Editorial

Penehcyclidine and awareness during anaesthesia: caution with zero numerators

In this issue of the journal, Wang et al. [1] make the extraordinary claim that a novel anticholinergic drug, penehcyclidine hydrochloride (PHC), can reduce, if not abolish, the occurrence of accidental intra-operative awareness with explicit recall (subsequently referred to as 'recall') [2]. In a placebo group of 452 patients, five experienced recall compared with 0/456 in the PHC group.

A note on the scientific motivation for this study is warranted. Acetylcholine is produced by neuronal subpopulations in the basal forebrain and the laterodorsal/pedunculopontine tegmentum in the pons [3]. It is well known that states of cortical activation (e.g. during waking or rapid eye movement sleep) are associated with high cholinergic tone; acetylcholine also plays a role in memory. As such, the effects of an anticholinergic drug such as PHC could conceivably suppress consciousness and memory formation.

Penehcyclidine hydrochloride is not available in the UK or the USA, but before readers rush to the internet to order supplies, it is worth considering some of the limitations of Wang et al.'s paper. We served as reviewers of the original submission and although we raised some of the points below in criticism, other reviewers favoured publication of this research. Disagreements of this sort are a healthy aspect of peer review and we were invited to offer readers an alternative viewpoint. Our main concerns revolve around the methodology and the statistics – the readers are left to judge the impact that Wang et al.'s paper should have on clinical practice.

Methodological caution

One salient methodological concern in the work of Wang et al. relates to the rates of recall used as the basis for the power calculation. Large prospective studies in the US [4, 5] and Europe [6] have consistently established a rate of recall between 1 and 2 per 1000. However, as Wang et al. note, in China the rate of recall may be considerably higher, ranging from 4 to 10 per 1000 in patients at all risk levels and almost 50 per 1000 after cardiac surgery [7–9]. Wang et al. report that their power analysis was based on two of these studies, which assessed two distinct risk populations with rates at the extremes of this range, separated by one order of magnitude (0.41% and 4.7%). For their power calculation they additionally cite the work of Groesdonk et al. [10], which found a zero rate of true recall in 534 fast-track cardiac surgery patients, and Elhakim et al. [11], which found a rate of 2/25 (8%) after thoracic surgery. It is unclear how Wang et al. determined a single rate of recall from these diverse studies, some of which were obviously not intended to result in epidemiologic data. Furthermore, three of the four studies they cite focused on cardiothoracic patients and are unlikely to be applicable to women undergoing breast surgery with total intravenous anaesthesia (TIVA), which was the population chosen by Wang et al.

Both groups in the cohort studied by Wang et al. had anaesthetic dose targeted to a Bispectral Index (BIS) value between 40 and 60, with patients outside of this range (for two consecutive values) excluded. Notably, two patients in the PHC group were excluded for this reason (one with a BIS < 40 and one with a BIS > 60). Had either of these patients experienced recall, the conclusions of the study would have been radically altered (see below). For the remainder, 5/452 patients (1.1%) in the control group reported recall despite their BIS values remaining within the acceptable range as defined in this study. Although patients receiving TIVA may be at higher risk for recall [12], these findings are not consistent with

a large, multicentre, randomised controlled trial conducted on TIVA and BIS in China. Zhang et al. reported that 0.65% of patients receiving TIVA without BIS monitoring had recall, while those with BIS monitoring had a rate of 0.14% [13]. It is extremely surprising (and difficult to explain) that a comparable population (Chinese surgical patients with BIS monitoring), could exhibit a rate of recall an order of magnitude lower than that reported by Wang et al. Although the authors used a committee of three experts to assess if any recall represented 'definite', 'possible' or 'no' awareness, they do not clearly present how many patients fell into which category (though we must assume that all of the five patients reported were classed as 'definite').

Finally, Wang et al.'s methodology was confined to two interviews, one postoperatively and one at up to 48 h after surgery. Most studies employ a third interview from one week to one month after surgery, that might detect recall more often [4, 6]. This is especially important to consider since, in one group, a zero numerator was obtained.

Statistical caution

The main concerns here are encapsulated in the penultimate paragraph of Wang et al.'s paper; namely, that the authors found a 'zero numerator' and also applied a 'one-tailed test'.

On the problem of zero numerators

Zero numerators are much beloved of some aspirant authors and the problems they cause to data interpretation have been discussed extensively in the pages of this journal and elsewhere [14–17]. A researcher may conduct an observational study or a comparative trial of an intervention and find that, in the test group of interest, there are no complications or failures at all (i.e. the zero numerator). The researcher will therefore conclude that the intervention was extremely favourable. Wang et al. did this comparing two ratios: 0/456 and 5/452 and found, unsurprisingly, that they are different ($p = 0.03$; one-tailed Fisher's exact test).

For reasons previously explained in some detail [17], this conclusion is rarely, if ever, correct. Whenever there is a zero numerator, what really matters is the upper limit of the 95% confidence interval (in this case, using binomial confidence intervals). For the data of Wang et al., this upper limit of 0/456 is in fact 3.7/456 (rounded up to 4/456), so it is this incidence that should form the basis of any comparison (which would not be statistically significant: $p = 0.495$, one-tailed Fisher's exact test).

This approach is emphasised in the so-called 'rule of three', which states that whenever there is a zero numerator, or whenever no complications or failures are observed, out of n observations, the upper 95% CI limit for the 'true' rate is approximately $3/n$; a simple approach that has its limitations [18], but is often surprisingly accurate [14, 15] (for Wang et al.'s dataset, the 95% CI is closer to 4, not 3).

In fact, Wang's data are exquisitely sensitive to the rate of recall in the test group. The 'rule of three' is overkill for their data; even if just one patient in the PHC group experienced recall, the comparison jumps from a 'significant' $p = 0.03$ to a 'non-significant' $p = 0.107$. Given our methodological concerns outlined above, it is not difficult to imagine either that the rate of recall in the PHC group was underestimated, or that the rate in the control group was overestimated.

On the dilemma of tails

Wang et al. performed a one-tailed statistical comparison on their data. A 'tail' can be understood by imagining a graph of the data distribution (i.e. a histogram of the frequency of an observation's occurrence against the values; Fig. 1).

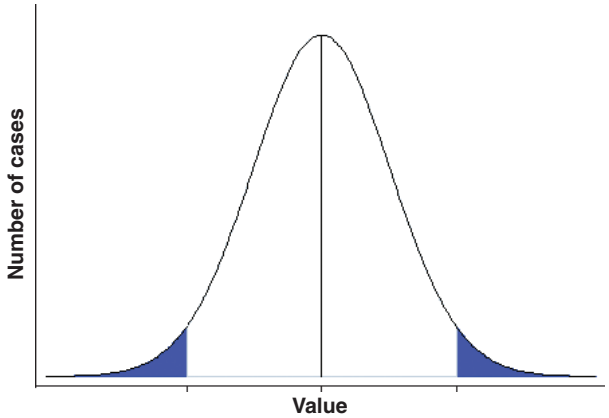

Figure 1 A normal distribution curve with the two tails (each representing ~2.5% of the data points) shown in blue.

The normal distribution is well known, and the data points at the very limits of the curve at either end form the two 'tails'. A two-tailed comparison tests for the possibility of a relationship in both directions. For example, we may wish to compare the mean of a sample against a given value x using a t-test. Our null hypothesis is that the mean = x. A two-tailed comparison will test whether the mean is both significantly > x and significantly < x. In the calculation, values from both 'tails' are used.

On the other hand, a one-tailed comparison tests for the possibility of a relationship only in one direction and completely disregards the possibility of a relationship in the other direction. In the example above, comparing the mean of a sample against x, a one-tailed comparison tests whether the mean is significantly > x or significantly < x, but not both possibilities at once. A one-tailed comparison thus provides more power to detect an effect in one direction by not testing the effect in the other direction.

It is possible to justify the use of one-tailed comparisons in some circumstances, especially where the overall direction of intervention is itself not in doubt, but the magnitude of effect may be of more interest. The choice of test should be based on the hypothesis so that for novel, unusual therapies about which very little is known (such as PHC), where the intervention could result in a change in either direction, a two-tailed comparison would usually be regarded as essential [19–22].

The use of two-tailed comparisons in Fisher's exact test requires a little explanation. The contribution of Fisher to statistics has been discussed before in the pages of this journal [23]. His 'exact test' was originally designed to assess if a colleague could tell whether milk or tea had been poured first into her cup, and it follows a hypergeometric distribution (Fig. 2). This is the distribution obtained when performing binomial sampling without replacement; for example, the probabilities of obtaining 0, 1, 2, 3, 4 and 5 white balls from an urn of 15 red and 5 white balls. Fig. 2 shows that this distribution is, unlike the normal distribution, unequal in its tails.

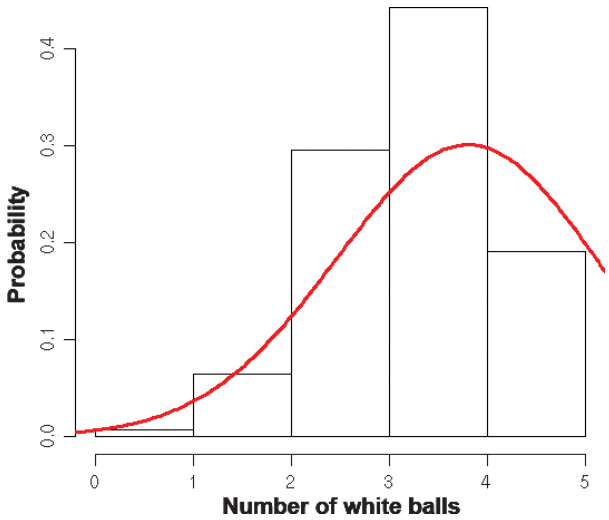

Figure 2 Hypergeometric distribution. In an urn of 15 red and 5 white balls, a sample of 5 balls is taken without replacement; the graph shows the probability of obtaining 0, 1, 2, 3, 4 and 5 white balls in the sample. Note the asymmetric distribution.

Table 1 The data of Wang et al. laid out in a 2 × 2 table. Numbers in black are Wang et al.'s reported data; numbers in red offer an example of the single (more extreme) tail that they used in their one-tailed test. Numbers in blue represent a less extreme tail, the probabilities for which would be included in a two-tailed test.

	PHC	Control
Aware	0,1,2,3,4 → 5 → 6,7,8….452	0
Not aware	452,451,…→ 447 → 446,445,….0	456

With a zero numerator, the data of Wang et al. were at one extreme of this distribution. Table 1 shows the data of Wang et al. displayed in a 2 × 2 format. The values in the top row are 0 vs 5. Theoretically more extreme ratios would be 0 vs 6, 0 vs 7, 0 vs 8,…0 vs 452 (Table 1; red text). The one-sided comparison that Wang et al. made tested whether 0/456 is less than 5/452, 6/452, 7/452,…452/452, which of course it is. Had they conducted a two-tailed test, they would have additionally examined the comparison of less extreme ratios; namely 0/456 vs 4/452, 3/452,

2/452, 1/452 and 0/452 (Table 1; blue text). For Fisher's exact test, a two-tailed comparison may not always yield a p value exactly double that of a one-sided comparison, but in this case, a proper, two-tailed comparison of the data in Table 1 yields a p value of 0.062, which is non-significant.

Summary

"*Le poids de la preuve pour une affirmation extraordinaire doit Átre proportionnel à son degré d'étrangeté*"

Marquis de Laplace [24]

Extraordinary claims need extraordinary evidence. Statistical significance (the p value) is in part an index of how confident we should be about the 'reality' of suggested differences. For significance obtained using a one-sided test that remains even when using a two-sided test, we can be fairly confident that any differences reported are likely and consistently 'real'. Significance obtained using a one-sided test that disappears when using a two-sided test should raise more than a kernel of doubt.

The methodological shortcomings may have missed some cases of recall in Wang et al.'s test group, and so led to the problem of a zero numerator in the PHC group. This was coupled with an unusually high rate of recall in the control group. Then, the only way that Wang et al. could obtain a significant result for PHC's effect was by performing a one-tailed test. If they had used any numerator other than 0 (e.g. applied the rule of three), or applied binomial confidence intervals, or correctly used a two-tailed comparison, the significant effect of PHC simply disappears. It is indeed an extraordinary claim to suggest that a single drug, used empirically, can eliminate recall after general anaesthesia. The evidence in support of that suggestion needs to be far more extraordinary, in our opinion, than has been presented by Wang et al. Nonetheless, it is an interesting idea and we look forward to further work in this compelling area of research.

Competing interests

JJP is Clinical Lead for the Fifth National Audit Project of the Royal College of Anaesthetists and the Association of Anaesthetists of Great Britain and Ireland (NAP5; Accidental Awareness during General Anaesthesia in the United Kingdom). There is no direct overlap of material presented in this editorial or the paper of Wang et al. with the work of NAP5. The views expressed are those of JJP as an individual and reviewer, and not those of NAP5 or its sponsors. No external funding or other competing interests declared.

J. J. Pandit
Professor & Consultant Anaesthetist
Nuffield Department of Anaesthetics
Oxford University Hospitals
Oxford, UK
Email: jaideep.pandit@dpag.ox.ac.uk
P. Picton
G. A. Mashour
Assistant Professors
Department of Anesthesiology
University of Michigan Medical School Ann Arbor
Michigan, USA

References

1. Wang J, Ren Y, Zhu Y, et al. Effect of penehyclidine hydrochloride on the incidence of intra-operative awareness in Chinese patients undergoing breast cancer surgery during general anaesthesia. *Anaesthesia* doi:10.1111/anae.12092
2. Wang M, Messina AG, Russell IF. The topography of awareness: a classification of intra-operative cognitive states. *Anaesthesia* 2012; **67**: 1197–201.
3. Woolf NJ, Butcher LL. Cholinergic systems mediate action from movement to higher consciousness. *Behavioural Brain Research* 2011; **221**: 488–98.
4. Sebel PS, Bowdle TA, Ghoneim MM, et al. The incidence of awareness during anesthesia: a multicenter United States study. *Anesthesia and Analgesia* 2004; **99**: 833–9.
5. Mashour GA, Shanks A, Tremper KK, et al. Prevention of intraoperative awareness with explicit recall in an unselected surgical population: a randomized comparative effectiveness trial. *Anesthesiology* 2012; **117**: 717–25.
6. Sandin RH, Enlund G, Samuelsson P, Lennmarken C. Awareness during anaesthesia: a prospective case study. *Lancet* 2000; **355**: 707–11.
7. Xu L, Wu AS, Yue Y. The incidence of intra-operative awareness during general anesthesia in China: a multi-center observational study. *Acta Anaesthesiologica Scandinavica* 2009; **53**: 873–82.
8. Wang E, Ye Z, Pan Y, et al. Incidence and risk factors of intraoperative awareness during general anesthesia. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* 2011; **36**: 671–5.
9. Wang Y, Yue Y, Sun YH, et al. Investigation and analysis of incidence of awareness in patients undergoing cardiac surgery in Beijing, China. *Chinese Medical Journal* 2005; **118**: 1190–4.
10. Groesdonk HV, Pietzner J, Borger MA, et al. The incidence of intraoperative awareness in cardiac surgery fast-track treatment. *Journal of Cardiothoracic and Vascular Anesthesia* 2010; **24**: 785–9.
11. Elhakim M, Abdelhamid D, Abdelfattach H, Magdy H, Elsayed A, Elshafei M. Effect of epidural dexmedetomidine on intraoperative awareness and postoperative pain after one-lung ventilation. *Acta Anaesthesiologica Scandinavica* 2010; **54**: 703–9.
12. Nordstr—m O, Engstr—m AM, Persson S, Sandin R. Incidence of awareness in total i.v. anaesthesia based on propofol,

alfentanil and neuromuscular blockade. *Acta Anaesthesiologica Scandinavica* 1997; **41**: 978–84.
13. Zhang C, Xu L, Ma YQ, et al. Bispectral index monitoring prevent awareness during total intravenous anesthesia: a prospective, randomized, double-blinded, multi-center controlled trial. *Chinese Medical Journal* 2011; **124**: 3664–9.
14. Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not yet occurred: a statistical reminder. *British Medical Journal* 1995; **311**: 619–20.
15. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything alright? *Journal of the American Medical Association* 1983; **249**: 1743–5.
16. Pandit JJ. Gambling with ethics? A statistical note on the Poisson (binomial) distribution. *Anaesthesia* 2008; **63**: 1163–6.
17. Pandit JJ. If it hasn't failed, does it work? On the worst we can expect' from observational trial results, with reference to airway management devices. *Anaesthesia* 2012; **67**: 578–83.
18. Winkler RL, Smith JE, Fryback DG. The role of informative priors in zero-numerator problems: being conservative versus being candid. *American Statistician* 2002; **56**: 1–4.
19. Bland JM, Altman DG. One and two sided tests of significance. *British Medical Journal* 1994; **309**: 248.
20. Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology* 2001; **54**: 109–10.
21. Koch GG. One-sided and two-sided tests and p values. *Journal of Biopharmaceutical Statistics* 1991; **1**: 161–70.
22. Peace KE. The alternative hypothesis: one-sided or two-sided? *Journal of Clinical Epidemiology* 1989; **42**: 473–6.
23. Pandit JJ. The analysis of variance in anaesthetic research: statistics, biography and history. *Anaesthesia* 2010; **65**: 1212–20.
24. Hahn R. *Pierre Simon Laplace, 1749-1827: a Determined Scientist*. Cambridge, Mass, USA: Harvard University Press, 2005.

doi:10.1111/anae.12114