EUROPEAN COMMISSION

Community Research

G-TwYST Stakeholder Consultation

Stakeholder comments and questions and project team responses on draft results of the 90-day animal toxicity study with GM maize (Study B)

G-TwYST
GMP Two Year Safety Testing
632165 B/2016/GLP

19 March 2018

Draft Report

G-TWYST
GM PLANTS TWO YEAR SAFETY TESTING

Acknowledgment and Disclaimer

The authors of this document thank all project partners for their valuable contributions and comments on draft versions of this document.

This document expresses the view of the G-TwYST consortium, and does not reflect an official opinion of the European Commission. Responsibility for the information and views expressed therein lies entirely with the authors.

In the whole document, the acronym "**G-TwYST**" has been used to refer to the project.

Editor: Armin Spök

Authors of G-TwYST Team Responses: Roger Allison, Pablo Steinberg, Hilko van der Voet, Ralf Wilhelm.

"GMP Two Year Safety Testing" (G-TwYST) is a Collaborative Project of the Seventh Framework Programme of the European Community for Research, Technological Development and Demonstration Activities.

Grant agreement no: 632165

Project duration: 21 April 2014 – 20 April 2018

Project website: www.g-twyst.eu

Table of Content

1. Introduction to the Draft Report

This document presents the comments on the study plan provided by stakeholders in writing in response to documents summarising draft results of a 90-day animal toxicity study conducted with GM maize along with responses to those comments by the G-TwYST team members. The document is still in a draft status as more comments are still pending and will be added at a later stage.

Objective

Openness and inclusiveness are key elements in G-TwYST's stakeholder involvement activities. The objective of this consultation was to trigger comments and questions from stakeholders which could than be used to improve the final report.

Scope

The consultation was to focus on issues related to the draft results of a 90-day animal toxicity study conducted with 11% and 33 % inclusion rate of GM maize NK603 (G-TwYST Study B).

Process

A broad invitation was circulated to some 700 stakeholder contacts representing Competent Authorities, both at the EU and the national level of all 28 Member States, industry, farming and professional organisations, civil society organisations (CSOs) and scientists. They received a first announcement in December 2017 followed by a more detailed invitation with a preliminary agenda in January 2018. Participation in the workshop was open to anyone representing one of the stakeholder categories mentioned above. 35 stakeholders registered for the written consultation and/or the workshop and signed a Non-Disclosure Agreement in order not to jeopardize academic journal publication in the next steps. Registered participants were provided access to the following documents on a password protected server:

> Goedhart, W. & van der Voet, H. (2017). G TwYST Study B. a 90-day toxicity study in rats fed GM maize NK603. Statistical report. Report 31.10.17, Biometris, Wageningen, The Netherlands.
>
> Goedhart, W. & van der Voet, H. (2017). G TwYST Study B. a 90-day toxicity study in rats fed GM maize NK603. Statistical report Appendices.
>
> Alison, R.H. (2018). Histopathology Phase Report to:90-Day subchronic toxicity study in rats fed GM maize NK 603.
>
> G-TwYST Study B data: Study design, randomisation, outliers, feed consumption, haematology, organ weight, blood & urine, clinical chemistry, phagocytes, CD markers (spleen, thymus, lymph nodes, bone marrow), proliferative activity of lymphocytes

Response

Four sets of comments (Competent Authority, industry, research, and CSO) comprising a total of 17 comments were received. The relatively lower number of comments received is likely to be attribute to time pressure, the absence of a draft interpretation of the data by the project team, as well as the lack of novelty of this type of 90-day study. Extensive stakeholder consultations were held between 2012 and 2015 on 90-day animal toxicity studies with GM maize MON810 in course of the GRACE project.

The 17 comments were organised according to categories and anonymized. The organisations that provided comments will be acknowledged in the final version of this report. This version will then also include the comments on the draft results of the long-term animal toxicity studies conducted in G-TwYST.

2. Stakeholder Comments and Project Team Responses

Topic	Subtopic	Comment	Comment made by	Response
General	Non-standard endpoints	Stakeholder consultation for the project has been robust and resulted in a 90-day study design that included standard and non-standard endpoints per the international standard for 90-day studies: OECD Test Guideline 408, "Repeated Dose 90-day Oral Toxicity Study in Rodents". It is clear the studies are designed to allow evaluation of the product's safety. However, the use of non-standard endpoints and the use of non-standard statistical methods (equivalence testing) may complicate interpretation of results to an unknown degree due to an absence of data demonstrating the normal range of variability in these parameters for rats of the same strain, age, and length of study (i.e. historical control data).	**Industry**	G-TwYST is a research project, therefore not only standard protocols have been followed, but also several potentially useful options have been investigated. This is true both for the choice of endpoints and for the choice of statistical methods. The main research object is the methodology to conduct safety assessments using animal feeding studies. There is no implication that any of the methods should be part of regulatory practice without further evaluation. All non-standard endpoints and non-standard statistical methods have been added with the purpose to allow a better and where possible more direct interpretation of the study results. The new statistical method of equivalence testing helps in the interpretation of results because it scales the observed confidence intervals for all differences to the observation ranges observed in historical data. It is then easy to see when results (even if they give significant differences) are still compatible with variation in the historical data (green area in the plots). There was indeed a lack of historical data for several endpoints, and for those cases the method could not be applied for the current study results.
General	Risk hypothesis	According to EFSA (2011) "Toxicological assessment should demonstrate that the intended effect(s) of the genetic modification has no adverse effects on human and animal health, and that the unintended effect(s), which have been identified or assumed to have occurred based on the preceding comparative molecular, compositional or phenotypic analyses, have no adverse effects on human and animal health." Hence feeding studies should	**Competent Authority**	In an explanatory statement, EFSA (2014) has made a distinction between two scenarios: scenario 1, where relevant changes and/or specific hazards are identified, and scenario 2, where this is not the case. In G-TwYST we are operating in scenario 2, therefore no 'hazard hypotheses' were developed EFSA (2014) has indicated that in this case the absence of specific toxicological liability does not allow to formulate a specific (hazard) hypothesis and to define an effect size of potential toxicological relevance. However, whereas there are no direct hazard hypotheses regarding effect sizes with a toxicological relevance, it is still possible to

Topic	Subtopic	Comment	Comment made by	Response
		analyze those hazard hypotheses developed at previous steps of the risk assessment. For us it is not clear if and how specific hazard hypotheses were developed and analyzed in the course of the G-TwYST feeding studies. We would like to know if hazard hypotheses regarding food-feed safety were developed based on i) the intended expression of CP4EPSPS (molecular level), ii) intended change in agronomic practice, i.e. glyphosate application (phenotypic level) or iii) unintended alterations in plant composition (compositional level). If hazard hypotheses were developed we would like to know how these hypotheses were addressed within the feeding studies and if certain variables were selected to analyze these hazard hypotheses. The analyzed variables at the G-TwYST studies differed in part from those of the GRACE studies and it would improve transparency if the reason for the alterations in selected variables would be provided. EFSA (2011). Scientific Opinion on Guidance for risk assessment of food and feed from genetically modified plants. EFSA Journal 2011;9(5): 2150.		formulate hypotheses about effect sizes that would be in agreement with differences observed in historical data sets for varieties with a history of safe use. This is the essence of the proposed statistical method of equivalence testing. Note that non-equivalence does not imply toxicological relevance. However, equivalence does imply toxicological non-relevance. Some variables related to possible immunological and endocrinological functions analysed in the frame of the G-TwYST project were indeed different from those in the GRACE project. The reason for analyzing these additional variables was that the evaluation of these endpoints was considered extremely important by several stakeholders at the first G-TwYST Stahekolder Meeting.
Data interpreta-tion		According to Goedhart & van der Voet (2017) the data interpretation of results were based on the following	**Competent Authority**	a. In the absence of biologically or toxicologically relevant effect sizes in subchronic studies, G-TwYST has operated under scenario 2 of EFSA (2014), and has as a main approach compared observed differences with

Topic	Subtopic	Comment	Comment made by	Response
approach		a. Results from G-TwYST studies were compared to thresholds derived from Hong et al. (2017). It remains unclear if these thresholds address specific hazards. Hong et al (2017) stated that thresholds (i.e. effect sizes for the biological alterations) "should not be considered synonymous with biologically or toxicologically relevant effects in subchronic studies", hence it should be discussed how the application of those thresholds contribute to the assessment of food-feed safety. b. Alterations in several parameters were combined and associated with certain pathological endpoints (here: "liver disorder" and "kidney disorder"). We appreciate this kind of functional analysis. However the reasons for the selection of these specific pathological endpoints should be explained. Were those pathological endpoints addressed due to an a priori defined hazard hypothesis or did the study results indicate an effect on those pathological endpoints? Hong B, Du Y, Mukerji P, Roper JM, Appenzeller LM (2017). Safety assessment of food and feed from GM crops in Europe: Evaluating EFSA's alternative framework for the rat 90-day feeding study. Journal of Agricultural and Food Chemistry, 65(27): 5545-5560.		variation observed for non-GM varieties in historical studies with the same species in the same test facility. As an alternative approach, G-TwYST has also used the 'targeted effect sizes' used by Hong et al. (2017), which originate from toxicological considerations regarding longer-term endocrine, chronic or carcinogenicity studies. We agree that a discussion about the application of these thresholds in the case of subchronic studies is needed, and the reported results may be helpful in that discussion. It was observed, for example, that the bandwidth within the 'targeted effect sizes' is typically larger than the variation observed in the historical data of non-GM varieties. b. The triplets of variables analyses under the labels "liver disorder" and "kidney disorder" were defined a priori, and were not based on observed results.

Topic	Subtopic	Comment	Comment made by	Response
Statistics	Non-standard statistical analysis	Additionally, it should be noted that the first 90-day study (and presumably the other studies within G-TwYST) include novel approaches to statistical evaluation of the study data. While innovation certainly can be desirable, the rationale to include and select these alternate methods of statistical analysis should be explained. Several statistical reanalyses were conducted with the data, and re-testing in this manner complicates interpretation of the results and may inflate error rates (e.g., false positives). EuropaBio would like to have a better understanding of why these approaches were deemed appropriate by the project coordinators and how their performance will be evaluated compared to the standard approaches. Perhaps more importantly, the publication of these results could be viewed as an endorsement of the non-standard endpoints and non-standard statistical methods of this particular study with implied advocacy towards inclusion in the OECD 408 guideline. EuropaBio considers that the added value of the proposed additional items would require extensive testing and that evaluation in one study is not sufficient for these elements to be proposed to be added to the current 90-day protocol performed for risk assessment purposes.	**Industry**	Novel statistical approaches have been developed in G-TwYST as a response to the confusion caused by the traditional approach which focuses on testing the statistical significance of differences. In the traditional difference tests, stating that there is a difference whereas in reality this is not the case is considered the error of the first kind (and statistical procedures are designed to control the error rate α). In contrast, in safety assessments a failure to find a difference (and thus declare safety) should be considered as the error of the first kind. This leads to the general statistical framework of equivalence testing, which has well-known uses in regulatory practice for medicines (FDA, EMA). The G-TwYST innovations have adapted such methods for the specific field of GM safety testing. G-TwYST is not presenting a safety assessment as such, but provides a scala of statistical approaches, both traditional and novel, to be evaluated. There is no implication that all of these statistical approaches should be used together in future safety assessments, and therefore there is also no inflation of error rates due to using multiple methods. The evaluation of the different approaches should be made in a wider discussion outside this project, and the final phase of G-TwYST can only initiate this discussion, e.g. through these comments and in the stakeholder meeting. Of course, as scientists, we endorse our innovative method of equivalence testing. However, we fully realise that further extensive testing and possibly fine-tuning, e.g. for other types of data such as pland omics data, is needed before this method could become part of regulatory protocols.

Topic	Subtopic	Comment	Comment made by	Response
Statistics	Non-standard statistical analysis	Six types of statistical methods/assessments have been performed. I propose that it is not the intention to set a standard with this. For any kind of study a correct statistical set-up and assessment needs to be performed, and it should suffice to do one type of statistical analysis, unless there are scientific reasons to doubt the robustness of the outcome of that assessment.	**Research Organisation**	G-TwYST is a research project, therefore not only standard protocols are being followed, but also several potentially useful options are being investigated. There is no intention to suggest that all six types of statisitcal methods should become part of regulatory practice.
Statistics	Equivalence testing	The proof of equivalence is not a proof of safety. However toxicological studies should indicate safety of GM-derived food/feed. What is the informative value of the equivalence test in terms of hazard identification?	**Competent Authority**	Equivalence is not the same as safety. Specifically, a limit of equivalence will typically be smaller than a toxicologically relevant effect size, and consequently even a proof of non-equivalence is still saying nothing about safety. However, if the equivalence is established relative to non-GM feeds with a history of safe use, then a proof of equivalence implies a proof of safety, as far as expressed in the investigated variables. Novel statistical approaches for equivalence testing have been developed in G-TwYST as a response to the confusion caused by the traditional approach which focuses on testing the statistical significance of differences. In the traditional difference tests, stating that there is a difference whereas in reality this is not the case is the error of the first kind (and statistical procedures are designed to control the error rate α). In contrast, in safety assessments a failure to find a difference (and thus declare safety) should be considered as the error of the first kind. This leads to the general statistical framework of equivalence testing, which has well-known uses in regulatory practice for medicines (FDA, EMA). The G-TwYST innovations have adapted such methods for the specific field of GM safety testing.

Topic	Subtopic	Comment	Comment made by	Response
Statistics	Equivalence testing	Test of equivalence in plant composition is used as indicator for further studies (e.g. lack of equivalence in compositional studies could indicate the need for feeding studies). However, if the concept of equivalence testing is applied for feeding studies there are no further studies foreseen in the food-feed safety assessment after the feeding studies. What happens if equivalence is rejected for certain endpoints?	**Competent Authority**	First, note that the equivalence tests focus on a null hypothesis of non-equivalence, that we seek to reject in favour of a conclusion of equivalence. Whereas the same results (as shown in the figures) can in principle also be used for a test of non-equivalence (i.e. if the whole interval would lie outside the [-1,+1] range on the ELSD scale), this in fact never occurred in the current results, and such a test of non-equivalence is also not considered as providing useful information. In fact, equivalence was more likely than not for 100% of the tested endpoints. In 6% of the 320 endpoints tested there was insufficient evidence in the data to allow rejection of the null hypothesis of non-equivalence. In the current study no explicit statistical methods for multiplicity correction were applied as these are not yet well developed in the field of equivalence testing. We note that an error rate of 6% is close to the nominal level of 5% that was the basis of the equivalence tests. The specific endpoints where equivalence could not be shown in the statistical test should be inspected jointly in an asssessment by toxicologists. It can be noted that in most of these cases the failure to prove equivalence was for endpoints where the residual standard deviation in the current study was substantially larger than in the available historical data. Therefore the current study may simply not have had enough precision, or alternatively, the precision in the historical dataset may have been accidentically higher than is considered normal. It should be noted that in the current research project G-TwYST, the only available historical data for animals of the same strain and in the same test facility were those from the earlier GRACE project. These data might be non-optimal, e.g. because of the non-routine character of any EU research work, which means that our current use should only be seen as a pilot study and not as a real safety assessment.

Topic	Subtopic	Comment	Comment made by	Response
				If it would be impossible to conclude equivalence and if also toxicologically disturbing effects or patterns of effects would be found, then a decision would need to be made by risk managers on the further process.
Statistics	Equivalence testing	Goedhart & van der Voet (2017) concluded that "Only in three of these cases (Lymphocytes males NK33- and growthRate females NK11- and NK33+) there was both a significant difference and a failure to show equivalence". Please discuss these findings with regard to food-feed safety. Goedhart, Paul W., & von der Voet, Hilko. 2017. G-TxYST Study B a 90-day toxicity study in rats fed GM maize NK603 Statistical report. Biometris report 31.10.17.	**Competent Authority**	It can be noted that for these endpoints (Lymphocytes in males, and growthRate in females) the residual standard deviation in the current study happened to be much higher than in the historical data set (see Figure 17). Therefore the failure to show equivalence is related to the lower precision for these measurements in the current study.
Statistics	Equivalence testing	You suggested that "in cases when a differences is found or an equivalence cannot be shown, the other variables [secondary variables associated with a certain toxicological effect] may provide further interpretation to the toxicologist" (Goedhart & van der Voet, 2017). Although this approach seems to be reasonable, it was not applied for most of the significant differences in the studies. What kind of trigger do you use to carry out a more functional and systemic analysis of results?	**Competent Authority**	Such interpretations are indeed needed, but were made outside of the statistical report by the toxicologist who could also consider the results from histopathology.

Topic	Subtopic	Comment	Comment made by	Response
Statistics	Equivalence testing	In the compositional analysis reference varieties are grown and analyzed simultaneously together with the test organism and the comparator. For equivalence testing historical data from databases are not used any more as they include unjustified high levels of variance. In feeding studies you suggest to use reference studies which will be performed independently from the test studies. Which requirements are set to ensure comparability of the different studies?	**Competent Authority**	From a statistical point of view, the situation in compositional analysis, where a reasonably large number of reference varieties are included in the experiments, is preferable over the situation of animal tests where this is not feasible. However, in principle it should be possible that a routine facility for animal testing would build up a representative reference data set from a series of similar studies with animals of the same strain, without the expectation of unjustified high levels of variance. Having said this, we can also observe that in this animal study the reference between group residual standard error is always smaller than the within group residual standard error (%R/S in Appendix 9). Other than in plant compositional studies, this could perhaps show that different reference (non-GM) feeds have negligible differences in the investigated animal endpoints. If risk managers would decide that this is indeed the case, the approach to equivalence testing can be simplified as was illustrated in our published paper on the method (section 2.7 in van der Voet et al. 2017). This would avoid the need to include many reference varieties in the historical data.
Statistics		It is understandable that the cage is taken as the experimental unit. But as a consequence the values of the individual animals in the cage were averaged. But what if there is a large variation in the values between the individual animals in the cage? Would it in that case be correct to use the averages? Have the researchers checked that there is not a large variation in the values between individual animals in the cage? And that the variation seen in the treated cages are	**Research Organisation**	According to the EFSA Guidance (2011) the cage should be considered as the experimental unit. No statistical analysis of the within-cage variation was performed. As follow-up on this comment, we have additionally tested the equality of within-cage variances for the four GM groups relative to the Control group for study B. Among the 624 tests (78 variables, 4 GM groups, 2 sexes) an F test for equality of within-cage variances was significant in 6.6% of cases, i.e. close to the nominal test level of 5%. Larger or smaller variances in comparison to the Control group were observed approximately equally often for the significant cases.

Topic	Subtopic	Comment	Comment made by	Response
		similar to the variation in the control cages?		
Feed	Feed purity	In the first stakeholder consultation we pointed out that all diet components except from the GM test material should be free from GMO, i.e. the control maize as well as all other diet ingredients (soy meal etc.). Has the purity of feed been analysed?In the first stakeholder consultation we pointed out that all diet components except from the GM test material should be free from GMO, i.e. the control maize as well as all other diet ingredients (soy meal etc.). Has the purity of feed been analysed?	**Competent Authority**	The diets are tested for GMOs. Nevertheless, since the maize has been purchased through the international market traces of other events (typical for Canada) have been found in the harvests. Other ingredients should have been GMO-free (as contracted) but due to the fixed project period and financial limits there has been no timely buffer to pre-test and repeat the production of diets and to postpone any trial. Hence the feed was produced "just in time" as affordable. The results of feed analyses (to be published) show that besides the expected occurrence of NK603 in some samples, there were also traces of other maize varieties present, both in maize kernels and in diets. These additional detects were so minor that they were actually below the limit of quantification, hence their presence can only be described in qualitative terms (demonstrated but not quantifiable). Moreover, this adventitious presence of GM materials is a common phenomenon in animal feeds, which are routinely analysed by the RIKILT lab. As previously argued for GRACE, the trace levels of admixture are unlikely to impact on the outcomes of the feeding studies.
Histo-pathology		It is stated that the death of a rat in treatment group 3 is an incidental finding, unrelated to the experimental treatment. On the basis of what does one come to that conclusion?	**Research Organisation**	1) The neoplasm is a malignant lymphoma. 1) This is one of a small number of spontaneous neoplasms, which are found at very low incidence in Wistar rats on 13 week studies (Toxicologic Pathology 45: 64-75, 2017). 2) It would be very difficult to induce this type of neoplasm in a rat of this age, even with powerful known carcinogens. 3) No more neoplasms of this type were present in the other 13 week study, at the 12 month sacrifice of study A, or among the 24 month sacrificed animals examined to date. Therefore, it is reasonable to consider this tumor an incidental neoplasm.

Topic	Subtopic	Comment	Comment made by	Response
Animal welfare		The G-TwYST project is relevant to our organisations due to involvement of animals used in feeding studies with genetically modified (GM) maize. We would like to focus our comments on the animal welfare issues we identified. Due to the fact, that there were no treatment-related necropsy or histopathological findings following the administration of genetically modified NK603 maize or genetically modified NK603 maize plus Roundup to rats for 90 days in the G-TwYST-project, we call on the EC to act immediately and initiate a research project to develop and/or identify suitable animal-free testing methods to replace feeding trials for testing of GM food and feed to fulfil the requirements of Directive 2010/63/EU. This would be a great chance to show where in silico, in vitro and other animal-free testing methods have a great potential and are superior to animal testing in terms of validity, significance reliability of data. It is the opportunity to send out a strong signal in favour of animal-free testing methods to the scientific and regulatory communities.	**CSO**	The project team is aware of the European Commission policy on more ethical use of animals in testing (3R principles) as well as of the Implementing Regulation (EU) No 503/2013 which mandatorily demands the perfomance of 90-day rodent feeding studies for the risk assessment of GM plants. The EU-funded project GRACE already published conclusions and recommendations on the added value of perfoming rodent feeding trials in the context of the risk assessment of GM plants. The conclusions can be found at the GRACE website: http://www.grace-fp7.eu/sites/default/files/GRACE_Conclusions%20&Recommendations.pdf. Together with the French project GMO90+ and now the EU-funded project G-TwySST a broad data set is provided to elucidate the perfomance and value of such feeding trials including extension of test periods, additional parameters, and advanced data analyses. Only partly the issue of alternatives to animal tests has been touched. A critical issue is to define which information for the risk assessment is actual missing and, in case, what is the appropriate test procedure needed to inform a risk assessment. The projects provide open access to the their data to support discussions and decision making on these issues with a sound scientifc data base and highly appreciate and encourage its use.

CSO: Civil Society Organisation