Dealing with multiple comparisons:

To adjust or not to adjust

Let's start by saying that dealing with multiple comparisons is tricky, perhaps not mathematically, but logically for sure. The wrapping of this problem should be labeled with at least two warnings, one from Oscar Wilde: 'the truth is rarely simple and never pure' and the other from Einstein: 'as far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality'. Einstein's statement was made in a lecture on geometry in relation to experience, partly highlighting his insight that Euclidean geometry fails to explain reality, so any branch of mathematics that is less intuitively obvious in its relation to reality than geometry has an even larger gap to close.

»In assessment of evidence our intention matters»

Multiple comparisons arise naturally in most scientific studies because of the need to capture or convey a lot of information with many variables. But there are many steps taken during a scientific process that also have multiplicity implications, such as variables to be included in the analysis, subgroup analysis, etc. Even a simple decision as deciding on a proper transformation has a multiplicity implication as we can search for a transformation that produces a more significant result.

First, the easy part: mathematical certainties can be derived assuming the null hypotheses are true, such that, for example, 'at 5% significance level, for every 100 tests we shall get 5 spuriously significant tests on average.' Alternatively, assuming independence and still under the null, it is virtually certain (probability =1-(1-0.05)100=0.994) that the test among the 100 tests with the smallest p-value will have a raw (i.e. unadjusted) p-value <0.05. This means that if we are naïve we would be easily misled by false positives, hence the usual warning about data fishing.

In practice, to account for multiplicity, one can simply adjust the p-value by multiplying it by the number of tests (denoted by M), or dividing the nominal significance level by M. So instead of using a level of 0.05, we would use the level 0.05/M, or, alternatively, instead of using the raw p-value, we would use the adjusted p-value obtained by multiplying the raw p-value with M. This makes it harder to declare significance. Using this so-called Bonferroni correction, we can guarantee that if we only reject hypotheses with adjusted p-value 0.05, say, then the probability of making any false rejection among all tests is ≤ 0.05. To highlight the mathematical certainty, suppose we have M null hypotheses, where a subset containing M0 hypotheses are true; let A be the event that we obtain at least one false rejection, then

Prob(A) = Prob(adjusted p-value ≤ 0.05 for at least one hypothesis)
= Prob(p-value ≤ 0.05/M for at least one hypothesis)
≤ M0*0.05/M
≤ 0.05.

Due to its simplicity, this procedure is commonly used in practice. If we want to protect against false positives, it seems reasonable to apply this procedure. The literature on multiple testing is enormous and there are numerous alternative methods. Many are improvements of the Bonferroni correction, indicating that »»»

DEALING WITH MULTIPLE COMPARISONS: TO ADJUST OR NOT TO ADJUST

multiple-testing correction is taken seriously, and when applied, it is mathematically straightforward.

Example 1: Cancer drug study. Table 1 shows the p-values in a study of a metastatic cancer drug vs placebo for ten patient characteristics. The most significant raw p-value is 0.007 for the Karnofsky index (let's keep it mysterious for now), but when corrected using the Bonferroni method, nothing is significant at the 5% level. (The last column actually gives an estimated number of false positives when we consider the variable and those above it as significant, so we do not truncate it at 1.00. We will use this later.)▫

Yet the statistical concerns are not shared universally among scientists. The clearest objection was formulated by Rothman (1990), who declared 'No adjustments are needed for multiple comparisons'. Being the first editor of the journal Epidemiology, Rothman is one of the most influential (and highly cited) epidemiologists, so it is worth understanding his arguments. Briefly, two non-mathematical presumptions are needed for application of multiplicity adjustments: (i) the 'universal null hypothesis', covering all hypotheses under consideration, is a reasonable state of nature, so that chance does cause many unexpected findings, and (ii) no one would want to investigate further something caused by chance. He argued that in a scientific process these presumptions are not true, that 'chance' is not a scientific explanation, so scientists should 'grasp at every opportunity' to understand unusual findings, and that the possibility of being misled is part of the trial-and-error process of science.

It is still the case today that most statistical results in epidemiology and medical literature are rarely adjusted for multiple comparisons, with notable exceptions in clinical trials and high-throughput molecular studies; see more below. In clinical trials, multiplicity issues arise in, for example, the choice of hypotheses to be tested, sidedness of the test, interim analyses during the trial, main analysis plan, subgroup analyses after the trial, etc. The US Food and Drug Administration's (FDA) "Guidance for Industry: E9 Statistical Principles for Clinical Trials" stipulates that these should be addressed explicitly in advance. In particular, for multiple comparisons, "adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan." Furthermore, since September 2007, in the so-called FDAAA 801 Requirement, any clinical trial of drugs or medical interventions that will seek FDA approval must be registered when the trial begins (see: www.clinicaltrials.gov), hence limiting or avoiding completely the reporting of unplanned analyses. Why are the strict guidelines not universally adopted in science in general? Imagine asking scientists to register all the hypotheses and analysis plans – including interim analyses in advance.

Example 1: Cancer drug study (continued). In the cancer drug study, suppose we know nothing about the variables prior to the study, so for us all these variables assume equal status. Then it would be naive to take the raw p-values seriously, thus forcing us to accept the adjustment and the overall null result. But suppose prior to collecting the data, because this is a study on end-stage metastatic patients, we declared that the Karnofsky index, a generalized measure of functional performance, was of primary interest, while the other variables were of secondary interest. Then, no adjustment would be necessary. Thus, as accepted in clinical trials, in assessment of evidence our intention matters. This does not feel controversial. We note that, for justifying the decision not to adjust, the primary interest in the index does not require prior knowledge or data or anything sensible; in principle we only need to declare it in advance.

Now suppose we accept we know nothing prior to the study, hence the multiplicity adjustment, but another research group who is working on the Karnofsky index contacts us to share the data. We agree to give them that variable only. Now, for them, it seems reasonable that no adjustment is needed and to conclude that the Karnofsky index is statistically significant. So, here we have a seeming paradox that, with the same data, two research groups can claim different evidence: one group cannot claim significance, but the other can. But what if we contact them? Now it seems the adjustment must be used, since we could have contacted any group working with the most significant variable. This means the mechanism of contact becomes an issue, but in practice the appearance of the 'second group' in the scene can of course be completely haphazard, e.g. via a chance encounter at a party, a friend of a friend, etc. In this social contact, how do we keep track of who brings up the topic first? It would be impossible to formalize such a process. ▫

Example 2: Single test. Surprisingly, the logical issue associated with the application of multiplicity adjustment arises even when we only perform a single test (Berger and Berry, 1987; Pawitan, 2001, p.202). Suppose a scientist comes to a statistician with a study (say with sample n=100), and the statistician performs a single test and obtains z= 2.1 (p-value=0.036). It seems uncontroversial to claim significance at α=0.05. Yes, but wait … what did the scientist plan to do were the result not significant? Suppose he planned to collect more data. So, his actual procedure is as follows:

- Collect n=100 and test if $|z_1| > c$, if significant stop.
- Otherwise, collect 100 more, and test if the overall $|z_2| > c$.
- To get an overall significance level α=0.05, we must use c=2.18, such that

$$\text{Prob}(|z_1| > c) + \text{Prob}(|z_1| \le c \text{ and } |z_2| > c) = 0.05.$$

So, the observed z=2.1 is actually not significant! As before, the statistical significance is affected not just by the data, but also by the intention of the scientist, but in this case it feels disturbing because the second stage of the study is still only a thought. (Before the reader accuses this example as perverse, adjusting for the intention here follows the FDA guideline on interim analysis of clinical trials.) ▫

Table 1. p-values in a study of a metastatic cancer drug vs placebo for ten patient characteristics.

Variables	p-value	10*p-value
1 Karnofsky index	0.007	0.07
2 Body weight	0.013	0.13
3 Tricep skin-fold	0.091	0.91
4 Hemoglobin concentration	0.236	2.36
5 Erythr sedimentation rate	0.350	3.50
6 Albumin in serum	0.525	5.25
7 Creatinine in serum	0.535	5.35
8 Bilirubin in serum	0.662	6.62
9 S-alkaline phosphatase	0.823	8.23
10 Alanine aminotransferase	0.908	9.08

These examples highlight a generic logical question: to what collection of tests do we need to apply the adjustment? To cover all tests relating to the same biological process? All tests in a single paper? If the latter, then theoretically we can avoid multiplicity correction by splitting the results into separate papers. The primary problem is that we are using the same statistic (p-value) both as a measure of evidence in a specific dataset (statistical distance between the hypothesis and the data) and as a measure of uncertainty (decision-making error rates) over hypothetical repetitions of the study. In the former adjustment for multiplicity is not an issue, but in the latter it is. However, the latter also requires a precise setup of how the study repetitions are to be done, and here intention matters, for example, in deciding which tests are to be included or prioritized. The examples show that any test can legitimately belong to distinct collections with distinct repetition studies that depend on the perspective of the experimenters. (This is not strange, e.g. a person may belong to distinct clubs with conflicting rules.) In Example 2, on seeing the observed data, the statistician's immediate reaction is to imagine hypothetical repetitions involving one test, but the scientist's intention involves hypothetical repetitions with two tests. Whose perspective is correct?

In view of this logical difficulty, how are we to react to potentially conflicting conclusions from unadjusted and adjusted analyses? We are put into this corner by an implicit demand that we make a decision. When a study is only performed once, i.e. the hypothetical repetitions remain hypothetical, we are in a never-ending unsolvable logical puzzle on how to decide. How do we break this puzzle? In clinical trials, for the key hypothesis, we break it by decree (e.g. FDA Guidelines), in essence limiting or avoiding the issue by stating the hypothesis and analysis methods in advance. Science does not follow such strict rules, but we need replication or validation studies to confirm interesting discoveries. Before further confirmation, discoveries are considered provisional and, in contrast to clinical trials, it is not necessary to make a decision about the true state of nature. However, we also note that, to confirm a discovery, it is not necessary to perform exactly the same experiment as before (i.e. the hypothetical repetitions). For example, a discovery in an observational study in humans will be substantially more credible if validated biologically in mice, and vice versa.

Eventually, in treating a study as a screening tool to identify interesting discoveries, we have to go back to the basic trade-off between type-I (false positive) and type-II (false negative) errors: protecting against one will increase the other. Strict adherence to multiple-testing adjustment protects against inflation of type-I errors and increases type-II errors, but who decides which error is more important? Different areas of science may evolve differently depending on their experience with these two errors. For example, molecular epidemiology went through the lamented candidate-gene approach to complex diseases, roughly from 1980s up to early 2000s (Hirshhorn et al, 2002; Chabris et al, 2012), where few findings were replicated.

Example 3: Publication bias and the Winner's curse. The human genome is a rich source of variables/genes. For many complex phenotypes, suppose there is no real effect or, more likely, the individual-gene effects are so tiny that the power of fundable studies is too small to detect anything. Say there are 100 research groups investigating different genes and phenotypes; each group is essentially generating random numbers. At 5% level, there will be 5 lucky groups with significant results: these are much more likely to get published, and then fail to replicate. If these are really 100 distinct groups, what can we do about this problem? No system of publication now can communicate so many negative findings to balance the false positives, so the problem seems to be an inevitable price of the scientific process.▫

Since high-throughput molecular studies, particularly the genome-wide association studies

References

Berger, JO and Berry, D (1987). The relevance of stopping rules in statistical inference. In Statistical Decision Theory and Related Topics IV (S Gupta and J Berger, Eds). New York: Springer-Verlag.

Chabris CF, et al. (2012). Most reported genetic associations with general intelligence are probably false positives. Psychological Science, 23(11), 1314-1323.

Goeman J and Solari A (2011). Multiple testing for exploratory research, with discussion. Statistical Science, 26 (4), 584-597.

Hirschhorn JN, et al. (2002). A comprehensive review of genetic association studies. Genetics in Medicine 4(2), 45-61.

Lee W, et al. (2012). Estimating the number of true discoveries in genome-wide association studies. Statistics in Medicine, 31(11-12), 1177-1189

Pawitan Y (2001). In All Likelihood. Oxford: Oxford University Press.

Rothman, KJ (1990). No Adjustments are Needed for Multiple Comparisons. Epidemiology, 1 (1), 43-46.

(GWAS), came into the scene, a single study/paper routinely performs millions of tests of single nucleotide polymorphisms (SNPs), and the genome-wide significance-level based on the Bonferroni correction is the accepted method of dealing with the huge multiplicity problem. One may argue correctly that we would be missing a lot of signals (type-II errors), but since the field had seen a lot of wasted effort at replicating false leads during the candidate-gene era, the consensus on the use of Bonferroni correction seems unchallenged. The problem in molecular epidemiology in the genomic era is that there are simply too many potential leads, so one needs a method to limit them. Note, however, that the p-values for each phenotype are usually adjusted separately, e.g. if there are one million SNPs, then the 5% significance level is divided by one million for each phenotype, regardless of the number of phenotypes, so the multiplicity adjustment is not followed consistently.

Multiple testing ideas are useful during an exploratory phase of a study. Goeman and Solari (2011) identified three sensible requirements for a multiple correction procedure: (i) not too strict, i.e. should allow possibilities of false positives; (ii) post-hoc, i.e. should allow choice after seeing the data; (iii) flexible, i.e. should allow whatever results to pursue, not just the significant ones. Instead of focusing on the probability of making any false positives, which is too strict, we can instead estimate and provide confidence lower-bounds for the number of true discoveries (=true positives). To emphasize its post-hoc feature, they called their procedure 'cherry picking'. Recent developments, for example in false discovery rate (FDR) estimation, are also in line with these requirements (e.g. Lee et al, 2012). The purpose is more to set realistic expectations for further studies rather than making final decisions about the true state of nature.

Example 1: Cancer drug study (continued). Suppose this study was performed on end-stage cachexic patients, which are characterized by severe wasting/loss of body mass and functional impairment. The top three variables in Table 1 are then of special interest, but this interest can be decided post-hoc (after seeing the data). From the third column in Table 1(10*p-value), the estimated number of true discoveries is 3-0.91≈2 (Lee et al, 2012). An application of the cherry-picking procedure (Goeman and Solari, 2011) gives 95% guarantee of at least 1 true difference from the top three findings, and 75% guarantee of at least one true difference from the second and third tests. So, even if we do not want to specify any hypothesis in advance, the analysis indicates the drug is worth studying further. ▫

To conclude, we do not mean to sound skeptical of the use of multiplicity adjustment, but only to emphasize the nuances when it is applied to real studies. We highlight some logical problems with its formal use when assessing scientific evidence, thus partly explain the lack of universal acceptance. There are areas that must pay close attention to the multiplicity adjustment, for example clinical trials, where type-I errors can have enormous human or financial costs, or high-throughput molecular studies with a flood of false positives when traditional significance levels are used. On the other hand, in areas of science where potential leads for discoveries are not abundant, formal multiplicity adjustment is not the norm; findings are considered provisional until further confirmation, and the inherently skeptical scientists pragmatically accept the trial-and-error aspect of the scientific process.

YUDI PAWITAN AND ARVID SJÖLANDER
KAROLINSKA INSTITUTET