Safety Assessments and Multiplicity Adjustment: Comments on a Recent Paper

Many end points representing possible non-target effects have to be evaluated in safety assessments that compare new and accepted products. This is known as the multiple-comparison or multiplicity problem. One popular method to adjust statistical testing procedures for multiplicity is the false discovery rate (FDR) method, often implemented via adjustment of \( p \) values, such as for example provided in SAS procedure MULTTEST. FDR-adjusted \( p \) values are obtained by multiplication of the raw \( p \) values with factors between 1 and \( m \), where \( m \) is the number of tested hypotheses. Let \( p_1 \leq \ldots \leq p_m \) be the ordered \( p \) values for \( m \) end points. Then, the FDR-adjusted \( p \) values according to a linear step-up algorithm are sequentially calculated as

\[
p(\tilde{\alpha}) = \frac{p_j^{1/m}}{j/m} \leq \min\left(p(j) \frac{m}{j} \right) \leq \frac{p_j(m)}{j},
\]

for \( j = 1, \ldots, m \).

Recently, Hong et al. published an evaluation of the European Food Safety Authority (EFSA) framework for safety assessment of genetically modified (GM) crops using a rat 90 day feeding study, which is a compulsory part of the safety assessment according to current European Union (EU) legislation. The appropriateness of these animal studies and the EFSA framework on how to conduct such studies are both under discussion. For example, the EU research project GRACE (http://www.grace-fp7.eu) has performed and evaluated four 90 day and one 1 year study contributing to this discussion (see the study by Schmidt et al. and references therein). Another currently ongoing EU research project is G-TwYST (https://www.g-twyst.eu), which is evaluating two 90 day studies and one combined chronic/carcinogenicity (2 year) study. Hong et al. also assessed the appropriateness and applicability of the EFSA recommendations using a 90 day study and a battery of statistical approaches, including retrospective and prospective power analyses. This comment is not the place to give a full appraisal of all aspects of this discussion. The discussion here is restricted to just one element of the statistical approach used, which is the treatment of the multiplicity as a result of many end points. Hong et al. evaluated a very large number of end points and adjusted the \( p \) values of their tests according to the FDR method. The maximum number of end points for each of the sexes was \( m = 146 \); therefore, FDR-adjusted \( p \) values are between 1 and up to 146 times as large as the raw \( p \) values (FDR adjustment was performed for the set of all end points that were reported across sex and separately for the male- or female-specific comparisons and, thus, may have been lower in practice, but exact values are not given). The main result of Hong et al. regarding the comparisons between test and control groups is that “no treatment-related differences were observed”. This can be contrasted with the detailed comparisons in Appendix D of the paper, where 32 out of 816 of the 95% confidence intervals for observed differences do not contain the value 0 and, therefore, indicate significant differences in an unadjusted test. Note that the rate of significant results (3.9%) is close to the expected rate of 5% false positives that is expected under a null hypothesis of equality for all end points and, therefore, in itself, is not a reason for concern about safety. However, the reported absence of any statistically significant difference should be seen as the direct consequence of using the FDR adjustment. Clearly, with no “discovery” at all in this set of results, the false discovery rate is zero by definition, and in this respect, the methodology can be said to have operated very effectively. In summary, FDR adjustment is not a minor detail but is a main factor that determines the test results.

I have two serious concerns about the methodology in this paper. First, the use of standard FDR correction or any other multiple-testing scheme makes no sense in food safety testing. It controls false discoveries and is therefore connected to difference testing, where the null hypothesis is equality of means and false positives are considered the error of the first kind: you want to have a small probability of erroneously reporting a difference. This is useful in studies that set out to find differences between groups, perhaps to find new explanations for biological phenomena or effective treatments. However, in the context of safety or equivalence testing, the purpose is to demonstrate safety with a chosen confidence level. Therefore, the statistical hypotheses are reversed: the null hypothesis is that some difference exists, and we want to show equivalence by rejecting such a null hypothesis (some possible approaches allowing for end points with widely different variations have been described). In equivalence testing, the errors of the first kind are the false negatives rather than the false positives, to guarantee a small probability of erroneously reporting equivalence. Consequently, the commonly used methods for multiplicity correction including FDR are addressing the wrong type of error and should not be used in safety assessments.

Second, contrary to the tests used to report results by Hong et al., the statistical power analyses in the same paper (both prospective and retrospective) do not use FDR adjustments. Therefore, the results of these power analyses cannot be interpreted as statements about the statistical power obtained using the FDR-adjusted tests. Clearly, the statistical power for any end point separately is much lower than stated (because the \( p \) values are adjusted upward). The potential danger of the paper is the message that its approach would be an appropriate procedure because (1) a high power of the difference tests for the proposed effect sizes seems to be attained and, at the same time, (2) not a single statistically significant difference is obtained. However, the statistical approaches followed for 1 and 2 (without and with FDR correction, respectively) are inconsistent. It is misleading to present FDR-adjusted test results together with power analyses that do not incorporate these adjustments.

Received: October 3, 2017

Published: February 18, 2018
As an additional point, Hong et al. also claim that FDR adjustments would be endorsed by EFSA. Whereas EFSA in its guidance \(^1\) did acknowledge the multiplicity problem (“the issue of multiple testing [...] should be addressed”), they have however not given an endorsement of FDR adjustment. Instead, EFSA \(^2\) leaves the matter to the statistical analyst (“any methods used to adjust for multiplicity should also be clearly documented and referenced”). EFSA \(^3\) already concluded on this: “FDR as usually applied (i.e. in a context of difference testing) is a property of the subset of endpoints for which a significant difference has been found. It does not address the endpoints for which no significance has been found and therefore FDR applied to difference testing does not seem sufficient as a measure in GMO risk assessment. It could be of interest to adapt the FDR concept for equivalence testing, i.e. for a situation where hypotheses are reversed, but the GMO Panel is not aware that this has yet been done.” By now, alternative methods for multiple or multivariate equivalence testing for safety evaluations have been proposed, \(^{11-14}\) which are currently under debate.

Hilko van der Voet* 
Wageningen University & Research, 6708 PB Wageningen, Netherlands

**AUTHOR INFORMATION**

**Corresponding Author**
*E-mail: hilko.vandervoet@wur.nl.

**ORCID**
Hilko van der Voet: 0000-0003-0802-7382

**Notes**
The author declares no competing financial interest.

**REFERENCES**

(1) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 1995, 57, 289—300.

(2) Hong, B.; Du, Y. Z.; Mukerji, P.; Roper, J. M.; Appenzeller, L. M. Safety assessment of food and feed from GM crops in Europe: Evaluating EFSA’s alternative framework for the EU’s 90-day feeding study. *J. Agric. Food Chem.* 2017, 65, 5545—5560.

(3) European Food Safety Authority (EFSA). Guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed. *EFSA J.* 2011, 9, 2438.

(4) European Commission (EC). Commission Implementing Regulation (EU) No 503/2013 of 3 April 2013 on applications for authorisation of GM food and feed in accordance with Regulation (EC) No 1829/2003 of the European Parliament and of the Council and amending Commission Regulations (EC) No 641/2004 and (EC) No 1981/2006. *Off. J. Eur. Communities: Legis.* 2013, 56, 1—52.

(5) Schmidt, K.; Schmidtke, J.; Schmidt, P.; Kohl, C.; Wilhelm, R.; Schiemann, J.; van der Voet, H.; Steinberg, P. Variability of control data and relevance of observed group differences in five oral toxicity studies with genetically modified maize MON810 in rats. *Arch. Toxicol.* 2017, 91, 1977—2006.

(6) van der Voet, H.; Perry, J. N.; Amzal, B.; Paoletti, C. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. *BMC Biotechnol.* 2011, 11, 15.

(7) Meyners, M. Equivalence tests – a review. *Food Qual. Prefer.* 2012, 26, 231—245.

(8) Kang, Q.; Vahl, C. I. Statistical analysis in the safety evaluation of genetically-modified crops: Equivalence tests. *Crop Sci.* 2014, 54, 2183—2200.

(9) van der Voet, H.; Goedhart, P. W.; Schmidt, K. Equivalence testing using existing reference data: an example with genetically modified and conventional crops in animal feeding studies. *Food Chem. Toxicol.* 2017, 109, 472—485.

(10) European Food Safety Authority (EFSA). Statistical considerations for the safety evaluation of GM foods. *EFSA J.* 2010, 8, 1250.

(11) Qiu, J.; Cui, X. Q. Evaluation of a statistical equivalence test applied to microarray data. *J. Biopharm. Statist.* 2010, 20, 240—266.

(12) van Dijk, J. P.; de Mello, C. S.; Voorhuijzen, M. M.; Hutton, R. C. B.; Arisi, A. C. M.; Jansen, J. J.; Buydens, L. M. C.; van der Voet, H.; Kok, E. J. Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier. *Regul. Toxicol. Pharmacol.* 2014, 70, 297—303.

(13) Pallmann, P.; Jaki, T. Simultaneous confidence regions for multivariate bioequivalence. *Stat. Med.* 2017, 36, 4585—4603.

(14) Vahl, C. I.; Kang, Q. Statistical strategies for multiple testing in the safety evaluation of a genetically modified crop. *J. Agric. Sci.* 2017, 155, 812—831.