Osteoarthritis and Cartilage 24 (2016) 763–764

Osteoarthritis and Cartilage

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

Editorial

Multiple *P*-values and Bonferroni correction

Keywords:
Statistics
Hypothesis testing
Multiplicity
P-values
Exploratory
Confirmatory

Most manuscripts submitted to Osteoarthritis and Cartilage include multiple *P*-values. Many of these manuscripts present multiplicity corrected significance levels. Some of these manuscripts also describe how the correction was performed. Few manuscripts, however, motivate why multiplicity corrections have been performed. Multiple testing is a complex methodological area with potentially grave consequences for mistakes. This is an attempt to briefly explain the basics of the problem.

A familywise error rate (FWER)

When a test of a null hypothesis is performed at a significance level of α (in practice usually 0.05), the risk of a false positive outcome of the test, the type I error rate, is α. In other words, the risk of a false positive outcome of a single test is 5%.

When it is relevant to calculate a combined error measure for a group of hypotheses, for example when a trial has two endpoints instead of one or three treatment groups instead of two, a group of hypotheses can be defined as a family, and a FWER can be calculated. For a family of k independent null hypotheses tested at a significance level of α, the probability that at least one of them will be false positive is $1 - (1 - \alpha)^k$.

As an example, an experiment with three treatment groups (A, B, C) and 5 different endpoints can be analysed with 15 independent null hypotheses, i.e., with three pairwise tests (A–B, A–C, and B–C) for each one of the 5 endpoints. When a family of 15 independent null hypotheses is tested at a significance level of 0.05, the FWER becomes 0.54, meaning that the risk of at least one false positive hypothesis test in the family is as high as 54%.

To achieve a FWER of 0.05 the significance level for each of the multiple hypothesis test must be set lower than 0.05. The Bonferroni method is used to calculate this test-specific significance level, calculated as α/k. In the above example the corrected test-specific significance level would be $0.05/15 = 0.0033$.

These calculations are based on the assumption that the tests are independent. With correlated tests the calculations become more complicated. Such problems are, for example, not uncommon in association tests of genetic data because of linkage disequilibrium between nearby markers and correlation between traits and models. Special methods have been developed for this[1].

Strategies for addressing multiplicity issues

With familywise multiplicity correction, the test-specific significance level depends on how the family is defined. For example, if only 5 of the 15 independent hypotheses in the above example were included in the test family, the Bonferroni corrected significance level would be 0.01 instead of 0.0033.

It is thus crucial that the definition of the test family is consistent with the research question[2]. A strategy for addressing multiplicity issues (and defining a test family) is therefore an important part of the development of study design.

For example, in the previous experiment with three treatment groups and 5 endpoints, one strategy may be to use Bonferroni corrections for the multiplicity created by the three treatment groups, i.e., to test all 5 endpoints with a test-specific significance level of $0.05/3 = 0.017$. This approach is very common in analyses of laboratory experiments. However, also the multiple testing of the 5 endpoints increases the type I error rate. An alternative strategy, protecting the overall type I error rate, could include the additional requirement that each one of the 5 endpoints must be statistically significant in order to claim a positive outcome of the experiment. A third strategy could be based on defining one endpoint as primary and conditioning the outcome of the entire experiment on the primary endpoint. Also other possibilities, such as the closed test procedure[3], exists.

The statistical principles for developing a multiplicity strategy are clearly described in regulatory authorities' guidelines[4] as well as in other statistical literature[5,6].

Disadvantages with Bonferroni

Bonferroni correction is an often used but controversial method. The methodology for multiplicity corrections has also been improved and newer alternatives[7] can be recommended as they are less conservative. However, multiplicity correction of the significance level always protects the FWER at the expense of the statistical power.

To maintain the risk of a false negative outcome (the type II error) at a reasonable level, the sample size needs to be increased

http://dx.doi.org/10.1016/j.joca.2016.01.008

when multiplicity corrections are performed. This is usually an economical and logistic disadvantage. Randomized clinical trials are therefore often designed with a multiplicity strategy that avoids multiplicity corrections[8].

Confirmation and exploration

The purpose of a confirmatory study, as a randomized clinical trial, is to test a pre-specified key hypothesis. A study protocol presenting the hypothesis and the statistical analysis prior to inspection of data is considered necessary, and if a single hypothesis test is not sufficient, multiplicity corrections are expected.

In contrast, an exploratory study does not need pre-specified hypotheses and includes typically large numbers of data-generated hypothesis tests. Coherent multiplicity corrections are therefore not possible, and even if one is performed it is unclear how the outcome of tests of data dependent hypotheses can be interpreted as compared to tests of pre-specified hypotheses[6].

Observational studies are typically exploratory, but randomized trials and laboratory experiments can be performed either as confirmatory or exploratory. It is important to acknowledge the difference between the two approaches. Test results from exploratory studies are generally not expected to include multiplicity correction, even if exploratory laboratory experiments often do.

Concluding recommendations

When research findings are published, the limitations of the empirical support of exploratory studies as compared to confirmatory ones should be clearly acknowledged in order to maintain consistency between the statistical results and the author's conclusions and to avoid misleading the reader.

In addition, irrespective of the type of study, if multiplicity corrections are used, the reasons for performing these correction should be explained, and the error rates used in the evaluation of the outcome, should be clearly presented to the reader.

Author contributions

JR designed the work, drafted it, revised it critically for important intellectual content, approved the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of interest

None.

Acknowledgements

The author is grateful for valuable suggestions and constructive criticism from the editor in chief.

References

1. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. Am J Hum Genet 2007;81:1158–68.
2. O'Brien PC. The appropriateness of analysis of variance and multiple comparison procedures. Biometrics 1983;39:787–94.
3. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976;63:655–60.
4. EMA. Points to Consider on Multiplicity Issues in Clinical Trials. London: European Medicines Agency; 19 September 2002. CPMP/EWP/908/99.
5. Gregori D, Baldi I, Frigo AC, Guardabasso V. Why, when and how to adjust for multiplicity in clinical trials: a perspective on regulatory activities. Biomed Stat Clin Epidemiol 2010;4:27–34.
6. Bender R, Lange S. Adjusting for multiple testing – when and how? J Clin Epidemiol 2001;54:343–9.
7. Levin B. Annotation: on the Holm, Simes, and Hochberg multiple test procedures. Am J Public Health 1996;86:628–9.
8. EMA. ICH E9 Statistical Principles for Clinical Trials. London: European Medicines Agency; September 1998. CPMP/ICH/363/96.

J. Ranstam
Mdas AB, Rotfruktsgatan 12B, SE-27154 Ystad, Sweden
E-mail address: jonas.ranstam@gmail.com.

14 January 2016