Ophthalmic statistics note 7: multiple hypothesis testing—to adjust or not to adjust

Valentina Cipriani,[1,2] Ana Quartilho,[1] Catey Bunce,[1,3] Nick Freemantle,[4] Caroline J Doré,[5] on behalf of the Ophthalmic Statistics Group

DEFINING THE PROBLEM

Investigating multiple research questions, or hypotheses, within one study is a common scenario in biomedical research with many examples in ophthalmology. As the number of statistical tests increases, the overall chance that we draw an erroneous conclusion in our study gets higher in a predictable manner. Each statistical test conducted at the conventional 5% significance level (α) has a one in 20 chance (or 0.05 probability) of appearing significant simply due to chance (a type I error) and a $1-0.05=0.95$ probability of being non-significant. If we test two independent true null hypotheses, the probability that neither test will be significant is $0.95\times0.95=0.90$. Likewise, if we test 14 independent hypotheses, the probability that none will be significant is $0.95^{14}=0.49$, and the probability that at least one will be significant is $1-0.49=0.51$, that is, we are more likely than not to find at least one test significant. In other words, if we go on carrying out tests of significance we are very likely to find a spurious significant result. In the field of statistics, this phenomenon is known as the problem of *multiple testing* or the *multiplicity problem*.[1]

Consider the ABC study which compared bevacizumab for neovascular age-related macular degeneration (nAMD) with standard National Health Service (NHS) care.[2] This study was conducted on 131 patients and found that 21 (32%) of patients treated with bevacizumab gained ≥15 letters compared with two (3%) of those in the standard care group with an OR of 18.1 (95% CI 3.6 to 91.2; $p<0.001$). The primary objective of this study was to determine whether bevacizumab was superior to standard NHS care and this single test of significance provided strong evidence. Closer inspection of the study reveals however that a variety of different treatments were used within the NHS standard care arm (sham injections, photodynamic treatment with verteporfin, intravitreal pegaptanib) and it was natural that investigators would wish to establish evidence of efficacy between bevacizumab and each of these alternative treatment modalities. Similarly, while the study had revealed evidence of a treatment effect on visual acuity, investigators were interested also to examine efficacy on other measures of visual function such as contrast sensitivity.[3] Clinical trials can be expensive and it would seem very wasteful and indeed perhaps unethical not to explore the data further. However, a single question at the outset has led to many questions of interest and many tests of significance being proposed.

Multiplicity may arise due to several different issues, including:

1. multiple outcomes (visual acuity, contrast sensitivity, quality of life)
2. subgroups (was the nAMD classic or occult)
3. multiple time points (the data were assessed at 1 year, was there evidence of an effect at 6 months?)
4. multiple questions (initially our scenario compared bevacizumab with standard care, but standard care could be sham injections, photodynamic treatment with verteporfin, intravitreal pegaptanib).

While clinical trialists may contend with multiplicity in the order of tens or perhaps 100s, genetic statisticians are dealing with multiplicity in the order of thousands or indeed millions. A special case of issue 4 (above), for example, might be large-scale genetic studies, including genome-wide association studies (GWASs) where thousands if not millions of single-nucleotide polymorphisms (SNPs) across the genome are genotyped simultaneously in a large set of cases and controls. A genetic association test that looks for a different allele frequency between cases and controls is then performed on each SNP and a corresponding p value calculated. For example, in a GWAS of age-related macular degeneration conducted in the UK on 743 advanced cases and 1598 controls, a genetic association test was performed on each of the 488 867 SNPs that passed quality control and a total of 26 116 tests with a p value <0.05 was observed.[4–6] This number is close to what would be expected (ie, $488\,867\times0.05=24\,443$) to show a significant result by chance alone when there is in fact no genetic association. Multiplicity issues can arise in all areas of medical research.

HOW TO ADJUST

When adjusting for multiplicity, a more stringent significance threshold than the usual $\alpha=0.05$ is used, so that rejecting the null hypothesis becomes more difficult and some protection against false-positive findings is gained. A plethora of statistical procedures has been developed to calculate an adjusted significance threshold and guidance on the most appropriate method (s) should be sought from a statistician at the study design stage and incorporated in the protocol.[7] One of the simplest approaches is the Bonferroni method that consists in setting the adjusted significance threshold for each test to 0.05/M, where M is the total number of independent tests to be performed. This way the probability of having at least one false-positive result in the study ('study-wise' α) when the null hypothesis is true for all M tests is no more than 0.05. Its simplicity has attracted criticism and more complex procedures may be more appropriate.[8] In the special case of GWASs where a very large number of genetic markers are tested, strict adjustment has become standard practice and a *genome-wide significance level* threshold of 5×10^{-8} obtained from simulation studies that emulated an infinitely dense SNP map has been widely adopted regardless of the actual SNP density of the study.[9 10]

It is important to remember that an inevitable consequence of reducing the risk of false-positive findings is that the risk of missing true-positive findings will increase. This is one of the reasons that the Bonferroni correction has attracted criticism. It is often termed overly conservative.

SHOULD WE ADJUST?

In clinical trials the question of multiplicity is usually addressed through the

[1]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK; [2]UCL Genetics Institute, London, UK; [3]London School of Hygiene & Tropical Medicine, London, UK; [4]Department of Primary Care and Population Health, University College London, London, UK; [5]UCL Comprehensive Clinical Trials Unit, University College London, London, UK

Correspondence to Dr Valentina Cipriani, NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, EC1V 9EL, UK; v.cipriani@ucl.ac.uk

identification a priori of a primary outcome on which the overall conclusion of the trial is judged. This simple approach preserves the study-wise α, relegating secondary outcomes to descriptors of how an intervention works or exploratory findings. Thus, in a trial where the primary outcome is statistically significant, we conclude that a difference between the experimental conditions has been established, and the results of secondary outcomes can be considered 'nominally' (ie, without regard to multiplicity) as indications of how the treatment appears to be working. If the primary outcome is not significant the results of the trial may be inconclusive, depending on the size of the CI around the estimated treatment effect,[11] and any apparently significant secondary outcomes must be viewed only as exploratory findings requiring confirmation in future trials. For randomised trials, the ICH E9 provides useful guidance on preserving study-wise α and the criteria we might use to decide upon the primary outcome.[12]

There are cases, however, where multiplicity should not be seen as a problem and where no adjustment is necessary.[7]

When designing a study it is essential to have a clear idea of the type of evidence that we are seeking and where the question lies in the research continuum. Are we seeking confirmation of a well described and understood effect, or are we exploring questions that would lead to further research? Studies wishing to provide a definitive answer are *confirmatory*, whereas studies which generate new hypotheses are *exploratory*. Since evidence from confirmatory studies can impact standard care, it is essential to control for the increased error arising from multiple statistical tests and adjustment for multiple testing is mandatory. Adjustment may not, however, be necessary for exploratory studies (figure 1).[7] Typically these studies are the first of their kind and because of this are small in size. Adjusting p values in this context might stop development of a promising treatment that could have a positive finding in a future definitive study. We would, however, urge those conducting such studies to follow good practice recommendations for pilot studies and focus more on CIs and estimation than on tests of significance.[13] Clearly, many studies involve combinations of confirmatory and exploratory objectives, as illustrated in our flow chart (figure 1). It is good practice to refrain from conducting an unnecessarily large number of hypothesis tests and instead try and reduce their number by defining the goals of the study. If we think about the study in advance we can give some questions higher priority. For example, if we have multiple time points in a study we can state in the study protocol that the outcome at 12 months is most clinically relevant and is the primary outcome while the outcomes at 24 and 36 months are secondary endpoints and interpreted as exploratory. This way we have a single primary outcome and avoid multiplicity issues.[7] Similarly, in the ABC study, the primary objective was to compare bevacizumab with standard NHS care: the comparisons with each individual treatment were viewed as secondary objectives.

Questions which arise after the primary analysis of the trial has been conducted are termed 'post hoc'. Clearly these cannot be accounted for in advance within the protocol. Multiple adjustment for these may not be necessary provided that such findings are clearly reported as post hoc, indicating that caution is required when interpreting them.

Finally, while the discussion of multiplicity naturally focuses on α levels and thus p values, these have somewhat surprising characteristics which challenge our

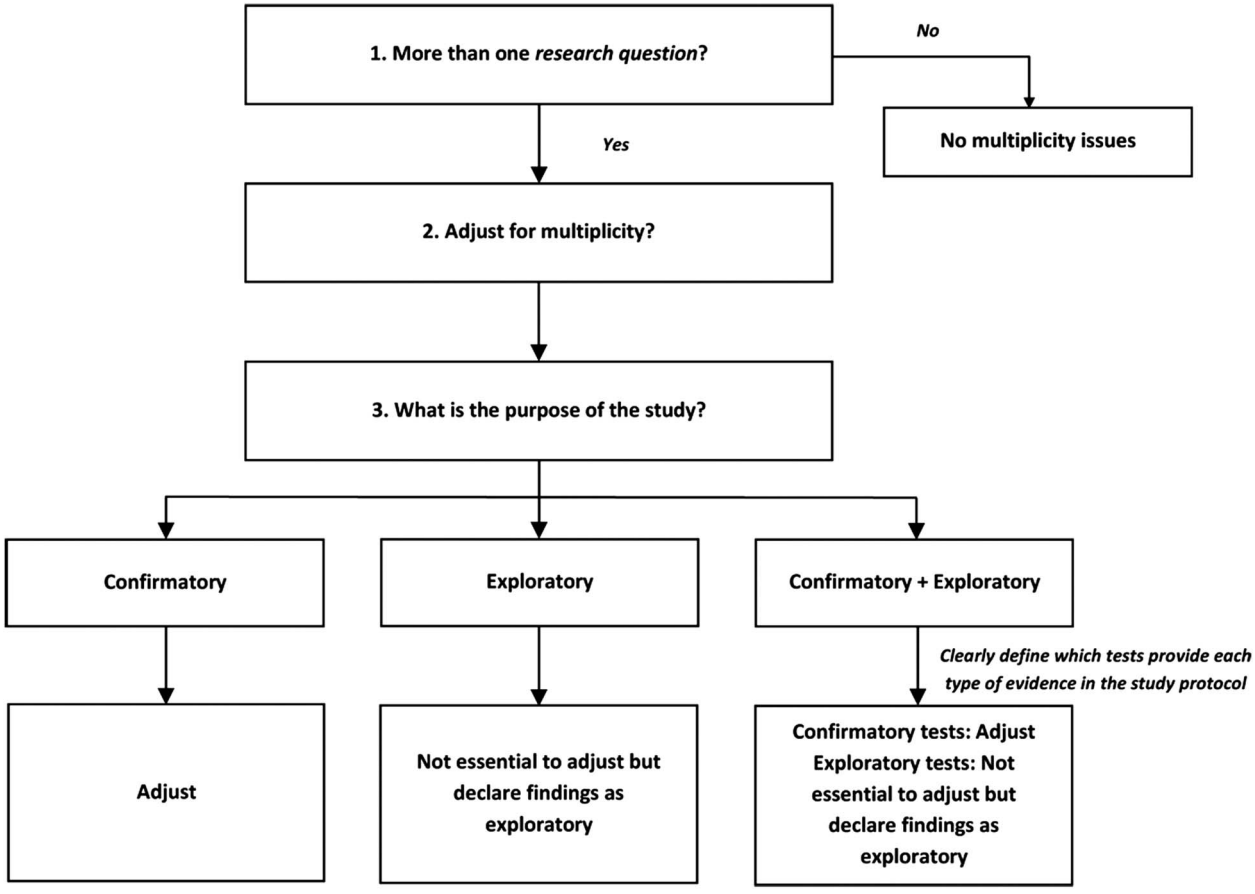

Figure 1 Flow diagram of adjustment for multiplicity.

interpretation, and it is generally preferable and more informative to conceive and present our results as estimates and CIs.[14]

For example, it is tempting to refer to a p value of 0.07 as 'bordering on significance' inferring, perhaps, that more data would inevitably yield a significant finding. In reality, even doubling the size of the trial would lead to a non-significant result about 27% of the time.[14] CIs provide an intuitive view of uncertainty which is lessened by increasing knowledge.

LESSON LEARNT

- Performing multiple hypothesis tests within a study will increase the number of false-positive findings.
- Adjustment for multiple testing consists of setting a more stringent threshold for significance than the usual 5% for each test performed.
- Adjustment for multiplicity is not always necessary.
- If the study is confirmatory, adjust results for multiplicity (figure 1).
- If the study is exploratory and results are declared as exploratory, adjustment for multiplicity is not essential (figure 1).
- Decisions regarding adjustment for multiple testing should be made prior to the start of the study and clearly stated and justified in the study protocol.

Collaborators Jonathan Cook, David Crabb, Phillippa Cumberland, Gabriela Czanner, Paul Donachie, Andrew Elders, Marta Garcia Finana, Rachel Nash, Neil O'Leary, Chris Rogers, Selvaraj Sivasubramaniam, Simon Skene, Irene Stratton, Luke Saunders, Wen Xing, Haogang Zhu.

Contributors VC, AQ designed and drafted the paper. VC, AQ, CB, NF and CJD reviewed and revised the paper.

Funding The posts of VC, AQ and CB are partly funded by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology.

Disclaimer The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

To cite Cipriani V, Quartilho A, Bunce C, *et al*. *Br J Ophthalmol* 2015;**99**:1155–1157.

Published Online First 25 June 2015

Br J Ophthalmol 2015;**99**:1155–1157. doi:10.1136/bjophthalmol-2015-306784

REFERENCES

1 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.
2 Tufail A, Patel PJ, Egan C, *et al*. Bevacizumab for neovascular age related macular degeneration (ABC Trial): multicentre randomised double masked study. *BMJ* 2010;340:c2459.
3 Patel PJ, Chen FK, Da Cruz L, *et al*., ABC Trial Study Group. Contrast sensitivity outcomes in the ABC trial: a randomized trial of bevacizumab for neovascular age-related macular degeneration. *Invest Ophthalmol Vis Sci* 2011;52:3089–93.
4 Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7:781–91.
5 Cipriani V, Leung H-T, Plagnol V, *et al*. Genome-wide association study of age-related macular degeneration identifies associated variants in the TNXB-FKBPL-NOTCH4 region of chromosome 6p21.3. *Hum Mol Genet* 2012;21:4138–50.
6 Anderson CA, Pettersson FH, Clarke GM, *et al*. Data quality control in genetic case-control association studies. *Nat Protoc* 2010;5:1564–73.
7 Bender R, Lange S. Adjusting for multiple testing—when and how? *J Clin Epidemiol* 2001;54:343–9.
8 Perneger Thomas V. What's wrong with Bonferroni adjustments. *BMJ* 1998;316:1236.
9 Pe'er I, Yelensky R, Altshuler D, *et al*. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008;32:381–5.
10 Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014;15:335–46.
11 Freemantle N. Interpreting the results of secondary endpoints and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ* 2001;322:989–91.
12 International Conference on harmonisation of technical requirements for registration of pharmaceuticals for human use (ICH) E9, Statistical Principles for Clinical Trials. February 1998. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf
13 Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract* 2004;10:307–12.
14 Wood J, Freemantle N, King M, *et al*. The trap of trends to statistical significance: how likely it really is that a near significant P value becomes more significant with extra data. *BMJ* 2014;348:g2215.