The Director General

Maisons-Alfort, 27th January 2011

OPINION
# of the French Agency for Food, Environmental and Occupational Health & Safety

Recommendations for carrying out statistical analyses of data from 90-day rat feeding studies in the context of marketing authorisation applications for GM organisms.

CONTEXT OF THE FORMAL INTERNAL REQUEST

The French Food Safety Agency (AFSSA) responded to an internal request to evaluate the current statistical procedure used for the analysis of data from the 90-day rat feeding studies, to achieve the following objectives of:

1) Reviewing the statistical test methods that are or could be used to analyse the 90-day rat feeding studies, emphasising their strengths and weaknesses;

2) Proposing appropriate statistical tools and methodology that fit the objectives of the 90-day rat feeding studies;

3) Investigating the relevance of these tools based on data supplied by an applicant for a marketing authorisation for a given GM organism;

4) Drawing up guidelines for applicants to follow in order to facilitate assessment of the quality of the experimental protocols applied and the results obtained.

BACKGROUND

Risk assessment concerning food and feed derived from genetically modified plants (GMPs) follows the strategy of "substantial equivalence[1]" by comparing molecular and agronomical characteristics and chemical composition with a view to assessing the relative safety of GMPs in comparison to their non-transgenic equivalents[2].

In 2006, EFSA published a Guideline document (EFSA 2006; EFSA 2008) defining studies to be supplied by applicants for marketing authorisations, including 90-day rodent feeding studies, on a case-by-case basis.

More recently, it specified recommendations on the performance of field trials and the analysis of data using appropriate statistical models for comparing the composition of GM plants with that of their conventional counterparts (EFSA 2009).

In 2002, AFSSA identified the sensitive elements of food and feed derived from GMO safety risk assessment; it particularly noted the importance of defining the sample size of populations studied in animal tests, which determines the statistical power (AFSSA 2002).

[1] **Substantial equivalence**: the notion of substantial equivalence expresses the idea that existing organisms used as foods or from which foods are derived can serve as a basis for comparison when assessing the safety of novel or modified foods (OECD 1993; WHO/FAO 2000).

[2] Equivalent or comparator: correspond to a non-GM plant with comparable genetic background.

In its Opinion of 2007, the Agency strongly recommended implementing a 90-day rat feeding study for primary genetic transformation events (AFSSA 2007).

The repeated dose 90-day oral toxicity study in rodents recommended by OECD standard 408 (OECD 1998) was originally designed to evaluate chemicals. It was later applied to the study of "novel foods" prior to their marketing authorisation, as well as to GMOs.
The Agency recently started working on a project to more effectively adapt the protocol for these studies to the particular specificity of GMOs. In the course of this work, it became apparent that it is necessary to focus particularly on aspects related to the statistical analysis of the data, as the available guidelines did not seem sufficiently explicit on this subject.

The purpose of this Opinion is to make recommendations for the statistical analysis of data from the 90-day rat feeding studies when assessing the safety of GM plants. These recommendations could be applied in all situations where this test is used.

ORGANISATION OF THE EXPERT APPRAISAL

The Agency set up a Working Group including statisticians and toxicologists.
The proposed methodology is described in detail in a report published by the Working Group entitled **"Best practice of statistical analysis of data from 90-day rat feeding studies in the context of the marketing authorisation of GMOs"**. This report was validated by the "Biotechnology" Expert Group at a meeting on 16 December 2010.
This Opinion is based on the Working Group's report.

DISCUSSION

We shall deal in turn with 1) a review of the protocols and statistical analyses of published studies relating to the safety assessment of GMPs, 2) the description of available statistical methods, with their strengths and weaknesses and 3) the application of the methodology deemed to be the most relevant to a real case of a particular GMP.

1. Principles of the OECD 408 protocol and current application to GM plants

The experimental protocol currently used is that laid out in the OECD 408 standard, initially designed for assessing the toxicity of chemicals. It recommends populations of at least 10 animals per sex and per group, with 3 doses of the test substance and a control group.
The study includes clinical, haematological, biochemical and histological examinations.

A review[3] of the studies published in the literature for GM plants using this protocol brought out the following points:

- the species used is the rat, most often the "Sprague-Dawley" rat;
- populations of 10 to 20 animals of each sex per group are used;
- a maximum of two doses is used (e.g. for maize: 11% and 33%);

The maximum rate incorporated in the diets depends on the vegetable species, with the maximum dose being limited by the need to maintain a balanced diet.

- the control groups always include rats fed the same "doses" of near-isogenic plants (and sometimes with commercial varieties) as with GMOs;
- the haematological and biochemical examinations (about 50 parameters) are carried out over one or two periods;
- at the end of the study, the organs of all the animals are weighed and gross necropsy and histopathology performed according to the usual principles.

[3] 17 articles published in the scientific literature were analysed.

From a statistical perspective, the OECD 408 protocol does not specify accurate methods. In the publications analysed, the authors use statistical parametric and non-parametric difference tests. They do not consider notions of statistical power or equivalence. Data concerning growth and consumption are not dealt with according to a model suitable for repeated measurements over time.

2. Strengths and weaknesses of the different applicable statistical methods

2.1 Descriptive analysis of the data

- **Identifying atypical data (outliers)**

Before performing any statistical analysis, it is essential to assess the quality of the data observed and to describe their variability.
'Outliers'[4] can have a considerable effect on some statistical parameters, such as the mean or the variance, and bias the results of parametric statistical tests. Such outliers can be identified with appropriate statistical methods (Grubbs or Dixon tests) and plotted graphically to show clearly how they are distributed across the different groups (treatment, sex) or the treatment period (mid time or end of the study). Toxicologists can then decide whether they should be excluded or included with consideration for the plausible cause of the observation of such extreme values. If they are not excluded, then appropriate statistical methods must be used.

- **Transforming the data**

Parametric tests comparing means and assessing confidence intervals for means assume that the variables studied follow normal distributions. In cases where this hypothesis is unrealistic, it is possible to transform the data to approach a normal distribution. Transformation is often necessary when small samples are involved.

2.2 Analysis of the biological parameters excluding rat weight data

2.2.1 Difference tests

- **The hypotheses tested**

A statistical test is a decision support tool for rejecting or validating a hypothesis based on observations. Traditionally, statistical tests known as 'difference testing' have two possible outcomes: rejecting or accepting the null hypothesis which, in the case of 90-day rat feeding studies, is the absence of any effect of a diet based on GMPs compared to a diet based on control plants.
The conclusions of the tests are subject to two types of error. The risk of a Type I error (named α) is the probability of wrongly concluding that there is a difference between animals subjected to the diet containing GM plants and those subjected to the control plant (false positive tests).
The risk of a Type II error (named β) is the probability of not detecting a difference between animals that have been fed a diet containing the GM plant and those fed with the control plant (false negative test). Controlling the risk of Type II error can help reduce the risk of false negative results.
The risk of Type I and Type II errors work in opposing directions. In general, if the risk of Type I error can be fixed (it is most often set at a low value, such as 5%), the risk of Type II error cannot be controlled as it depends in particular on the tested effect size. The statistical power, equal to 1-β, measures the probability of correctly concluding that a given effect actually exists.
In order to reduce the risk of Type II error (β), ANSES therefore suggests increasing the risk of Type I error (α) (to 10% instead of 5%) which will lead toxicologists to examine a higher number of statistically significant differences. This means a higher level of vigilance, thus reducing the probability of not detecting a real difference (reducing the number of "false negatives").

- **Parametric and non-parametric tests**

Difference tests can be conducted with parametric methods when the structure of the data allows it (symmetrical or Gaussian distribution, constant variance). Data can sometimes be transformed to approach a Gaussian distribution. Generally, parametric tests are more powerful and make it possible to fit the analysis to the experimental protocol of the study (for example, by taking into

[4] An *outlier* is a value that stands apart from all the other observations in the sample. They can be the result of the variability inherent in the biological criterion measured, a measurement error or an execution error.

account repeated data about a given animal or by testing any interactions between the different variables of the study). When the conditions for applying parametric tests are not satisfied even after data transformation, resorting to non-parametric tests can be worthwhile.

Parametric and non-parametric difference tests are designed to reveal the differences between animals having undergone different experimental treatments, but cannot provide proof that there is no difference.

- **Consequence of the multiplicity of tests on Type I risk of error**

A large number of statistical tests on a given dataset rapidly increase the Type I risk of error, i.e. the probability of observing falsely significant differences. Corrections such as the False Discovery Rate (FDR) proposed by Benjamini (Benjamini and Hochberg 1995) can be used to estimate the proportion of false positives in the results.

- **Consequence of a lack of power**

The absence of significant test differences could be the result of a lack of power and cannot be interpreted as the absence of effect of the treatment. Low power can be the result of the variability of the measurements or of low sample size. The power can also depend on the value fixed for Type I (α) risk of error.
Ideally, power analysis is carried out *a priori* as a function of the desired effect size from which the sample size leading to a sufficient power can be determined.
However, statistical power analysis can also be done after the study has been carried out (*a posteriori*), to help interpret the absence of significant effect. For this, the minimum detectable effect size, when comparing groups of animals fed on a diet containing a GMP with those fed on a diet containing the control plant, is calculated for the given power (usually 80%).

2.2.2 Non-difference or equivalence test

EFSA recently recommended using equivalence tests for comparative analyses of the chemical composition of GM plants and controls, cultivated simultaneously in field trials under different environmental conditions (EFSA 2009). It also suggested using them to assess environmental risk and more specifically the potential impact of GM organisms on non-targeted organisms (EFSA 2010).

It is difficult to use equivalence tests for analysing data from toxicological tests on laboratory animals as the results can only be interpreted if a large number of animals per group and per sex are observed.

2.3 Analysis of weight data

If rats are weighed repeatedly over time, mixed models must be used that are suitable for repeated data from the same animal. Various software programs are available for analysing growth curves.

2.4 Conclusions for this analysis

Considering the different points examined, and with the aim of optimising the interpretation of these tests, ANSES suggests increasing the Type I (α) risk of errors, while also using sufficiently large animal populations in experimental studies to lead to a satisfactory statistical power. By increasing the power of the tests, the number of statistically significant differences, potentially indicating risks to health, would be increased. The risk of returning false negatives is reduced.
Section 3, below, tests this methodology with a real dataset in order to verify its operational feasibility and estimate the minimum size of animal populations necessary to attain sufficient statistical power.

3- Application of the proposed statistical methods to the data from a study supplied with an application for marketing authorisation of a GMO

The difference tests were applied to the data from a study carried out with genetically modified maize (MON810), supplied by the applicant to support a request for authorisation to market this maize in the European Union. The results show the importance of an in-depth descriptive analysis

to identify potential outliers in the data and verify the conditions under which the parametric tests are applied.
In this study, most of the experimental data were transformed in order to approximate a Gaussian distribution.
To boost the vigilance of toxicologists, and also to increase the statistical power of the test, the Type I (α) risk of errors was set at 10% instead of the usual 5%.
Under these conditions, the results of parametric tests revealed 33 significant differences out of 432 'GMO diet versus near-isogenic diet' comparisons. After controlling for FDR, no differences were deemed significant. The weight data were processed by a Mitscherlich-type nonlinear mixed model that took data repetition into account, and no significant difference was found.

For each parameter, calculations of statistical power were carried out on the basis of a detectable effect size, equivalent to the value of one standard deviation calculated from data from the groups fed with commercial varieties, or from historical data taken from the literature. This is a new approach relative to current practice. On this basis, the statistical power of the difference tests appears insufficient for some parameters, leaving a risk of wrongly returning an absence of difference.
Since statistics are a decision support tool, toxicologists should examine any significant difference or any statistical power deemed insufficient by 1) considering the parameters individually, 2) analysing them according to current requirements of converging evidence and 3) taking into account histological data which are excluded from these statistical analyses. This approach led the toxicologists to conclude that there were no differences of toxicological significance between animals fed on the GM plant being studied and the control animals fed on near-isogenic plants.
Furthermore, complementary calculations showed that the values of statistical power reach acceptable levels (higher than 80%) when the number of animals per group and per sex is increased from 10 to 20. This increase in the sample size seems particularly appropriate considering that some studies already use groups of 20 animals per sex, and that some parameters are already measured on these 20 animals. Under these conditions, the detectable effect size is at least equal to one standard deviation of the control data.
Taking into account the evaluation of the statistical power, is an important recommendation of this Opinion. Increasing statistical power facilitates and reinforces the conclusions of the toxicologists.

On the basis of the previously-defined tolerance thresholds and for this dataset, implementing the statistical equivalence tests would require large populations (several hundred animals per group, i.e. at least 10 times the number recommended in the OECD protocol) in order to allow conclusions with sufficient statistical power (80%).
This conclusion, which raises ethical issues concerning animal welfare as against the relevance of the expected information, leads to further consideration of the applicability of the statistical equivalence test to the 90-day rat feeding studies.

CONCLUSION AND RECOMMENDATIONS

ANSES has already recommended, in the context of the safety assessment of a new genetically modified plant (a new genetic transformation event), that a 90-day rat feeding study be implemented.
This 90-day rat feeding study is based on the OECD 408 protocol. In the light of the risk assessment carried out by the Agency, it seems necessary to improve this protocol, particularly concerning 'statistical data analysis', which is an important part of the risk assessment. The aim of the recommendations listed below, based on an increase in the statistical power of the tests, is to propose a highly rigorous methodology for statistical data analysis.

As the standards for the risk assessment of GM organisms are set by the European Food Safety Authority (EFSA), the recommendations listed below, which aim to improve the statistical analysis of data, may be considered as a scientific contribution by ANSES to EFSA:

- Identify and analyse outliers, prior to statistical processing, using descriptive methods. These data must be closely analysed by toxicologists.

- Transform asymmetric distributions to approximate Gaussian distributions, especially when sample sizes are small.

- Use statistical models adapted to the experimental design and the measurements carried out by using:
  - mixed models (linear or non-linear) if measurements are repeated for the same individuals on different dates;
  - Gaussian models if the data follow symmetric distributions (after data transformation, if necessary);
  - non-parametric methods if the data follow asymmetric distributions.

- Accept a 10% Type I risk of errors (greater than the 5% usual level). The expected result will be an increase in the number of statistically significant differences; these will boost the vigilance of toxicologists for a greater number of parameters, whose relevance they will have to assess.

- Evaluate the consequences of the multiplicity of statistical difference tests and calculate the probability of obtaining false positives (FDR) in the result.

- Evaluate the statistical power of the difference tests by calculating for each biological parameter, the minimal detectable effect size between treatments leading to a statistically significant difference with a probability of 0.8 (i.e. a statistical power of 80%). The toxicologists will judge whether the detectable differences for a statistical power of 80% have toxicological relevance on the basis of their own expertise, and, when available, of historic data or data internal to the study (groups fed with commercial varieties).
  To achieve a statistical power of 80% for almost all parameters, with conservative tolerance thresholds equal to one standard deviation computed from data on groups fed with commercial varieties, it is recommended that 20 animals be used per group (one treatment, one sex, one dose). Using other thresholds could lead to different sample sizes.

- The conclusions of studies using the term ‘equivalence between the two diets’ must be justified by using equivalence test techniques. Further reflection is required on the applicability of this test to the 90-day rat feeding studies data.

- Facilitate the interpretation of results by illustrating them graphically, particularly concerning the effect sizes.

- Make the raw data available in electronic form to allow experts to carry out any further verification or analysis they may deem necessary.

Adopting these recommendations should lead to a clearer presentation of the results of the 90-day rat feeding studies, particularly by identifying uncertainties. Toxicologists will thus be able to interpret the results more rapidly and with greater objectivity and reliability.

Director General
Marc MORTUREUX

KEYWORDS

GMO, the 90-day rat feeding study, statistical analysis, OECD 408.

BIBLIOGRAPHY

AFSSA (2002) Evaluation des risques relatifs à la consommation de produits alimentaires composés ou issus d'organismes génétiquement modifiés (Risk assessment for the consumption of foods composed of or derived from GMOs).

AFSSA (2007) Avis relatif aux études de toxicité réalisées dans le cadre des demandes de mises sur le marché d'OGM (Opinion on toxicity studies carried out in the context of applications for marketing authorisations for GMOs).

Benjamini Y, Hochberg D (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289-300.

EFSA (2006) Guidance document of the Scientific Panel on Genetically Modified Organisms for the risk assessment of genetically modified plants and derived food and feed. *The EFSA Journal* **99**, 1-100.

EFSA (2008) **DRAFT** Updated Guidance document for the risk assessment of genetically modified plants and derived food and feed *The EFSA Journal* **727**, 1-135.

EFSA (2009) Scientific opinion on statistical considerations for the safety evaluation of GMOs, on request of EFSA. *The EFSA Journal* **1250**, 1-62.

EFSA (2010) Scientific opinion on the assessment of potential impacts of genetically modified plants on non-target organisms *The EFSA Journal* **8**, 1877.

OECD (1993) Safety evaluation of food derived by modern biotechnology : concept and principles. *Paris, France*.

OECD (1998) Guideline for the Testing of Chemicals : Repeated Dose 90-day Oral Toxicity Study in Rodents. *OECD*.

WHO/FAO (2000) Safety aspects of genetically modified foods of plant origin. Report of a joint FAO/WHO expert consultation on foods derived from Biotechnology. *WHO/FAO* **2000**(29).