doses to obtain benefits from capsaicin without producing collateral damage to dental pulps.

Conclusion

Inferior dental nerve infiltration with 1% capsaicin reduces SP expression in rat dental pulp tissue. This may provide a possible mechanism for controlling pulpal neurogenic inflammation to maintain pulp vitality when harmed by external irritants.

References

Awawdeh L, Lundy FT, Shaw C, Lamey PJ, Linden GJ, Kennedy JG (2002) Quantitative analysis of substance P, neurokinin A and calcitonin gene-related peptide in pulp tissue from painful and healthy human teeth. *International Endodontic Journal* **35**, 30–6.

Buck SH, Burks TF (1986) The neuropharmacology of capsaicin: review of some recent observations. *Pharmacological Reviews* **38**, 179–226.

Chancellor-Freeland C, Zhu GF, Kage R, Beller DI, Leeman SE, Black PH (1995) Substance P and stress-induced changes in macrophages. *Annals of the New York Academy of Sciences* **29**, 472–84.

Grutzner EH, Garry MG, Hargreaves KM (1992) Effect of injury on pulpal levels of immunoreactive substance P and immunoreactive calcitonin gene-related peptide. *Journal of Endodontics* **18**, 553–7.

Hargreaves KM, Swift JQ, Roszkowski MT, Bowles W, Garry MG, Jackson DL (1994) Pharmacology of peripheral neuropeptide and inflammatory mediator release. *Oral Surgery, Oral Medicine, and Oral Pathology* **78**, 503–10.

Hargreaves KM, Jackson DL, Bowles WR (2003) Adrenergic regulation of capsaicin-sensitive neurons in dental pulp. *Journal of Endodontics* **29**, 397–9.

Heyeraas KJ, Berggreen E (1999) Interstitial fluid pressure in normal and inflamed pulp. *Crit Rev Oral Biol Med* **10**, 328–36.

Hylden JL, Noguchi K, Ruda MA (1992) Neonatal capsaicin treatment attenuates spinal Fos activation and dynorphin gene expression following peripheral tissue inflammation and hyperalgesia. *Journal of Neuroscience* **12**, 1716–25.

Ichikawa H, Sugimoto T (2000) Vanilloid receptor 1-like receptor-immunoreactive primary sensory neurons in the rat trigeminal nervous system. *Neuroscience* **101**, 719–25.

Ikeda H, Tokita Y, Suda H (1997) Capsaicin-sensitive A delta fibers in cat tooth pulp. *Journal of Dental Research* **76**, 1341–9.

Jacobsen EB, Heyeraas KJ (1996) Effect of capsaicin treatment or inferior alveolar nerve resection on dentine formation and calcitonin gene-related peptide- and substance P-immunoreactive nerve fibers in rat molar pulp. *Archives of Oral Biology* **41**, 1121–31.

Kim S (1990) Neurovascular interactions in the dental pulp in health and inflammation. *Journal of Endodontics* **16**, 48–53.

Kyrkanides S, Tallents RH, Macher DJ, Olschowka JA, Stevens SY (2002) Temporomandibular joint nociception: effects of capsaicin on substance P-like immunoreactivity in the rabbit brain stem. *Journal of Orofacial Pain* **16**, 229–36.

Olgart LM, Kerezoudis NP (1994) Nerve–pulp interactions. *Arch Oral Biol* **39**(Suppl.), 47S–54S.

Payan DG (1989) Neuropeptides and inflammation: the role of substance P. *Annual Review of Medicine* **40**, 341–52.

Pernow B (1983) Substance P. *Pharmacological Reviews* **35**, 85–141.

Yoshimura M, Yonehara N, Ito T, Kawai Y, Tamura T (2000) Effects of topically applied capsaicin cream on neurogenic inflammation and thermal sensitivity in rats. *Japanese Journal of Pharmacology* **82**, 116–21.

COMMENTARY

The unit of analysis, and measurement of effect size

R. G. Newcombe

Statistical Advisor for the *International Endodontic Journal*

Correspondence: Robert G. Newcombe PhD CStat FFPH, Reader in Medical Statistics, Wales College of Medicine, Cardiff University, Health Park, Cardiff CF14 4XN, UK (Tel.: 0292074 2329; fax 029 2074 2898;; e-mail: newcombe@cardiff.ac.uk).

In refereeing the preceding article, two statistical issues arose, which are worth drawing to the attention of journal readers. The first relates to a flaw in the study design, which in my experience may be common in many areas of dental and medical research. The second relates to the appropriate quantification of effect sizes. The simple structure of the results presented and the extreme differences shown render this article an ideal vehicle to make these points.

With regard to the study design, I commented to this effect:

'There is one unfortunate aspect of the study design. The unit of data should be the animal, not the tooth. Two animals were allocated to each regime, so really the sample size per group is not 12 but 2, and from this viewpoint, the differences would no longer be regarded as statistically significant. A better design (if feasible) would be to take, say, six animals, identify six lower molars from each, and allocate two to each regime, preferably balancing this allocation with regard to tooth position. This would normally lead to a two-way analysis by animal and regime.

It would be quite justifiable to reject this article on account of this design flaw. Nevertheless, in view of the overwhelming differences between regimes shown, and dual to this, the very limited degree of variation within each column that could have arisen due to inter-animal variation, I consider the observed findings are more likely to be real than an artefact due to this design flaw'.

For an illustration of the fallacy of using the site, not the individual as the unit of data see Newcombe & Duff (1987).

With regard to characterizing differences between groups, the correct strategy is used viz. a one-way ANOVA approach first, comparing the three groups on an equal footing in order to minimize the impact of multiple testing, then having found highly significant differences, pairwise comparisons are performed. Non-parametric methods are chosen, in this instance primarily because the gross difference in spread between the groups invalidates the parametric ANOVA. But this is a clear case in which, having answered the first question 'Do the groups differ?' in the affirmative, the obvious next question is, 'Can we say by how much?'. The *P*-value produced by a significance test does not convey this information. Sometimes, when researchers quote a Mann–Whitney test, they also give an estimated median difference as an expression of the effect size; a confidence interval for this is available in some widely used statistical software such as Minitab. Nevertheless, these results have some unusual characteristics. The estimated median difference is not identical to the differences between the group medians. On small samples it is not possible to construct an interval with exactly 95% nominal confidence level – in this instance they are 95.4% intervals – and sometimes (although not in this case) the resulting intervals can be disconcertingly skew. Moreover, the reader may not be able to relate directly to what is conveyed by the statement that the median difference between groups 2 and 3 is 123.94 (95.4% confidence interval 104.92–142.00) – the units are valid ones but are unlikely to be sufficiently familiar to most readers to enable them to appraise how large a difference should be regarded as important. (Of course, in the light of the previous remarks, any confidence interval calculations are only valid if we are content to accept that all the values in a column can be treated as independent.)

An alternative approach which gets around these difficulties of interpretation is to use a *relative*, not an *absolute* measure of effect size. A suitable measure is obtained by dividing the Mann–Whitney *U*-statistic by the product of the sizes of the two samples, giving an index which we can call U/mn. This measure takes the value 1 (or 0) when there is no overlap between two samples, and 0.5 when they are identical. In this instance, all three groups have 12 observations each. All values in group 2 exceed all values in group 1, so here U takes its maximum possible value of $12 \times 12 = 144$, and $U/mn = 144/144 = 1.0$. Similarly, comparing groups 2 and 3 gives $U/mn = 1.0$. When we compare groups 1 and 3, we find there is a single overlap, viz. $0.109 > 0.102$, but apart from this, all values in group 1 are higher than all values in group 3. So here, $U = 144-1 = 143$ and $U/mn = 143/144 = 0.993$.

All these values taken at face value indicate a very high degree of separation between groups. But they are based on very small sample sizes. Confidence intervals here are very informative, showing how one could occasionally end up with such a high value of U/mn by chance when in the underlying populations, the degree of separation is substantially lower than this. Confidence intervals for U/mn are now available using a method that has been validated as satisfactory in such extreme cases as well as when the degree of separation is more moderate (Newcombe 2004a,b). A highly user-friendly Excel spreadsheet GENERALISEDMW1.xls which performs the calculations is available at the author's website: http://www.uwcm.ac.uk/study/medicine/epidemiology_statistics/research/statistics/newcombe.htm.

The spreadsheet requires the user to supply the sizes of the two groups to be compared and the value of U, which would normally be obtained using standard statistical software. For the contrast between groups 1 and 3, the

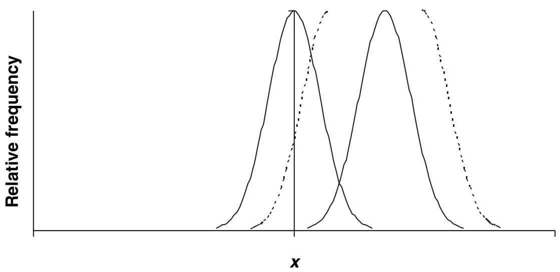

Figure 1 Illustration of degree of separation between two frequency distributions corresponding to U/mn = 0.993. Broken curves show confidence region.

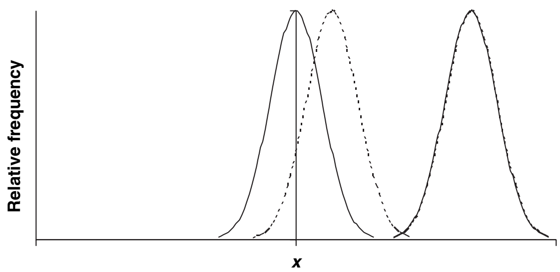

Figure 2 Illustration of degree of separation between two frequency distributions corresponding to U/mn = 1.0. Broken curve shows lower confidence bound.

spreadsheet calculates and displays $U/mn = 0.993$ with 95% confidence interval 0.819–0.9997. These indicate that at a conservative estimate, the degree of separation may be considerably lower than 1, in fact around 0.82, nevertheless this also represents considerable separation between two populations.

The spreadsheet also displays a graph (Fig. 1) which helps visualize the degree of separation. Two solid Gaussian distribution curves are displayed, in this instance with peaks separated by 3.5 standard deviations, which corresponds to $U/mn = 0.993$. The curve for which a central vertical line is displayed represents the reference distribution. The second solid curve represents the degree of separation from this. In this instance, with $U/mn > 0.5$, it is to the right of the reference curve. Two broken curves are also displayed, which demarcate a confidence region around the second solid curve. The left-hand, ascending broken curve is half of a Gaussian curve corresponding to a separation of $U/mn = 0.819$ from the reference curve. The right-hand, descending broken curve is half of a Gaussian curve corresponding to a separation of $U/mn = 0.9997$ from the reference curve. Generally, only half curves are plotted, to avoid producing a cluttered diagram.

For the contrast between group 2 and either group 1 or group 3, $U/mn = 1.0$, with 95% confidence interval 0.832–1.0, once again indicating that even at a conservative estimate, the data indicates a substantial degree of separation between groups. This extreme case is displayed (Fig. 2) as two Gaussian distributions separated by 6.7 standard deviations, which corresponds to $U/mn = 0.999999$; the two curves appear completely separate. The lower limit for U/mn, 0.832, is represented by a broken Gaussian curve corresponding to this degree of separation from the reference distribution; in this case, both ascending and descending parts of the broken curve are displayed. The upper limit is the same as the point estimate, $U/mn = 1.0$, and the curve representing this coincides with the second solid curve.

In both instances, the inference is that on the data actually collected, the degree of separation of the two distributions is either perfect or nearly so. However, at a conservative estimate, the value of U/mn in the wider population of specimens such as these could be as low as 0.82–0.83, which corresponds to a substantial shift, nevertheless with considerable overlap between the two distributions.

References

Newcombe RG (2004a) Interval estimation for U/mn, a generalisation of the Mann–Whitney statistic. *Presented at the International Biometric Society Conference*, Cairns, July 2004.

Newcombe RG (2004b) Confidence intervals for an effect size measure based on the Mann–Whitney statistic 2. Asymptotic methods and evaluation. *Statistics in Medicine*, in submission.

Newcombe RG, Duff GR (1987) Eyes or patients? Traps for the unwary in the statistical analysis of ophthalmological studies. *British Journal of Ophthalmology* **71**, 645–646.