Equivalence Testing Approaches in Genetically Modified Organism Risk Assessment

Hilko van der Voet* † and Claudia Paoletti‡

†Wageningen University & Research, 6708 PB Wageningen, Netherlands
‡European Food Safety Authority (EFSA), 43126 Parma, Italy

ABSTRACT: Since 2011, the European Food Safety Authority (EFSA) has implemented combined difference and equivalence testing of agronomic, phenotypic, and composition data in the risk assessment of genetically modified crops. A short perspective is provided on misunderstandings that have shown up in published criticisms of the approach to equivalence testing, different viewpoints regarding the questions to be answered, and new developments in statistical modeling.

KEYWORDS: European Food Safety Authority, genetically modified organism, risk assessment

Since 2011, the European Food Safety Authority (EFSA) has implemented difference and equivalence testing of crop agronomic, phenotypic, and composition data as complementary approaches in the risk assessment of genetically modified (GM) crops. In several papers written by biotech company scientists, even the term “inconsistent outcomes” for the obvious fact that, with a limited amount of data, different results are obtained across multiple studies, generalizing this to the statement “These results call into question the consistency and value of this approach”. In contrast, it should be considered normal that results are variable when a large number of hypotheses are tested repeatedly in different studies. A better statistical approach for the Herman et al. data, in line with EFSA guidance, would have been provided by a joint analysis of all four studies in their example. In a comment on the original publication of the EFSA method, industry scientists agreed that taking proper account of natural variation among commercial varieties was appropriate but still complained that the EFSA method leads to study-specific equivalence limits. At the time, the response to these comments already made clear that estimating natural variation requires data and that this inevitably leads to study-dependent results. In line with this, it has been found that equivalence tests for GMO safety using fixed limits is too rigid and that it is preferable to define safety ranges based on results measured in the same field trials, as was already concluded by industry researchers as well. This implies study-dependent outcomes of any approach.

A second point of criticism has been the low power of the method in specific case studies. In a recent example, a non-genetically engineered (non-GE) soybean variety (Maverick), used as an isoline for various GM lines in four field studies, was compared to selected reference varieties. Following the EFSA approach, it was classified in category I (equivalent) for 77, 75, 80, and 47% of the analytes and category II (equivalent more likely than not) for 9, 16, 9, and 16% of the analytes. This leaves 14, 9, 11, and 37% of the analytes for which further interpretation with respect to the consequences for safety is required according to EFSA. The interpretation of these simple results by the authors is mistaken. They claim that “failure to conclude equivalence [...] is a weakness of the method”, because “[t]he isoline is [...], by definition, equivalent to the non-GE crop”. However, this reasoning fails to distinguish between a state of nature (Maverick is a commercial variety and in that sense equivalent to other reference varieties) and the outcome of a statistical test (can we demonstrate with high confidence that Maverick is equivalent to the reference varieties?). The EFSA approach has been developed from the starting point that a risk assessor should always wish for a high confidence for equivalence statements, because the objective is to maximize consumer protection, i.e., minimize the risk of false negatives. This means that the method is designed to avoid too many false equivalence statements (error of the first kind, claiming equivalence while not true). This automatically implies the acceptance of a higher probability of not declaring equivalence, while in fact equivalence is true (error of the second kind). Equivalence tests are used as a screening method, with further attention needed in case equivalence is not established. This is a consistent and valid approach. The found percentages for category I cited above indicate the statistical power of the method when applied to the specific example. Therefore, the criticism boils down to disappointment that the statistical
power of the equivalence test in these case studies is not as high as, e.g., 95%. The expected power will vary with the position of the expected value for the test (in this example, the isoline) within the range of the reference values. To be specific, the power will be maximal if the test line has its expected value at the center of the reference range, and it will decrease if the test line expected value is displaced toward the border of this range. It is equally obvious that not all reference varieties will have their end points in the middle of the range. Some of them will be on the outside borders, and therefore, the power to show that these are equivalent will be lower. In conclusion, the equivalence test method will be most effective for test lines with expected values in the interior of the reference range but less so for test lines near the borders. Note again that a less than optimal power is not a requirement of any consistent statistical testing method and even less a “weakness of the method”. Rather, the causes for a less than optimal power are the use of reference varieties that (1) are asymmetrically distributed around the expected value of the test non-GM line and/or (2) have low variation among them.

The selection of the test line, isoline, and the reference varieties is not made by EFSA but by the applicant. The only requirements from EFSA are that reference varieties should be commercial lines and should be suitable for the environmental conditions and customary agronomic practices of the sites chosen. In the current regulatory system, applicants are responsible for the selection of appropriate non-GM test materials. A priori prescriptive guidance for such decisions cannot be made given the intrinsic complexity of the topic and the specificity of the considerations to be made on a case-by-case basis. Therefore, applicants are requested to provide explicit justification of the criteria followed to make this decision in each technical dossier. For a given test line, the choice of the isoline might be straightforward or not. In this specific case, it seems straightforward, because Maverick is a commercial variety that is both the parent and the backcrossing target of the genetic modifications.

Within the frame of EFSA guidance document requirements, the detailed justification of the criteria followed to select reference varieties is the responsibility of the applicant, because in most situations more than one choice is possible. In the paper on the Maverick case study, no information could be found how the reference varieties were selected and if other choices would have been possible. It might well be that, in this specific case study, there was little possibility to choose other non-GM reference varieties, because soybean is already GM for a great majority of the crop grown in the U.S. This point raises fundamental questions about the relation between the concept of safe use and the actual use of GM crops on a large scale. Such questions are outside the scope of the current discussion. As long as the requirement that reference varieties are non-GM remains, it may simply mean that there are fewer options for the definition of the reference set for equivalence testing and that optimal selections for specific GM test lines cannot be made. The consequence could very well be that the equivalence test might have a power lower than what would have been possible with a wider and more representative range of reference varieties. Note, however, that under all circumstances, the equivalence test remains useful by focusing on consumer protection (having a low error rate for false equivalence statements) and guarding against wrong conclusions from difference tests, which may show differences that are statistically significant but not relevant in terms of the reference variation or may fail to show differences that are relevant.

A third point of criticism is that the EFSA equivalence test focuses on a direct comparison of the GM line to the commercial reference varieties instead of a comparison to the near-isogenic comparator. This is especially relevant in cases where the near-isogenic comparator itself is very different from the non-GM reference variety population. In such cases, EFSA has concluded that “it may indicate that the GM plant and its conventional counterpart are derived from varieties with characteristics not present in the non-GM reference varieties and, consequently, the test material may not have been chosen appropriately”. The selection of the isoline and non-GM reference varieties should be performed with sufficient care to ensure, as much as possible, that the variation of commercial reference measurements is representative of the variability expected under the environmental conditions of the selected sites and that the expected isoline measurements are covered by the variability observed among the reference commercial varieties with a history of safe use. If not, then equivalence test results, such as reported in the case study by Herman et al., may be the consequence.

It is understandable that biotech companies and public authorities may have different a priori views on equivalence testing. For biotech companies, the main question is how the existing genotypes are being changed, and therefore, the focus is on the GM trait effect. For public authorities, the first question is how new GM crops compare to commercial reference varieties used in agricultural practice, for which a history of safe use can be assumed. Both are valid points of view answering fundamentally different questions. In the specific context of animal feeding studies, it has been proposed to compare the difference between the GM line and the isoline to typical differences between any two reference varieties. Such an approach may be explored further. Note that this would address the main concern of industry scientists. Currently, this new method is being adapted to crop composition field studies as well.

When the statistical approach described in the EFSA guidance document cannot be implemented, applicants should submit a proper statistical analysis and discuss its implications for the risk assessment. The Maverick example seems to be a case where the isoline is itself quite different from the selected non-GM references, at least for several analytes. An obvious question is whether this could have been prevented by a more careful selection of the non-GM reference varieties in the field studies, for which nothing is reported by the authors. If it would have been impossible to perform a more suitable selection of commercial references, EFSA allows for alternative approaches. In conclusion, the current EFSA approach is fully consistent and performs well when references are selected appropriately, but the approach was misinterpreted in the cited paper.

We do not want to convey the impression that statistical methods underlying risk assessment cannot or should not be developed further. Science will never stop developing, and consequently, science-based methodology must regularly be evaluated. In fact, much work has been performed over the past decade to propose improved methodology. For example, Kang and Vahl proposed a one-step approach rather than a two-step approach using a generalized fiducial inference method. This idea was taken up in the specific context of animal feeding studies, and another criterion was proposed to
obtain a sufficiently high power for the equivalence test. This criterion also implemented testing against the control instead of the references while still retaining the importance of the reference collection to estimate variation. With regard to the difficulty to obtain sufficient reference data in any single study, the use of databases was proposed in several publications. However, it should be evaluated if these databases are sufficiently representing the design of the experiments and allow for distinguishing genotypic from environmental variation. With regard to the multiplicity problem, it is currently being discussed whether univariate or multivariate methods are most suitable and whether the false discovery rate (FDR) is apt or inapt to control false non-discoveries. It may be time to compare and evaluate the different proposals and improve the science-based methodology if appropriate and where possible, ensuring an open and constructive discussion involving all stakeholders.

■ AUTHOR INFORMATION

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

ORCID

Hilko van der Voet: 0000-0003-0802-7382
Claudia Paoletti: 0000-0001-8807-8870

Notes

The authors declare no competing financial interest. Claudia Paoletti is employed by the EFSA. The position and opinion presented in this article are those of the authors and do not necessarily represent the views or scientific works of EFSA.

■ REFERENCES

(1) European Food Safety Authority (EFSA). Scientific Opinion on Guidance for risk assessment of food and feed from genetically modified plants. EFSA J. 2011, 9 (5), 2150.
(2) 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.
(3) Ward, K. J.; Nemeth, M. A.; Brownie, C.; Hong, B.; Herman, R. A.; Oberdoerfer, R. Comments on the paper “A statistical assessment of differences and equivalences between genetically modified and reference plant varieties” by van der Voet et al. 2011. BMC Biotechnol. 2012, 12, 13.
(4) Jiang, C.; Meng, C.; Schapaugh, A. Comparative analysis of genetically-modified crops: Part I. Conditional difference testing with a given genetic background. PloS One 2019, 14, e0210747.
(5) Herman, R. A.; Huang, E.; Fast, B. J.; Walker, C. EFSA genetically engineered crop composition equivalence approach: Performance and consistency. J. Agric. Food Chem. 2019, 67, 4080–4088.
(6) van der Voet, H.; Perry, J. N.; Amzal, B.; Paoletti, C. Response to Ward et al. 2012. BMC Biotechnol. 2012, 12, 13.
(7) Oberdoerfer, R. B.; Shillito, R. D.; de Beuckeleer, M.; Mitten, D. H. Rice (Oryza sativa L.) containing the bar gene is compositionally equivalent to the nontransgenic counterpart. J. Agric. Food Chem. 2005, 53, 1457–1465.
(8) Hothorn, L. A.; Oberdoerfer, R. Statistical analysis used in the nutritional assessment of novel food using the proof of safety. Regul. Toxicol. Pharmacol. 2006, 44, 125–135.
(9) European Food Safety Authority (EFSA). Guidance on the agronomic and phenotypic characterisation of genetically modified plants. EFSA J. 2015, 13 (6), 4128.
(10) 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.
(11) Engel, J.; et al. Equivalence tests using crop compositional data (manuscript in preparation).
(12) Kang, Q.; Vahl, C. I. Statistical analysis in the safety evaluation of genetically-modified crops: Equivalence tests. Crop Sci. 2014, 54, 2183–2200.
(13) Sult, T.; Barathet, V. J.; Bennett, L.; Edwards, A.; Fast, B.; Gillikin, N.; Launis, K.; New, S.; Rogers-Szuma, K.; Sabbatini, J.; Srinivasan, J. R.; Tilton, G. B.; Venkatesh, T. V. Report: Release of the International Life Sciences Institute Crop Composition Database Version 5. J. Food Compos. Anal. 2016, 51, 106–111.
(14) Paoletti, C.; Favilla, S.; Leo, A.; Neri, F. M.; Broll, H.; Fernandez, A. Variability of crop’s compositional characteristics: What do experimental data show? J. Agric. Food Chem. 2018, 66, 9507–9515.
(15) van Dijk, J. P.; Souza de Mello, C.; Voorhuijzen, M. M.; Hutten, R. C. B.; Maisonneuve Arisi, A. C.; 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.
(16) Pallmann, P.; Jaki, T. Simultaneous confidence regions for multivariate bioequivalence. Statistics in Medicine. 2017, 36, 4585–4603.
(17) Engel, J.; van der Voet, H. G-TwYST Harmonisation of Statistical Methods for Use of Omics Data in Food Safety Assessment; Biometris, Wageningen, Netherlands, 2018; Report 41.05.18, DOI: 10.18174/455159.
(18) European Food Safety Authority (EFSA); Aguilera, J.; Aguilera-Gomez, M.; Barrucci, F.; Cocconcelli, P. S.; Davies, H.; Denslow, N.; Lou Dorne, J.; Grohmann, L.; Herman, L.; Hogstrand, C.; Kass, G. E. N.; Kille, P.; Kleter, G.; Nogue, F.; Plant, N. J.; Ramon, M.; Schoonjans, R.; Waigmann, E.; Wright, M. C. EFSA Scientific Colloquium 24—Oomics in Risk Assessment: State of the Art and Next Steps 2018.
(19) Kok, E.; van Dijk, J.; Voorhuijzen, M.; Staats, M.; Slot, M.; Lommen, A.; Venema, D.; Pla, M.; Corujo, M.; Barros, E.; Hutten, R.; Jansen, J.; van der Voet, H. Omics analyses of potato plant materials using an improved one-class classification tool to identify aberrant compositional profiles in risk assessment procedures. Food Chem. 2019, 292, 350–358.
(20) 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.
(21) van der Voet, H. Safety Assessments and Multiplicity Adjustment: Comments on a Recent Paper. J. Agric. Food Chem. 2018, 66 (9), 2194–2195.