Pain 139 (2008) 481–482

PAIN

www.elsevier.com/locate/pain

Editorial

So, why *do* I have to correct for multiple comparisons? Concepts and commentary on Turk et al.

The paper by Turk et al. [3] covers an important topic that has generated a lot of discussion among statisticians and researchers over the past 40 years. To provide a framework, suppose that a researcher compares a treatment to a control on mean pain, in 20 different subgroups. Each comparison has an adequate sample size. Nevertheless, the difference between treatment and control is significant only in the "Philadelphia" subgroup, at $p = 0.03$.

The problem may appear obvious. "Of course", if you test in 20 subgroups you expect one to be significant by chance, even if there are no true differences. Therefore, one weakly significant comparison out of 20 could easily be a type I error (concluding an effect when there is none/false significance) and cannot be interpreted as a real effect. This standard reasoning is the basis for the Turk et al. paper in this issue.

But how exactly does the presence of 19 non-significant results alter the interpretation of the one that was significant? Why, for example, would the other results be relevant to a clinician who was only interested in Philadelphians? (See Walker [4] for a similar reasoning). Indeed, some commentators have argued against any correction for multiple comparisons. They suggest that we interpret each result as its own analysis, regardless of the number of others tested [2].

The answer lies in recognizing that the results of the other comparisons typically *do* have implications for the significant one, or the one of interest. In the example, unless there was good reason to expect this therapy to work better in Philadelphians than elsewhere, the fact that 19 other comparisons were non-significant is a strong argument that this treatment does not work. Anywhere! Thus, a smaller *p*-value is needed to provide adequate evidence for a difference in Philadelphians.

A similar argument applies to other multiple comparisons settings. Say, 20 variables compared between two groups on an exploratory basis. Or 20 theory-driven comparisons. In either case 19 are non-significant. The 20th comparison needs to work harder to convince a skeptic (by having a much smaller *p*-value!).

On the other hand, if it was truly likely that the treatment would work in Philadelphians much more than elsewhere, then that comparison would become a "primary hypothesis" and no correction for the number of others would be necessary. The fact that 19 truly irrelevant comparisons were non-significant would not change the prior probability that the one relevant hypothesis was true. Turk et al. recognize this implicitly in their discussion of primary endpoints. They correctly emphasize that the primary hypothesis must be specified and justified in advance. It cannot be decided or changed after the data are obtained, because doing so vitiates the logic that allows it to be tested without consideration of the results of other analyses.

Other than for well-defined primary hypotheses, the importance of being aware of type I error and taking steps to avoid it remains. Consider the following "real life" scenario: I found $p < 0.03$ for my comparison, only in the Philadelphia subgroup. That's bad enough. I tried breaking down the comparisons by age and ethnicity (not significant so I didn't "mention" these analyses). That's *worse*. And, my study might have been filed away, never published, had I not "found" anything to report. That's "file drawer bias". These are all pressures making "significant" *p*-values more likely to be false positives. We need to control type I error, or real effects will get lost in a mass of random garbage.

While there is no cure for the need to avoid falsely significant results (type I errors), recent research is increasing the number of options that are available. Geneticists may have literally thousands of comparisons in a single study. The so-called "empirical Bayesian" approaches have been developed to analyze such data without severely restrictive type I error controls. Effron et al. [1] have a much-cited paper on this topic. "False Discovery rate" methods, which Turk et al. mention briefly, provide yet another approach.

doi:10.1016/j.pain.2008.09.011

Turk et al. do a nice overview of commonly employed corrections for multiple comparisons, such as the Bonferroni correction and its recent modifications, as well as global procedures such as Hotelling's T^2.

Most of us learned from texts in which the Bonferroni correction had no direct competitors outside of the context of ANOVA. Today, there are several variants, some of which are more powerful (but may also sacrifice some control of type I error). Since a full Bonferroni is conservative if the hypothesis tests are correlated, I think these modern versions are fairly safe to use.

Most of us have also used the global procedures, like MANOVAs (with Hotelling's T^2 or equivalent statistics). These procedures avoid the multiplicity issue by considering all the hypotheses as a single set. When a global summary of effects is interpretable, these may be useful – the authors discuss their limitations.

Less familiar to most readers will be the "gatekeeping" approaches Turk et al. describe. These techniques are potentially helpful, but deal only with the situation in which hypotheses can be strictly and logically ordered. Otherwise, one risks, for example, seeing numerous interesting differences declared non-significant because one "primary" hypothesis exhibited $p = 0.07$, thereby stopping the analysis. Some more flexible approaches do exist, as Turk et al. mention, but these raise concerns as well. I would prefer an approach that does not depend so heavily on the honesty, logic, and attention to advance planning by the researchers.

The bottom line remains that good research deals effectively with chance. There are many ways to do that, but if we fail to use them, we undermine our credibility.

References

[1] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001;96:1151–60.

[2] Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6.

[3] Turk D, Dworkin RD, McDermott MP, Bellamy N, Burke LB, Chandler JM, et al. Analyzing multiple endpoints in clinical trials of pain treatments: IMMPACT recommendations. Pain 2008;139:485–93.

[4] Walker AM. Reporting the results of epidemiologic studies. Am J Public Health 1986;76:556–8.

Edward Gracely
Drexel University, College of Medicine, Family, Community, and Preventive Medicine, 2900 Queen Lane, Philadelphia, PA 19129, USA
Tel.: +1 215 991 8466; fax: +1 215 843 6028.
E-mail address: egracely@drexelmed.edu