DRAFT FOR PUBLIC CONSULTATION

SCIENTIFIC OPINION

EFSA guidance on repeated-dose 90-day oral toxicity study on whole food/feed in rodents[1]

EFSA Scientific Committee[2, 3]

European Food Safety Authority (EFSA), Parma, Italy

SUMMARY

Following a request from the European Commission the Scientific Committee was asked to develop principles and guidance for the establishment of protocols for 90-day feeding studies in rodents with whole food and feeds. The design of such protocols should be based on the specific properties of food/feed derived from genetically modified plants and other novel food under investigation and in line with the purpose of the study. In view of the multidisciplinary nature of this subject, the task was assigned to the Scientific Committee.

Risk assessment of food and feed comprises an integrated approach where information is required on a number of characteristics from various types of tests, including toxicity. Data generated from toxicity testing, whether collected from in vivo or in vitro studies provide fundamental information for carrying out a risk assessment of a food for human consumption, or of a feed for animals.

In specific cases, toxicity testing of the whole food/feed may be considered, depending on the type of the food/feed under investigation, its history of (safe) use, and the available toxicological information on the whole food/feed and its constituents. This guidance further develops the general procedure set out in the OECD Guideline for the Testing of Chemicals – Repeated Dose 90-day Oral Toxicity Study in Rodents (OECD TG 408), and provides specific advice for performing and reporting experiments carried out with whole food/feed.

Appropriate characterization of the whole food/feed to be tested is required and should include among others a description of the source, its composition, the manufacturing process, information on stability and the presence of chemical and/or microbiological contaminants.

[1] On request from the European Commission, Question No EFSA-Q-2009-00941, endorsed for public consultation on 22 June 2011.

[2] Scientific Committee members: Boris Antunović, Susan Barlow, Andrew Chesson, Albert Flynn, Anthony Hardy, Klaus-Dieter Jany, Michael-John Jeger, Ada Knaap, Harry Kuiper, John-Christian Larsen, David Lovell, Birgit Noerrung, Josef Schlatter, Vittorio Silano, Frans Smulders and Philippe Vannier. Correspondence: scientific.committee@efsa.europa.eu

[3] Acknowledgement: The Scientific Committee wishes to thank the members of the Working Group on 90-day feeding trials on whole food/feed: Gerrit Alink, Gianfranco Brambilla, Noel Albert Dierick, Michael Festing, Gerhard Flachowsky, Gijs Kleter, Joel Guillemain, Harry Kuiper, Morten Poulsen, Vittorio Silano and Hans Verhagen for the preparatory work on this scientific opinion and hearing expert Joe Perry and EFSA staff Saghir Bashir, David Carlander and Ellen van Haver for the support provided to this scientific opinion.

Suggested citation: EFSA Scientific Committee; 2011. Draft for public consultation – Scientific Opinion EFSA guidance on repeated-dose 90-day oral toxicity study on whole food/feed in rodents. [45 pp.] Available online: www.efsa.europa.eu

Preparation of appropriate test diets is a key element of the experiment with respect to the choice of the diet type, nutritional balance and necessary adjustments, processing, and storage. Since it is often not possible to include whole foods in an amount that will induce toxicity and thus to obtain a dose-response relationship, the use of two dose levels is recommended to maximise the power. The highest dose level of the whole food/feed that can be incorporated in the animal diet should not cause nutritional imbalance or metabolic disturbances in the test animal, and the lowest dose level should always be above the anticipated human/target animal intake level.

For ethical and scientific reasons the test animals should be housed two (of the same sex) per cage, which is the experimental unit. A randomised block design is suggested with the animals within a block being matched for age and weight (for each sex) and location within the animal house. A randomised block design helps to reduce variation.

Two examples of randomised block designs are provided which use 96 animals. For novel foods this corresponds to three treatment groups (low and high dose and control) in 8 blocks, and for GMO four treatment groups (low and high dose GMO and low and high dose isogenic comparator) in 6 blocks. Further increase in power of the experiment, when considered relevant, could be achieved by adding extra blocks to the experiment.

Due to the fact that a number of the variables (i.e. effect size, variability, significance level, power and the alternative hypothesis) will have to be estimated or assumed, the number of animals (sample size) will vary according to the choices and justifications made. The applicant should describe and justify the calculation of sample size and the values of the variables used in the protocol. In addition, the design of the experiment should be clearly described including whether it is a “completely randomised design ” or a “randomised block design” and the experimental unit should be specified (e.g. number of animals/cage).

It is emphasized that the biological relevance of observed differences should be assessed, even if some fail to reach the chosen level of statistical significance. This assessment should involve the use of point and interval (e.g. confidence) estimates in addition to the significance level.

The inclusion of reference groups, fed with a diet containing commercially available food/feed similar to the test food/feed, in the experimental design, in order to estimate the natural variability of test parameters is in general not recommended, since this would substantially increase the number of test animals. Historical control data on natural variations in values of test parameters should primarily be obtained from databases available in the actual testing facility, while data from literature might also be informative. Inclusion of reference groups may be considered if no acceptable historical background data is available.

A comprehensive set of end-points should be measured at the end of the 90-day period. An interim collection of data from blood samples should normally be taken after 45 days.

The study report of the experiment should include descriptive statistics and be presented in such a way as to facilitate interpretation. Graphical methods, particularly the presentation of means with confidence intervals, should be used. Consideration should be given to expressing results in terms of standardised effect sizes.

KEY WORDS

Repeated-dose toxicity study, 90-day, rodent, food, feed, testing, protocol, safety, risk assessment

69

70 **TABLE OF CONTENTS**

BACKGROUND AS PROVIDED BY THE EUROPEAN COMMISSION

Evaluation of the safety and nutritional properties of whole genetically modified (GM) and other novel foods/feeds is an important feature in the safety/nutritional assessment of these foods/feeds (Regulation (EC) No 1829/2003 on GM food/feed and feed and Regulation (EC) No 258/97 on Novel Foods under revision).

Commonly the safety assessment of these foods/feed comprises an extensive compositional analysis, an in-vitro/in-silico characterization and assessment of results obtained from animal tests with relevant purified compounds identified in them, like for instance newly expressed proteins or other constituents, rather than the toxicological/nutritional testing of the whole food/feeds themselves. In specific cases toxicity testing of the whole food/feed may be considered, depending on the type of the food/feed under investigation, its history of (safe) use, the available toxicological information, or remaining uncertainties. As of today, no standardised protocol or guidelines exist for this type of study and applicants are advised to adapt the OECD Test Guideline 408 (90-day oral toxicity study in rodents) designed for toxicity testing of single defined substances.

In March 2008, a report of the EFSA GMO Panel Working Group on animal feeding trials entitled "Safety and nutritional assessment of GM plants and derived food and feed: The role of animal feeding trials" was published . This publication treats this issue in more detail and recommends the development of supplementary guidelines for this type of study.

In order to provide rapidly guidance to applicants on this matter, it is appropriate that EFSA develops guidance for applicants on this matter. This work could also contribute to the establishment of such guidance at the international level.

TERMS OF REFERENCE AS PROVIDED BY THE EUROPEAN COMMISSION

EFSA is requested according to Article 29 of Regulation (EC) No 178/2002 to develop principles and guidance for the establishment of protocols for 90-day feeding studies in rodents with whole food and feeds. The design of such protocols should be based on the specific properties of the GM and other novel food/feed under investigation and in line with the purpose of the study. Specific attention will be paid to the development of protocols suitable for food/feed derived from GM plants.

Guidance should include among others considerations on:

- Study purpose and design
- Type of test, control and reference diets, analysis and storage
- Criteria for balancing diets,
- Types of test, control and reference groups,
- Dosage regimes and spiking,
- Toxicological and nutritional endpoints to be measured.
- Data collection, statistical analysis
- Quality assurance aspects

GUIDANCE

1. Introduction

Risk assessment of food and feed comprises an integrated approach where information is required on a number of characteristics from various types of tests, including toxicity. The data and information generated from toxicity testing, whether collected from in vivo or in vitro studies provide information for carrying out a risk assessment of a food for human consumption.

The OECD Guideline for the Testing of Chemicals – Repeated Dose 90-day Oral Toxicity Study in Rodents (OECD TG 408) provides information on possible hazards due to repeated exposure to chemicals over a prolonged period of time (90-days) covering post-weaning maturation and growth into adulthood (OECD, 1998). The OECD TG 408 is designed to provide information on toxic effects on the animals, to indicate target organs, and to establish a no-observed-adverse-effect level of exposure and to establish safety criteria for human exposure. Compared with the original guideline from 1981, the current version of the OECD TG 408 places additional emphasis on neurological endpoints and provides indication on immunological and reproductive effects.

Recently the French Agency for Food, Environmental and Occupational Health and Safety (ANSES) published an opinion with recommendations for carrying out statistical analyses of data from 90-day rat feeding studies in the context of marketing authorisation applications for GM organisms (ANSES, 2011). The ANSES opinion, based on using the data and study design of the MON810 study, was provided as a contribution to EFSA during the development of the current guidance.

Current experiences of EFSA from assessing repeated-dose 90-day oral toxicity studies indicate that a number of differences exists among the considered studies e.g. in experimental designs, test diets and dosage regimes, biological endpoints and statistical approaches.

Application of the OECD TG 408 for testing whole food/feed encounters a number of challenges. While single chemicals and simple chemical mixtures can be administered to the test animal at dose levels which are several times higher than the likely human exposure levels, this may not be possible with whole food or feed as these are bulky and can result in satiation and/or unbalanced diets if given at high levels. Therefore, careful consideration needs to be given to ways in which the design and analysis could be adjusted in order to increase the chance of detecting any toxic effects.

This guidance further develops the general procedure set out in OECD TG 408 and provides specific advice for performing and reporting experiments carried out with whole food/feed. The main modifications compared with OECD TG 408 are related to the preparation of the test diet (section 2), the housing of animals (section 4) and the experimental design and statistical methods (section 5) which accordingly are extensively discussed in the guidance. Endpoints to be measured are indicated in section 4 and section 6 discusses interpretation of the results. Section 7 gives information related to the uncertainty analysis. Finally, section 8 describes what should be reported in the study report, including the protocol used and the statistical analysis plan.

1.1. Objective of this guidance

This guidance aims to aid applicants in designing, conducting, analysing, reporting and interpreting repeated-dose 90-day oral toxicity studies of whole food/feed in rodents for the purpose of risk assessment. The guidance offers advice on key principles to minimise bias and maximise the precision to draw valid conclusions from the experiment. The guidance also provides additional information for the statistical analysis and aims to harmonise reporting of the results (the study report).

1.2. The term whole food/feed as used in this guidance

In the context of this guidance, whole food/feed refers to a product, intended to be ingested, which in general is composed of a multitude (up to thousands) of individual substances. Whole food/feed range from plant based products such as maize or potatoes to more refined products such as fruit juices or flour, to foods composed of or derived from microorganisms as well as animal-derived food products such as meat and milk.

The interpretation of the whole food/feed term as used in this guidance (sometimes also referred to as "whole product") aims to differentiate a whole food/feed from more purified single food/feed ingredients, consisting of one or few substances that in the context of animal testing could be administered at higher dietary levels.

It is expected that within the European regulatory context the guidance is specifically focused on testing whole food/feed derived from genetically modified organisms (GMOs) and those that fall under the novel food regulation (currently Regulation (EC) No 258/1997). The guidance would also be suitable to test e.g. whole food/feed products derived from animal cloning or GM animals.

1.3. Risk assessment strategies and animal feeding trials with whole food/feed

The risk assessment strategy for different types of whole food/feed requires specific information on the characteristics and properties of the food in question, e.g. information on the source material, production method and processing, on the composition and presence of contaminants, and the toxicological and nutritional properties. The information is generated from specific tests with food constituents. Repeated-dose 90-day oral toxicity studies with the whole food in rodents may be performed on a case-by-case basis to provide additional information for the risk assessment.

1.3.1. Food/feed derived from genetically modified organisms

Products under consideration include whole food/feed derived from GM plants with various input traits to introduce e.g. herbicide tolerance and/or insect resistance (including stacking of such events) and traits leading to improved responses to environmental stress conditions, or to improved nutritional/health characteristics (see further Table 1 of the EFSA Report on Animal Feeding Trials, 2008). Typical GM crops are maize, soybeans, oilseed rape and cotton. This category also includes genetically modified microorganisms and their products.

Furthermore, products under consideration may be derived from GM animals whose genetic material has been altered in a heritable way either through recombinant DNA or other in vitro nuclear techniques. Applications may include genetic modification of husbandry animals, fish, as well as crustaceans and molluscs, insects (for instance honey bees) and other invertebrates. Inserted traits can be related to more efficient or increased production of food, enhanced nutritional characteristics and wholesomeness of these foods, lower emissions to the environment or improvement of the health characteristics of the GM animal, including better resistance to abiotic stressors and pathogens, improved fertility and lower mortality.

Under certain conditions, 90-days toxicity studies in rodents with the whole food derived from the GMO may be considered. The purpose of such studies is to reassure that the GM food/feed is as safe and nutritious as its traditional comparator, rather than determining qualitative and quantitative intrinsic toxicity of defined food/feed constituents (EFSA, 2008; EFSA Panel on Genetically Modified Organisms (GMO), 2011).

1.3.2. Novel foods

Products under consideration are whole novel foods or food ingredients falling under Regulation (EC) No 258/1997. This may be the case for products with a new or intentionally modified primary molecular structure; whole novel foods or food ingredients consisting of or isolated from microorganisms, fungi or algae; whole novel foods or food ingredients consisting of or isolated from plants and food ingredients isolated from animals, whole novel foods or food ingredients to which has been applied a production process not currently used, where that process gives rise to significant changes in the composition or structure of the foods or food ingredients which affect their nutritional value, metabolism or level of undesirable substances. Examples of already authorised novel foods are noni juice, neptune krill oil, Salvia hispanica seeds and ice-structuring protein, salatrims and enova oil. The full list of currently authorised novel foods in EU is found at: http://ec.europa.eu/food/food/biotechnology/novelfood/authorisations_en.htm.

Safety assessment of a large number of novel foods for which a traditional counterpart exists, has been based on the acceptance of substantial equivalence of the novel food with already existing foods in terms of their composition, nutritional value, metabolism, use modalities, nature and levels of undesirable substances. Data from repeated-dose 90-day toxicity tests may be generated on a case-by-case basis.

1.3.3. Other types of food/feed that may be considered

In addition to the two above-mentioned main categories, there may be other types of whole products that could be considered to be tested according to this guidance. Examples of other products under consideration could be meat and milk products from animal clones or from offspring of animal clones (EFSA, 2008) or foods modified by nanotechnology.

2. Characterization of the whole food to be tested

Appropriate characterization of the food or feed to be tested is an integral part of the risk assessment. Critical elements for the characterisation of food/feed have been described in various documents (Regulation (EC) No 258/97; Verhagen et al., 2003; Agget et al., 2005; EFSA Panel on Genetically Modified Organisms (GMO), 2011). The following are examples of elements that should be addressed to obtain a complete analytical composition of the whole food/feed: name, source and specifications, composition, manufacturing processes, batch to batch variations, information on stability etc (for specific details, see EFSA guidance documents for the intended use).

2.1. Preparation of the diet for the testing

The performance of laboratory animal studies of whole food/feed meets a number of challenges since whole products, as covered in this guidance, are complex mixtures of compounds with very different biological characteristics. Food/feed are bulky and may have an effect on the satiety of animals and can therefore only be fed at relatively low multiples compared to their typical presence in the human/target animal diet. Moreover, there is a possibility that in attempting to maximise the dietary content of the whole product under investigation, nutritional imbalances may occur. These could lead to the appearance of effects which may not be related to the properties of the whole product being tested.

For whole food/feed where no adequate information exists on previous testing it could be necessary to perform a small preliminary tolerance test with a limited number of animals and with a short duration (1-2 weeks). The purpose of such pilot studies is to investigate whether the feed intake of the animals is appropriate, to get indications of the dose levels to be used in the 90-day study and to observe if any side effect occurs.

2.1.1. Formulation of nutritionally balanced test diets, matrix dependency

Before preparing the animal diet, it is necessary to have a complete analytical picture of the composition of the whole food/feed and, if available, the comparator. The whole product analysed should be a sample of that which is incorporated in the diet of the test animal. The compositional analyses should include macro- and micronutrients, other food/feed constituents, and chemical and microbiological contaminants.

It should be considered if the whole product contains inherent anti-nutritional components or minerals in a relatively high concentration (e.g. trypsin inhibitor in unprocessed soybean meal or glycol-alkaloids in potatoes). A high incorporation level of such a whole food/feed in the diet of the test animals can result in a nutritional or even toxic effect. These effects can be predicted from the compositional analysis, review of literature or preliminary studies and should be taken into account in the test diet formulation. The presence of anti-nutritional components, or other substances, in the whole products to be tested may be the limiting factor for determining its maximum inclusion level into the test diet.

Adjustments of the contents of nutritionally important ingredients should be considered if significant compositional differences exist between the whole food/feed and a potential comparator at the compositional analyses. If a natural comparator does not exist, the anticipated level of nutritionally important ingredients in the diets of the control and dose groups should be examined. If the levels of nutritionally important ingredients in the diets differ by more than 5 % between groups it is recommended to adjust the diets (FDA, 2000; Knudsen and Poulsen, 2007; Poulsen et al., 2007).

2.2. Choice of diet

There are several types of diets to which the whole product to be tested could be incorporated to form the animal test diet. The most common diets in animal studies are the following:

- Diets based on natural ingredients, mainly agricultural ingredients and by-products
- Purified diets (formerly known as semi-synthetic diets)
- Synthetic diets which are chemically designed
- Human-type diets

Natural-ingredient diets are formulated with agricultural ingredients like cereals, maize, soy etc. They are acceptable and palatable to most animal species. They include the commercially produced standard laboratory animal diets, known as chow diets, which often have been used for rodent feeding studies testing chemicals.

Purified diets are formulated with a more refined and restricted number of ingredients than the natural-ingredient diet. The ingredients are well-characterised and may include maize starch, soy oil, sucrose, casein, cellulose etc. Purified diets are most often preferred when whole foods and macro-ingredients are tested because it is easy to alter ingredients in this type of diet. It is therefore, in most cases, possible to achieve higher incorporation level of the whole product to be tested than in the natural-ingredient diet.

Synthetic diets are made from simple, elemental ingredients like amino acids and specific fatty acids and are used to test single chemically defined substances like a specific micronutrient or amino acid. The synthetic diet is expensive and rarely used.

A human-type diet should represent a balanced human meal and at the same time fulfil the nutritional requirements of the experimental animal. This type of diet is not used very often due to the lack of background experience and its complex nature.

2.2.1. Dosage regimes: routes of administration, dose range, levels and frequency

The whole product to be tested should preferably be incorporated in the diet and fed ad libitum. This will give the most optimal and relevant physiological intake scenario. In the case when the whole food/feed is given beside the diet or by gavage, the same kind of consideration about balancing the diet should be taken as in the case when the whole food is incorporated into the diet. Administration by gavage is not common for whole products but could be considered in certain instances due to poor palatability or stability, or in cases where an exact dosing is needed.

According to the OECD TG 408 at least three dose levels and a control should be used. Furthermore, it is stated that unless limited by the biological nature of the test substance, the highest dose level should induce toxicity but not death or severe suffering in test animals. A no-observed-adverse-effect level (NOAEL) should be observed at the lowest dose level. However, when testing whole food/feed this may not always be relevant since it is often not possible to include whole foods/feed in the test diet in an amount that will induce toxicity without causing nutritional imbalance or metabolic disturbance. Therefore the use of only two dose levels, high and low is recommended (see also section 5.2.1).

The highest dose level should correspond to the highest level of the whole product that can be incorporated in the animal diets without causing nutritional imbalance or metabolic disturbance (NRC, 1995). The lowest dose level should always be above the anticipated human intake, as otherwise the data obtained will be of no relevance in the assessment. When using this strategy, the recommended OECD TG 408 use of a two to four fold interval between the doses may not be optimal.

When high incorporation levels in the diet are used, it should be verified that they do not lead to nutritional imbalances. Nutritional differences above 5 % should be adjusted for in the total diet (see section 2.1). It should be scientifically justified why a higher incorporation level is not feasible. The highest level that can be used may be impacted by processing and should be assessed on a case-by-case basis. However, when untreated products are tested, the presence of anti-nutrient factors should be considered (e.g. trypsin inhibitors in case of soybeans).

When testing complex novel products of protein origin or with a high content of protein it is frequently the protein per se that is the limiting factor in the attempt to get as high an incorporation level as possible. For novel fats or products containing high levels of fats it is correspondingly the fat that is the limiting factor that in excess can cause an unbalanced diet. When testing meat based products, consideration should be given that rodents, albeit omnivores, are not adapted to a full meat based diet.

2.2.2. Processing of the test diet

The whole food or feed to be incorporated in the animal diet should be as similar as possible to the product that is to be consumed. Therefore, in some instances (e.g. for rice, potato, legumes etc.) a heat treatment (e.g. cooking) of the whole products may be necessary. Similarly, for feed, the use of pilot-processing to obtain the by-products that are marketed may be required, e.g. seed meal remaining after seed oil extraction. It should be considered that processing may lead to the formation of toxic compounds, like for instance acrylamide and Maillard reaction products, or may result in the destruction of anti-nutrient factors such as alpha-amylase inhibitors. The impact of whole food/feed processing on human/animal health could be assessed by testing an animal test diet with processed and unprocessed whole product.

In all cases, the use of heat treatment during the manufacturing of the diets and its impact has to be considered, (e.g. influence of steam pelleting and autoclaving). At the same time, it should be ensured that these changes in practice do not impact on the safety and quality of the product.

2.2.3. Analysis of biological and chemical contaminants in the test diet

The potential occurrence of biological, chemical and microbial contaminants in the test diet, should be controlled and results discussed and reported. Acceptable levels in rodent diets have been issued by different national bodies, framed within Good Laboratory Practices (GLP) guidelines (Clarke et al., 1977; Rao and Knapka, 1987; Stevens and Russel, 2007; Directive 2010/63/EU).

2.2.4. Storage of the test diet

Good manufacturing techniques and appropriate environmental storage conditions will minimize spoilage and degradation of the test diet. Guidance how to store feeding stuffs, preventing nutrients degradation and mould and insect colonisation and growth, are implemented under Hazard Analysis Critical Control Points schemes (e.g. TASCC, 2010).

3. Endpoints to be measured

The 90-days study in rodents should be conducted with the full range of observations as described in the OECD TG 408. Measured endpoints in the OECD 408 TG, in addition to general clinical observations, include e.g. food/feed and water intake, growth, haematology, blood clinical biochemistry, urinalysis, gross necropsy and histopathology.

In addition to the OECD TG 408 observations, additional parameters described in the more recent guideline on repeated-dose 28-day oral toxicity study in rodents (OECD test guideline 407) should be assessed. The additional parameters place more emphasis on endocrine-related endpoints (e.g. determination of thyroid hormones, gross necropsy and histopathology of tissues that are indicators of endocrine-related effects, and (as an option) assessment of oestrous cycles).

Other parameters could also be considered if there are indications that the whole product may have effects on e.g. the cardiovascular, nervous or immune system. If the whole product has been designed to have e.g. an impact on the gut microbial flora, this should be investigated.

Furthermore to what is recommended in the OECD TG 408, an interim collection of data from blood samples should normally take place after 45 days.

The endpoints should be reported for all animals, except for histopathology which initially should be performed on the control and high dose group. If histopathological differences are observed in the animals from the high dose group, those from the low dose group, and the isogenic dose group (when available) should also be examined.

The protocol should clearly specify all the endpoints to be measured and the times at which they have to be measured. The use of other methods and/or inclusion of additional endpoints should be justified. The results from the 90-day experiment may trigger the need to perform additional studies (see also section 7.1).

4. Animals for use in 90-day toxicity studies

The general principles for using laboratory animals should be adhered to. All studies should be carried out following OECD Good Laboratory Practice (GLP) guidelines (OECD, 1998) and taking account of animal welfare as outlined by the EFSA Scientific Panel on Animal Health and Welfare (AHAW) related to the aspects of the biology and welfare of animals used for experimental and other scientific purposes (EFSA, 2005) and of the EFSA Scientific Committee on existing approaches incorporating replacement, reduction and refinement of animal testing: applicability in food and feed risk assessment (EFSA, 2009a). All procedures should be approved by an ethics committee taking account of the "3Rs" (Replacement, Refinement and Reduction) (Russell and Burch, 1959).

Animals used in 90-day toxicity experiments should be healthy and free of the major pathogens. They should come from breeding colonies maintained to internationally recognised standards such as AAALAC (http://www.aaalac.org/) accreditation or its equivalent, with a routine health monitoring system which screens for pathogenic bacteria, viruses and parasites. The list of pathogens tested in the screening should be included in the study report, with an indication of those which were present and absent in the breeding colony from which the animals were obtained.

Weaning age animals should be acclimatised for a period of 5-15 days. Dosing should begin as soon as possible after acclimatisation and, in any case, before the animals are nine weeks old. At the commencement of the study the weight variation of animals should be minimal and not exceed ± 20 % of the mean weight of each sex.

4.1. Housing and maintenance

Rats and mice are social animals and housing them singly causes stress (Westenbroek et al. 2003; Leshem and Sherman, 2006). Stress sometimes leads to an increase in variability. For example, mice housed singly had a mean body weight of 46±5.8 g compared with those housed two per cage of vs 44.7±3.9 g (Chvedoff et al., 1980). This extra variability would translate into needing twice as many animals (30 vs 14) to detect a 5 g change in body weight using a two-sample t-test assuming a 90 % power and a 5 % significance level.

It is common practice to house animals individually when performing whole food/feed studies. However, to reduce stress and inter-individual variability, it is recommended, both for welfare and scientific reasons, that animals should normally be housed as pairs in a solid-bottomed cage unless a different system is scientifically justified. In mice it has been observed that housing two animals of the same sex (especially males) together may at sexual maturity cause aggressiveness. To minimise this risk, it is suggested that non-aggressive mice strains are selected and paired when they are received by the test facility. Aggressiveness from housing two rats per cage of the same sex is not observed, however, it is also suggested that pairing should take place when they are received. The aspect of aggressiveness, should be monitored during the experiment and reported at the end.

Housing animals in pairs has statistical implications. In a controlled experiment animals are assigned to the treatments at random, and it must be possible for any two animals (termed the experimental units, ExpU) to receive different treatments. Animals in the same cage cannot receive different treatments when these are supplied in the diet. However, cages can be independently assigned to treatments, so these are the ExpU.

4.2. Choice of stocks or strains of animals

There are two major classes of laboratory mice and rats used in research and testing: outbred stocks and isogenic strains (inbred and F1 hybrid) . Outbred “genetically undefined” stocks such as Sprague-Dawley and Wistar rats, and Swiss and CD-1 mice are produced in closed colonies where each individual is genetically unique. For example there is no definition of a “Sprague-Dawley” rat, and there are no genetic markers which define outbred stocks. Genetic quality control is therefore restricted to determining whether a stock has changed over a period of time and whether any two stocks are similar. Animals from such stocks tend to be phenotypically more variable than isogenic strains and the colony is less stable and can undergo quite rapid genetic change as a result of selective breeding, random genetic drift (particularly in small colonies) and undetected genetic contamination with animals from a different stock. The main advantages of outbred stocks are that they are more vigorous and cheaper than inbred strains and that they have a long tradition of use in toxicity testing.

Inbred strains and F1 hybrids (the first generation cross between two inbred strains) are “genetically defined” so that it is possible using genetic markers to determine whether an individual is e.g. an inbred F344 rat. There are more than 150 inbred rat strains and over 500 mouse strains used in

research throughout the world. Inbred strains are more stable than outbred stocks. They cannot be changed by selective breeding, although there are sub-lines of many of the most widely used inbred strains. These arose as a result of residual heterozygosity because some strains were not fully inbred at the time that different breeding colonies were established, and as a result of new mutations (Stevens et al., 2007). Further details of these two classes of stock and strains and the genetic nomenclature rules are given elsewhere (Festing and Lutz, 2010, 2011).

Most toxicity testing of foods, food constituents, food additives or food contaminants is done using outbred Wistar or Sprague-Dawley rats or CD-1 mice. It has been argued that the use of a small battery of inbred strains in a multi-strain assay would be more sensitive and would reduce the number of false negative results (Festing, 2010). However, given the large experience with using outbred stocks for testing of foods and food constituents, and the available data base on sensitivity and variation in test parameters of the test animals, continuing their use is recommended until evidence to justify a change becomes available. Chosen stocks or strains should be designated according to internationally accepted nomenclature rules. The reason for choosing a particular strain or stock should be clearly stated.

5. Experimental Design and Statistical Methods

In addition to the aspects below, further considerations when designing the experiment and applying statistical methods are provided in Appendix 1 – Statistical principles and good experimental design.

5.1. Confirmatory versus exploratory test

The applicant should clearly state the purpose of the study, e.g. confirmatory or exploratory, and the hypothesis to be tested in advance and documented in the protocol.

For confirmatory studies the power calculation, statistical analysis and statistical reporting should be directly related to the study objectives and statistical hypotheses. The statistical hypotheses (i.e. the null and alternative) should be clearly stated. The endpoint(s) of primary interest should be stated and the sample size/power should be calculated using a biologically relevant effect and its associated expected standard deviation. If the sample size/power is calculated on the basis of a standardised effect size then it should also be biologically relevant. In the event of multiple endpoints the issue of multiple testing (i.e. multiplicity) should be addressed.

Exploratory experiments can be seen as hypothesis generating that can be verified in future experiments/studies. The objectives of an exploratory experiment should be clearly stated in a clear and concise manner. In contrast to confirmatory experiments the hypotheses may be difficult to state in advance of the experiment and might be generated by exploratory analysis. As such analyses are data dependent, caution should be taken interpreting the results and drawing strong conclusions.

5.2. Experimental design considerations

The objectives of a proposed experiment should be clearly stated. It should be designed to be unbiased, with no systematic differences among groups apart from the treatment. This is mainly controlled by assigning animals to the treatments at random (randomisation), by housing the animals at random within the animal house (as far as this is practical), by making measurements in random order and by blinding the staff to the treatment group to which a subject belongs (especially important for behavioural measures, ophthalmology and pathology measures).

The experiment should be powerful: if there is a true difference between the treatment groups, then the experiment should have a good chance of detecting it. Power depends on controlling inter-individual variations, on the magnitude of the difference between the treated groups, on sample size, and on the

acceptable levels of false positive (usually set at 5 %) and false negative (often set at 10-20 %) results. The experiment should also have a wide range of applicability. For example, sex-dependent effects may be present, so both sexes should be included. Finally, it should be simple in order to minimise the chance of mistakes being made. Further details are given in Appendix 1.

5.2.1. Formal experimental designs – Randomised block design

The randomised block design involves splitting the experiment up into a number of "mini-experiments" or blocks, which are then re-combined in the statistical analysis. Animals within these blocks can be matched both for initial characteristics such as body weight, and for other possible sources of variation such as location within the animal house (blocks could be housed in different rooms) or the timing of making the measurements/determinations of the end points (e.g. blocks could be processed on different days). As a result, randomised block designs can often be substantially more powerful than a completely randomised design, depending on the magnitude of these sources of variation and are therefore recommended for the experimental design. The randomised block design is described in more detail in Appendix 1 and examples are provided in Appendix 2.

5.2.2. Inclusion of control/reference groups and historical data

Negative control groups are intended to demonstrate the normal state of the animal for comparison with data from treated groups. They also enable comparisons to be made with historical data from previous studies. The negative controls should be like the treated groups in all ways apart from the treatment.

Positive control groups are intended either to demonstrate susceptibility of the animal to a specific toxic effect or to compare the response of the test material in treated animals to that of animals treated with a chemical with known toxicity similar to that of the test material. They can also be used to show whether an experiment has been conducted sufficiently well to be able to detect toxic effects of only moderate severity. In order to assess the sensitivity of the test system, spiking (positive control) of a diet with a particular compound may be considered. For instance in case of food/feed derived from a GM crop which expresses a lectin, this compound may be added separately to one of the test diets in order to discriminate between adverse effects possibly induced by the lectin, and those effects induced as a result of the genetic modification. The procedure of spiking has to be decided on a case-by-case basis and will only be meaningful if the spiked component possesses a toxic potential at the typical level of expression in the whole food/feed. If a positive control group is to be used it should be scientifically justified in the study protocol and in the study report.

In addition to control groups, reference groups may be included which are fed a diet composed of commercially available material similar to the test food/feed, with a known toxicological database, and history of safe use. For instance in case of GM maize, commercially available non-GM maize varieties may be used. The main purpose of reference groups is to show the range of normal values of test parameters found under the conditions of the experiment.

The use of reference groups would substantially increase the number of animals used and there is no assurance that they would help in the interpretation of the results (there may, for example, be no important differences among them). In the ANSES assessment of MON810 data the variability within the reference groups was so low that its usefulness to define the range of normal value was limited (ANSES, 2011). Thus, for ethical, economic and scientific reasons the use of reference groups is not recommended unless there is no acceptable historical background data available.

Historical control data on natural variations in values of test parameters should primarily be obtained from databases available in the actual testing facility. Data should have been obtained from several studies during the last 5 years prior to the study, on the same strain, taking into account genetic drift. Data from literature might be added if thought to be informative. A major difficulty in using historical

control data is the comparability of these data with the data obtained from the study actually performed with respect to, among others, test animal strains used, dietary factors, experimental environmental conditions etc. Therefore a careful evaluation by the applicant on the use of historical control data is required.

5.2.3. Specification of the experimental unit as a cage

Toxicity studies differ in the number of animals allocated to each cage which range from usually one up to five animals. However, for reasons given in section 5.1 the animals should be housed two per cage in order to minimise stress and any resulting inter-individual variability. The cage with two animals in it will then become the experimental unit (ExpU), i.e. the entity that is randomised to the treatment groups.

The statistical analysis should first test whether there are cage effects. If not, then the statistical analysis can be based on the individual animals. However, if there are significant cage effects, then the analysis should be based on the mean of the two animals within each cage. Housing more than two animals per cage will result in a reduction in the number of ExpU ("n"), so it should be avoided. Animals should be individually identified and all data should be collected separately for each animal, except for food and water consumption, and urine and faeces.

5.2.4. Determination of sample size and power

An appropriate sample size (number of ExpU) can be estimated from a number of variables; consideration of the effect size of scientific interest (the "signal"), the variability of the experimental material (the "noise"), the significance level (usually set at 5 %), the power (often set at 80-9 0%) and the alternative hypothesis (one or two sided) (for additional details see Appendix 1). The relationships among these variables is shown graphically in Figure 1, where the effect size is specified in terms of standard deviations. This is known as the "standardised effect size" (difference between treatment groups divided by its SD) and can be regarded as a signal/noise ratio.

In many cases it may be challenging to specify the magnitude and standard deviation of the biologically relevant difference between the treated and control groups for a given endpoint. This becomes even more challenging when considering multiple endpoints which is the case in toxicological studies. For confirmatory trials, the endpoints of interest should be identified prior to designing the experiment.

For exploratory trials an approach is proposed which aims at designing experiments to detect a standardised effect size of about 1, whilst aiming to achieve 80-90 % power. Standardised effect sizes of up to this magnitude seem to have little biological relevance in relation to toxicity. For example, a number of responses of this magnitude in the trials of MON863 were judged not to be toxicologically significant (EFSA, 2007a, 2007b). In the ANSES assessment of MON810, the detectable effect size that was determined to calculate power of the test of difference is at least equal to one standard deviation of the control data (ANSES, 2011). The applicant should justify their choice of selected standardised effect size that they wish to detect.

The assumptions used for the calculations of the graphs in Figure 1 do not include a sex by treatment interaction. However, the analysis should investigate sex by treatment interactions, or any other interactions that are deemed to be of importance.

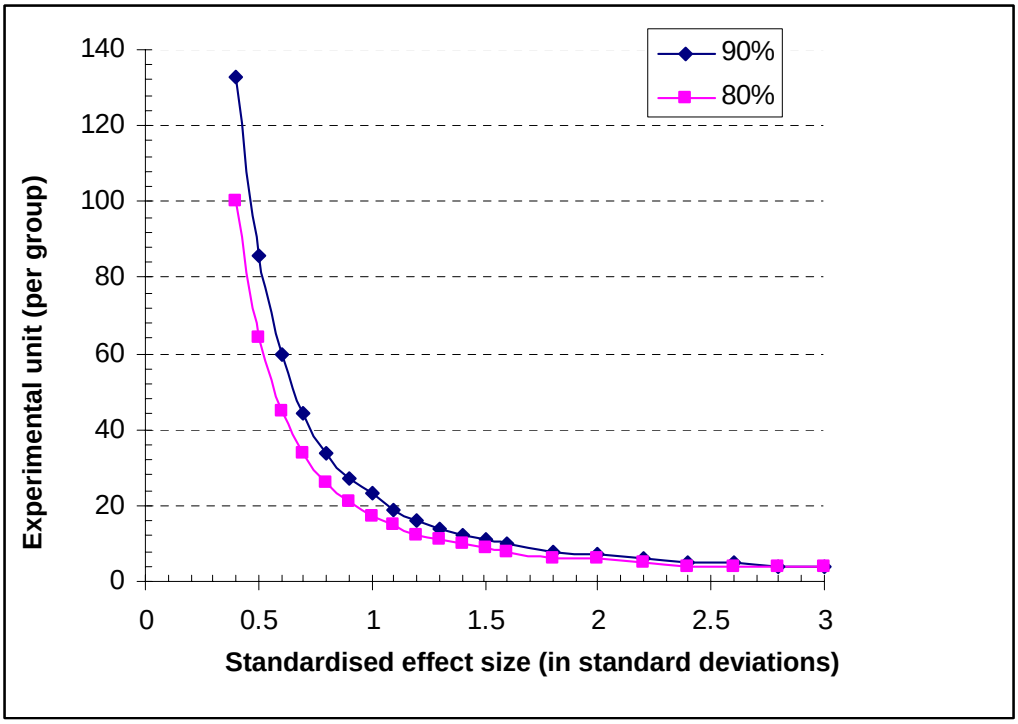

Figure 1: Number of experimental units needed per treatment group as a function of Standardised Effect Size for an 80 % and 90 % power and 5 % significance level using a two-sided t-test. This will approximate the situation in a 2 (treatments) x 2 (sexes) factorial design.

OECD TG 408 suggests for the testing of chemicals the use of 80 animals (without giving any sample size justification) comprising both sexes in four treatment groups (low, medium and high dose, and the control group). These four treatment groups each consist of 20 animals . It should be noted that this is a "completely randomised design" and not a "randomised block design". It can be seen from Figure 1 that the OECD test guideline (assuming a completely randomised design, with four treatment groups of 20 animals, one animal per cage), has about an 80 % chance of detecting a standardised effect size of 0.9 standard deviations and a 90 % chance to detect a standardised effect size of 1.1 standard deviations (SD) assuming no sex by treatment interaction. ANSES bases their recommendations for the number of animals, 20 animals per group and per sex, which corresponds to 40 experimental units per treatment group, on a similar design (i.e. a randomised design with one animal per cage) (ANSES, 2011).

This guidance proposes a strategy for the experimental design which aims to maximise the power of the experiment to detect a standardised effect size of one standard deviation (1 SD), while avoiding for ethical reasons the use of a substantially higher number of test animals. The maximisation of the power is achieved by reducing the number of dose groups to low and high dose groups in order to maximise the numbers of control and top dose animals (see section 2.2.1). The power is further increased by decreasing the inter-individual variability by housing animals in pairs (see section 4.1) and by applying a randomised block design (see section 5.2.1).

Power can also be increased by increasing the sample size. However, as can be seen from Figure 1, the number of animals needed to detect a standardised effect size of much less than 1 SD increases exponentially.

Based on these considerations, examples of experimental designs for novel food and GM food are provided below. The examples are illustrative of designs that aim to detect standardised effect size of around 1 SD. Alternative designs and/or standardised effect sizes can be used, provided scientific justification is given. Even the use of either of the two example designs should be scientifically justified.

630 An example of a randomised block design for testing novel foods, designed to maximise power,
631 involves eight blocks, each of six cages with two animals per cage. Each block includes a control, low
632 and high dose in both males and females. There will therefore be 16 experimental units (corresponding
633 to 32 animals, 16 cages with two animals each (i.e. 16 per sex)) at each of the three treatment groups
634 and a grand total of 96 animals. The analysis strategy should start by testing the high dose against the
635 control and if it is statistically significant and biologically meaningful then the low dose should be
636 tested against the control. This analysis strategy partially addresses the issue of multiplicity (Hochberg
637 1988). This design has 80 % power to detect an effect size of 1.02 SD and 90 % power to detect an
638 effect size of 1.18 SD.

639 In the case of GM foods an example of a randomised block design involves four treatment groups: a
640 low and high level of the test food and the same levels of the isogenic comparator, tested in both sexes
641 (eight cages total). There are six blocks, each of eight cages. There will therefore be 12 experimental
642 units per treatment group (i.e. 12 cages with two animals each (24 animals, 12 per sex). In this case 48
643 animals (half of total number) will receive the GM food and half the isogenic comparator. This
644 involves a total of 96 animals. The analysis strategy should start by testing GMO food against isogenic
645 comparator where the high and low dose groups are taken into consideration in the statistical analysis .
646 This design has 80 % power to detect an effect size of 0.83 SD and 90 % power to detect an effect size
647 of 0.96 SD. Further details about the design and analysis of the two examples are given in Appendix
648 2.

649 When the predicted standardised effect size of the endpoints is estimated to be larger than 1 SD based
650 on the choice of values for the endpoints, the applicant should consider increasing the power by using
651 additional animals by adding extra blocks to the design (i.e. aiming towards a detectable standardised
652 effect size of one). Any increase in the use of animals should be justified and carefully balanced with
653 the expected outcome. Detectable standardised effect sizes for the examples when additional blocks
654 are used are found in Appendix 2.

655 The power of the experiment can also be increased by using a higher significance level than 5 %
656 which is the statistical level most commonly used in biological research. By using a higher
657 significance level, e.g. 10 %, effects that fail to reach statistical significance at 5 % would then be
658 considered statistical significant. This would lead to an increased number of statistically significant
659 results to be toxicologically addressed to assess their biological significance. However, a higher
660 statistical significance level will also increase the number of false positive results impacting the
661 toxicological assessment workload. For the assessment, strong emphasis should be placed on the
662 biological relevance of any observed differences whether or not they reach the chosen level of
663 statistical significance. This is best done by looking at the point and interval (e.g. confidence)
664 estimates and not by just focussing on the P-value.

665 Due to the fact that a number of the variables indicated above (i.e. effect size, variability, significance
666 level, power and the alternative hypothesis) will have to be estimated or assumed, the number of
667 experimental units (sample size) will vary according to the choices and justifications made. In many
668 cases it will be challenging to specify how large a difference between the treated and control means for
669 each parameter measured is likely to be important, and there may be no accurate estimates of the
670 standard deviation of each parameter.

671 For some endpoints there could be a sex by treatment interaction, or other types of interactions.
672 Experiments can be powered to consider such interactions, but in practice this is difficult to do without
673 knowing the exact nature of these interactions. To increase the power to detect interactions additional
674 blocks can be added although difficulties in calculating the associated power are acknowledged.
675 Interactions can be tested by using a higher significance level (e.g. 10 %) whilst keeping in mind the
676 issues raised above.

677 When estimating treatment effects in the presence of sex by treatment interactions all the data should
678 be modelled together to maximise the power (treatment effects for each sex should be estimated

separately). In the event that the sample contains siblings, modelling all the data together allows them to be clustered (which will partially reduce the power). Analysing the data for each sex assumes independence between the sexes and is therefore not recommended. .

In the protocol the applicant should justify the sample size calculation including the variables (i.e. effect size, variability, significance level, power and the alternative hypothesis) . In addition, the design of the experiment should also be clearly described including whether it is a "completely randomised design" or a "randomised block design" and the ExpU should be specified (e.g. number of animals/cage).

5.3. Reporting the analysis conducted and reporting of the results

The reporting of the statistical analysis should be consistent with the protocol and the statistical analysis plan. The result should be presented in a consistent and clear manner to facilitate the interpretation by the risk assessors. All the important details about the experiment design and an overview of statistical methods, including the design, and analysis, should be documented in a protocol prior to the start of the trial. Details of the statistical analysis should be documented in a statistical analysis plan (SAP) prior to the completion of the study. It should be signed and dated by at least the responsible statistician. The full statistical analysis performed according to the statistical analysis plan should be included in the study report (or can be written as a separate report and annexed to the study report). Any unplanned analysis should also be detailed in the final study report.

5.3.1. Specification of the methods of statistical analysis and presentation of the results

Suggested steps in the statistical analysis, such as the screening of the data for outliers, transformation of scale where necessary and the choice of the most appropriate statistical tests taking account of the distribution of the observations are given in appendices 1, 3 and 4. These also provide suggestions for the presentation of the data which should include summary statistics such as means and standard deviations as well as measures of the magnitude of differences between groups as assessed using confidence intervals where possible. The EFSA GMO panel has discussed in detail many of the considerations which need to be taken into account in the statistical analysis of data resulting from field trials involving GM plants (EFSA Panel on Genetically Modified Organisms (GMO), 2010).

The statistical analysis should include an assessment of the differences between males and females for each parameter. The parameters with a possible difference between the sexes should be documented in the protocol. The statistical analysis plan should detail all the analysis methods with all the results reported in the final report. Sex-limited traits (i.e. ones such as testis and uterus weights) which can only be measured in one sex should be analysed using an appropriately reduced analysis of variance. Any statistically significant interactions, particularly those involving treatment and gender should be fully explored using sub-group analyses. Failure to find sex differences for parameters where such differences are commonly found would suggest that the investigators have failed to control inter-individual variation or make the measurements accurately, suggesting that the experiment is of poor quality.

The separate analysis of many parameters, most of which are not expected to differ between treatment groups, may result in a large number of statistical tests. This will lead to the issue of multiple testing (multiplicity) and therefore it should be addressed by the applicant in the protocol, statistical analysis plan and study report. Any methods used to adjust for multiplicity should also be clearly documented and referenced. With a randomised block design, block is also a random factor which should be included in the model.

The protocol should describe the intended methods of statistical analysis and the methods employed to minimise the bias (see Appendix 2 for further details). The statistical analysis should provide the full

details of the intended analysis including full descriptions of the statistical models fitted. The following should be addressed in the statistical analysis plan:

- Key objectives of the analysis (including whether the analysis should be considered as confirmatory or exploratory).
- Hypothesis to be tested (clarify if testing is for superiority or equivalence)
- The presentation of summary statistics (means, medians, standard deviations etc)
- Clear specifications of all models including the adjustments for covariates including interactions
- Longitudinal or repeated data should be modelled using appropriate techniques (e.g. linear or non-linear mixed models)
- Choice of appropriate statistical methods including parametric and non-parametric methods
- All assumptions should be clearly stated
- The separate analysis of growth data
- Handling of missing data
- The identification and handling of outliers
- Data transformations, where appropriate
- Interim analyses and data monitoring
- Multiple comparison/multiplicity
- Examination of subgroups

Where the statistical analysis is conducted by sex the results should be presented consistently for each sex and for both sexes combined to assist the risk assessor.

5.3.2. Descriptive statistics

Descriptive statistics should be presented for all environmental and analysis variables (endpoints). The summary statistics should include the mean, standard deviation, median, lower quartile, upper quartile, minimum and maximum. Table 1 in Appendix 4 presents an example of how summary statistics can be presented.

The use of graphical methods such as plots of means and 95 % confidence intervals for each group, and/or box and whisker plots is encouraged.

5.3.3. Analysis of results

The results from the statistical analysis should be presented in the original units and in terms of the standardised effect size using point and intervals estimates (e.g. confidence) as presented in Table 2 and Table 3 in Appendix 4.

If the results are also expressed in terms of standardised effect sizes (differences between treatment groups)/SD with 95 % confidence intervals (Nakagawa and Cuthill, 2007), then this ratio is in standard deviation units (the signal/noise ratio, see section 5.2.4) and all parameters can be shown on the same graph. This makes it easier to see the pattern of response across a range of parameters. As many parameters are likely to be correlated there may be a slight excess of statistically significant comparisons above what would be expected from the use of a 5 % significance level. However, the biological relevance of all statistically significant differences as well as the point and interval (e.g. confidence) estimates of any responses (some of which may not reach statistical significance) should be considered by an appropriately qualified toxicologist (see section 6).

5.3.4. Individual data

All individual data should be provided.

6. Interpretation of results of animal studies

Interpretation of data from the animal feeding trials requires extensive expertise in many different scientific fields like e.g. toxicology, chemistry, biological chemistry, animal nutrition and an understanding of statistics. Any effects observed in the animals should be evaluated in order to assess their relevance for the safety of the whole food for humans or of the whole feed for target animal species.

Observed differences in test parameters between treated and control groups must be investigated, discussed and reported in the study report with respect to a number of considerations as indicated in the following sections (6.1 to 6.7)

6.1. Dose-related trends

The magnitude of the effect is expected to increase with the dose level in severity and/or incidence, thus providing an indication for a causal effect, although it is recognized that where small effects are being investigated such a trend may not be observed. Absence of a dose-response relationship may be due to the limited dose range applied or may indicate that the effect is accidental or spurious. When a difference is only noticed at the highest dose level, factors like type and magnitude of the finding, frequency, normal trends and ranges, correlation with other findings should be considered to determine whether a treatment relation exists or a casual artifact has occurred. Supportive data for a possible causality between the test food/feed and effects in test animals may include, for example, additional toxicity (if available) or predictive data from in vitro and in silico experiments.

6.2. Possible interrelationships between test parameters

Changes in organ weights should be normalized to body weight/brain weight in order to eliminate influence of normal variation in animal growth. Furthermore changes in body weight may be the result of a changed intake of a more or less palatable diet.

The change of an isolated parameter is often of limited interest and the conclusion on biological significance depends on several parameters (haematology, biochemistry and pathology). Observed changes in single test parameters may be interconnected thus strengthening the indication that an effect has occurred as a result of the treatment. For example, liver damage, observed as a change in histopathology, gross pathology, and organ weights, may also be evident from changed levels in serum of liver-derived enzymes, or bilirubin. Detection of toxic responses in the blood by hematological analysis may be interlinked with results from the analysis of bone marrow, spleen, lymph nodes and mononuclear phagocyte system (reticuloendothelial tissue) of various organs and tissues.

6.3. Occurrence of effects in both genders

Effects often occur in both male and females animals, but in certain cases one gender may be more sensitive than the other due to differences for example in detoxification mechanisms, or due to differences in hormonal metabolism (endocrine effects) (see also section 5.4.1).

6.4. Reproducibility

Differences observed in treated animals may also have been observed in other studies in the same or in another animal species.

6.5. Animal species specificity of effects

Certain effects may be specific for the test species but not of value for humans or other species (for example nephro-pathological effects of hydrocarbons in rodents due to accumulation of a male-specific rat protein, which is absent in humans).

6.6. Background range of variability

If the change observed in a certain parameter falls within the background range of variability, this may indicate that the investigated food/feed does not cause a health problem. However further aspects should be considered in relation to gender specificity or linkage with other changes in order to exclude potential adverse effects upon consumption of the food.

7. Assumptions and uncertainty analysis

With respect to the overall risk evaluation of the results obtained from the animal feeding trial, it should be indicated what assumptions have been made during the risk assessment in order to predict the probability of occurrence and severity of adverse effect(s) in a given population, and the nature and magnitude of uncertainties associated with establishing these risks.

Although it may be impossible to identify all the uncertainties, each scientific output should describe the types of uncertainties encountered and considered during the different risk assessment steps, and indicate their relative importance and influence on the assessment outcome (EFSA, 2009b).

Any uncertainties in the design of the experimental model which might influence the power of the experiment should be highlighted and quantified as far as possible. In particular, attention should be paid to the specificity (choice of the test animal species) and sensitivity of the test model and to uncertainties related to extrapolation of results to humans or target animal species exposed to the whole food/feed under investigation. Distinction should be made between uncertainties that reflect natural variations in biological parameters (including variations in susceptibility in populations), and possible differences in responses between species.

7.1. Additional animal studies

Results from the 90-day study may trigger additional studies. It is also noted that the subchronic, 90-day rodent feeding study is not designed to detect effects on reproduction or development, other than effects on adult reproductive organ weights and histopathology, and, therefore, also depending on the outcome of the 90-day feeding study, further animal studies on potential effects on reproduction/fertility may be required.

8. Study performance and documentation

8.1. Study performance

The specific procedures (including quality control) used to implement and adhere to the principles outlined in this guidance are the responsibility of the sponsor of the study. The sponsor should also ensure that the team conducting the experiment are appropriately qualified and experienced. The sponsor is also charged with ensuring that all the important details about the experiment, including the design, conduct and analysis, are documented in a protocol and reported in the study report.

8.2. Protocol

The protocol should be written and signed off prior to the start of the experiment by the study team who are suitably qualified and adequately experienced. All the important details about the experiment design and an overview of statistical methods, including the design, and analysis, should be documented in a protocol. Any amendments should also be documented and signed off.

8.3. Statistical Analysis Plan

The statistical analysis plan (SAP) should be written and signed off prior to the end of the experiment by the study team who are suitably qualified and adequately experienced. Any amendments should also be documented and signed off.

8.4. Statistical Report

A statistical report should be written with all the analysis results as documented in the SAP. The programs, logs and outputs should be provided for the purposes of the review.

8.5. Full Study Report

The outcome of the study should be provided to the risk assessor in an integrated full study report describing all the steps of the study. The investigator should provide the protocol developed for the study, the statistical analysis plan for the statistical assessment of the data which should be developed before the end of the actual experiment. The protocol, statistical analysis plan and the statistical analysis report could be annexed to the study report. An outline for the study report, protocol and statistical analysis plan is provided in Appendix 3.

The aim of the integrated full study report is to provide all necessary information required by the risk assessor in a comprehensive way with clear presentation of the results of the study. The study report should include description and aim of the experiment, methods, results, tables and figures, analyses performed, discussion of the results and references.

CONCLUSION OF THE GUIDANCE

The safety assessment of GM food/feed and novel foods is comprised of an extensive compositional analysis and a toxicological and nutritional characterization of specific compounds identified in these whole products, rather than of the toxicological/nutritional testing of the whole products themselves. However, testing of the whole food/feed may be necessary depending on the available information, and therefore the development of principles and practical rules to perform animal feeding trials with such products, is of great importance.

Appropriate characterization of the whole food/feed to be tested is required and should include among others a description of the source, its composition, the manufacturing process, information on stability and the presence of chemical and/or microbiological contaminants. Furthermore, preparation of appropriate test diets is a key element of the animal feeding trial with respect to the choice of the diet type, nutritional balance and necessary adjustments, processing, and storage. The goal is to achieve as high level as possible of the whole food to be incorporated in the animal diets without causing nutritional imbalance or metabolic disturbance

There are two major classes of laboratory mice and rats used in research and testing: outbred stocks and isogenic strains (inbred and F1 hybrid). Given the large experience with using outbred stocks for testing of foods and food constituents, and the available data base on sensitivity and variation in test parameters of the test animals, their use is recommended.

For ethical and scientific reasons the test animals should be housed two (of the same sex) per cage. The experimental unit (ExpU) is a cage containing two animals which should be individually identified with separate records. Animals should be less than nine weeks old at the start of the experiment, be healthy and free from pathological micro-organisms.

A randomised block design should normally be used with the animals within a block being matched for age and weight (for each sex) and location within the animal house. This design helps to reduce uncontrollable variation especially when the experiment needs to be housed in more than one room or spread over a period of time. Further increase in power of the experiment, when considered relevant, could be achieved by adding extra blocks to the randomised block designs.

Due to the fact that a number of the variables (i.e. effect size, variability, significance level, power and the alternative hypothesis) will have to be estimated or assumed, the number of animals (sample size) will vary according to the choices and justifications made. The applicant should describe and justify the calculation of sample size and the values of the variables used in the protocol. In addition, the design of the experiment should be clearly described including whether it is a "completely randomised design " or a "randomised block design" and the experimental unit should be specified (e.g. number of animals/cage).

It is important to identify and limit the impact of any potential sources of bias as completely as possible. The presence of bias is likely to seriously compromise the ability to draw valid conclusions from the experiment.

Examples of experimental design for testing whole food/feed are provided which use 96 animals. When needed additional animals can be added in blocks. The examples are illustrative of designs that aim to detect standardised effect size around one standard deviation. Alternative designs and/or standardised effect sizes can be used, provided scientific justification is given.

Animals should remain on the test diets for a period of 90 days. A comprehensive set of end-points should be measured at the end of this period. An interim collection of data from blood samples should normally be taken after 45 days. All animals should be weighed once per week.

Since it is often not possible to include whole foods in an amount that will induce toxicity and thus to obtain a dose-response relationship, the application of two dose levels is recommended to maximise the power. The highest dose level of the whole food/feed that can be incorporated in the animal diet should not cause nutritional imbalance or metabolic disturbances in the test animal, and the lowest dose level should always be above the anticipated human/target animal intake level.

The inclusion of reference groups in the experimental design, fed with a diet containing commercially available food/feed similar to the test food/feed, , in order to estimate the natural variability of test parameters, is in general not recommended. Historical background data on variations in test parameter values should in principle be obtained from existing databases available in the testing facility or in the public domain. Inclusion may be considered if no acceptable historical background data available.

A statistical analysis of the differences between males and females should be included as a check on the quality of the study, with the results being included in the study report. The gender differences should be discussed in relation to historical data. When estimating treatment effects in the presence of sex by treatment interactions all the data should be modelled together to maximise the power (treatment effects for each sex should be estimated separately). In the event that the sample contains siblings, modelling all the data together allows them to be clustered (which will partially reduce the power). Analysing the data for each sex assumes independence between the sexes and is therefore not recommended.

It is emphasized that the biological relevance of any observed differences whether or not they reach the chosen level of statistical significance. This assessment should involve the use of point and interval (e.g. confidence) estimates in addition to the significance level.

Equivalence between two diets can only be concluded from an experiment designed to test for equivalence using appropriate statistical methods. Equivalence cannot be concluded by observing "non-significant" P-values from an experiment designed for superiority (i.e. absence of evidence is not evidence of absence).

The study report should include descriptive statistics including the number in each group, means, standard deviations, medians, lower quartiles, upper quartiles, minimums, maximums and the 95 % confidence intervals separately for each parameter and treatment group, by gender. Confidence intervals and P-values should be shown for every comparison. Results should be presented in such a way as to facilitate interpretation. Graphical methods, particularly the presentation of means with confidence intervals, should be used. Consideration should be given to expressing results in terms of standardised effect sizes. Any strong correlations between parameters should be noted.

944 **REFERENCES**

945 Aggett P.J., Antoine J.M., Asp N.G., Bellisle F., Contor L., Cummings J.H., Howlett J., Müller D.J.G.,
946 Persin C., Pijls L.T.J., Rechkemmer G., Tuijtelaars S. & Verhagen H. (2005). Process for the
947 Assessment of Scientific Support for Claims on Foods (PASSCLAIM): Consensus on criteria.
948 Eur.J.Nutr. 44 (supplement 1), 5-30.

949 ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail),
950 2011. Opinion of the French Agency for Food, Environmental and Occupational Health and Safety
951 - Recommendations for carrying our statistical analyses of data from 90-day rat feeding studies in
952 the context of marketing authorisation applications for GM organisms. Request no. 2009-SA-0285.

953 Chvedoff M, Clarke M, Faccini JM, Irisari E and Monro AM, 1980. Effects on mice of numbers of
954 animal per cage: an 18-month study. (preliminary results). Archives of Toxicology, Supplement 4,
955 435-438.

956 Clarke, HE, Coates ME, EVA JK, Ford DJ, Milner CK, O'Donoghue PN, Scott PP and Ward RJ,
957 1977. Dietary standards for laboratory animals: report of the Laboratory Animals Centre Diets
958 Advisory Committee. Laboratory Animals 11, 1-28.

959 EFSA (European Food Safety Authority), 2005. Opinion of the Scientific Panel on Animal Health and
960 Welfare on a request from the Commission related to "Aspects of the biology and welfare of
961 animals used for experimental and other scientific purposes". The EFSA Journal, 292, 1-46
962 http://www.efsa.europa.eu/en/efsajournal/doc/292.pdf

963 EFSA (European Food Safety Authority), 2007a. Statement on the Analysis of Data from a 90-Day
964 Rat Feeding Study with MON 863 Maize by the Scientific Panel on Genetically Modified
965 Organisms (GMO). European Food Safety Authority,
966 http://www.efsa.europa.eu/EFSA/efsa_locale-1178620753812_1178621169104.htm

967 EFSA (European Food Safety Authority), 2007b. EFSA Review of Statistical Analyses conducted for
968 the Assessment of the MON 863 90-Day Rat Feeding Study. European Food Safety Authority,
969 http://www.efsa.europa.eu/EFSA/efsa_locale-1178620753812_1178621342614.htm

970 EFSA GMO Panel Working Group on Animal Feeding Trials, 2008. Safety and nutritional assessment
971 of GM plants and derived foodand feed: The role of animal feeding trials. Food and Chemical
972 Toxicology 46:S2-S70.

973 EFSA (European Food Safety Authority), 2009a. Opinion of the Scientific Committee on a request
974 from EFSA on existing approaches incorporating replacement, reduction and refinement of animal
975 testing: applicability in food and feed risk assessment. The EFSA Journal, 1052, 1-77.

976 EFSA (European Food Safety Authority), 2009b. Transparency in Risk Assessment – Scientific
977 Aspects, Guidance of the Scientific Committee on Transparency in the Scientific Aspects of Risk
978 Assessments carried out by EFSA. Part 2: General Principles. The EFSA Journal, 1051, 1-22.

979 EFSA Panel on Genetically Modified Organisms (GMO), 2010. Scientific Opinion on Statistical
980 considerations for the safety evaluation of GMOs. EFSA Journal 8(1):1250, pp 1-59.

981 EFSA Panel on Genetically Modified Organisms (GMO), 2011. Scientific Opinion on Guidance for
982 risk assessment of food and feed from genetically modified plants. EFSA Journal 9(5): 2150. Pp 1-
983 37 pp.

984 FDA (Food and Drug Administration) Redbook, 2000. Guidance for Industry and Other Stakeholders,
985 Toxicological Principles for the Safety Assessment of Food Ingredients. July 2000; Revised July
986 2007. U.S. Department of Health and Human Services, Food and Drug Administration, Center for
987 Food Safety and Applied Nutrition. http://www.cfsan.fda.gov/guidance.html.

988 Festing MF, 2010. Inbred strains should replace outbred stocks in toxicology, safety testing, and drug
989 development. Toxicol.Pathol. 38(5):681-90.

990 Festing MFW and Lutz C, 2010. Introduction to laboratory animal genetics. In: Hubrecht R, Kirkwood
991 J, editors. The care and Management of Laboratory and Other Research Animals. 8th. ed. Oxford,
992 Ames: Wiley-Blackwell, pp. 37-60.

993 Festing MFW and Lutz C, 2011. Laboratory Animal Genetics and Genetic Quality Control. In: Hau J,
994 Schapiro SJ, editors. Handbook of Laboratory Animals Science. 3rd. ed. CRC Press, 209-50.

995 Hochberg, 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika
996 75(4):800-802.

997 Knudsen I and Poulsen M, 2007. Comparative Safety Testing of Genetically Modified Foods in a 90-
998 Day Rat Feeding Study Design Allowing the Distinction between Primary and Secondary Effects
999 of the New Genetic Event. Regulatory Toxicology and Pharmacology, 49, 53-62.

1000 Leshem M and Sherman M, 2006. Troubles shared are troubles halved: stress in rats is reduced in
1001 proportion to social propinquity. Physiol Behav, 89(3):399-401.

1002 Nakagawa S and Cuthill IC, 2007. Effect size, confidence interval and statistical significance: a
1003 practical guide for biologists. Biol Rev Camb Philos Soc, 82:591-605.

1004 NRC (National Research Council), 2005. Nutrient requirements of laboratory animals. Fourth revised
1005 edition, 1995.

1006 OECD Guideline for the Testing of Chemicals – Repeated Dose 90-day Oral Toxicity Study in
1007 Rodents, 408, 1998. http://browse.oecdbookshop.org/oecd/pdfs/free/9740801e.pdf

1008 OECD Guideline for the Testing of Chemicals – Repeated Dose 28-day Oral Toxicity Study in
1009 Rodents, 407, 2008. http://browse.oecdbookshop.org/oecd/pdfs/free/9740701e.pdf

1010 OECD Series on Principles of Good Laboratory Practice (GLP) and Compliance Monitoring, 1998.
1011 http://www.oecd.org/document/63/0,2340,en_2649_34381_2346175_1_1_1_1,00.html

1012 Poulsen M, Schrøder M, Wilcks A, Kroghsbo S, Lindecrona RH, Miller A, Frenzel T, Danier J,
1013 Rychlik M, Shu Q, Emami K, Taylor M, Gatehouse A, Engel KH, Knudsen I, 2007. Safety testing
1014 of GM-rice expressing PHA-E lectin using a new animal test design. Food/feed Chem. Toxicol. 45,
1015 364-377.

1016 Rao GN and Knapka JJ, 1987. Contaminant and Nutrient Concentrations of Natural Ingredient Rat and
1017 Mouse Diet Used in Chemical Toxicology Studies. Fundam Appl. Toxicol. 9, 329–338.

1018 Russell WMS and Burch RL, 1959. The principles of humane experimental technique. Potters Bar,
1019 England: Special Edition, Universities Federation for Animal Welfare.

1020 Stevens JC, Banks GT, Festing MF, Fisher EM, 2007. Quiet mutations in inbred strains of mice.
1021 Trends Mol.Med. 13(12):512-9.

1022 Stevens KA and Russel RJ, 2007. Chapter 10: Nutrition. The mouse in the biomedical research. Eds
1023 JG Fox, S Barthold, M Davisson, CE Newcomer, FW Quimby, A Smith. II Edition Academic
1024 Press.

1025 TASCC (Trade Assurance Scheme for Combinable Crops), 2010. Code of Practice for the Storage of
1026 Combinable Crops and Animal Feeds. Effective from July 1st 2010.
1027 http://www.agindustries.org.uk/document.aspx?fn=load&media_id=3734&publicationId=2150

1028 Verhagen H, Aruoma OI, van Delft JHM., Dragsted LO, Ferguson LR, Knasmüller S, Pool-Zobel BL,
1029 Poulsen HE, Williamson G, Yannai S, 2003. Editorial - The 10 basic requirements for a scientific
1030 paper reporting antioxidant, antimutagenic or anticarcinogenic potential of test substances in in
1031 vitro experiments and animal studies in vivo. Food and Chemical Toxicology 41: 603-610.

1032 Westenbroek C, Ter Horst GJ, Roos MH, Kuipers SD, Trentani A, den Boer JA, 2003. Gender-
1033 specific effects of social housing in rats after chronic mild stress exposure. Prog.
1034 Neuropsychopharmacol. Biol. Psychiatry 27(1):21-30.

1035

APPENDICES

APPENDIX 1 – STATISTICAL PRINCIPLES AND GOOD EXPERIMENTAL DESIGN

This statistical Appendix gives further details of the principles of experimental design and the reasons for the suggested modifications to the OECD 408 for assessing the possible toxicity of novel and genetically modified food using a repeated-dose 90-day oral toxicity in rodents. It is recommended that any planned toxicity test should be preceded by one or more pilot studies. These should be used to test the logistics of the proposed study, ensure that the staff are adequately trained and that all apparatus is available and all the proposed measurements can be made to the required level of accuracy. It can also provide preliminary information on dose levels and inter-individual variability.

1. Controlled experiments

There is an extensive literature on methods of designing and analysing formal experiments (Fisher, 1960; Cox, 1958; Cochran et al., 1957; Montgomery, 1984; Mead, 1988). The principles of the design, statistical analysis and interpretation of experiments relating specifically to assessing GM foods have been reviewed by the (EFSA, 2007a, 2007b, GMO Panel Working Group on Animal Feeding Trials, 2008; Hartnell, 2007).

2. Types of comparison: Superiority vs equivalence

If the trial objective is to show a clear toxicological or beneficial effect of the whole food compared to a control group then the experiment should be designed for superiority (i.e. testing for a difference). If the trial objective is to show "toxicological equivalence" of the whole food compared to a control group then the experiment should be designed for equivalence. Equivalence cannot be concluded based on an observed non-significant p-value when testing a superiority null hypothesis.

The sample size calculations, analysis, reporting and interpretation should reflect the chosen objective in the appropriate sections of the protocol, statistical analysis plan and study report. Confidence intervals for the treatment effect compared to the control group should always be presented.

3. Blinding and randomization

Blinding/masking staff during the experiment reduces the risk of any unconscious or conscious bias as a result of the way the animals are handled and/or assessed due to the knowledge of the treatment groups. Partial blinding/masking (e.g., knowing that certain animals are in the same treatment group if the feed/food is labelled A, B, C, etc) could also lead to similar problems. The level of blinding should be detailed in the protocol with complete information including the people (including roles) who were blinded and those who were not. All measures taken to minimise bias and how they are to be assessed should also be detailed in the protocol.

The details of the randomisation methods, associated procedures and staff with access to the coding list should be documented. Details about blocking and stratification should also be given. The method used to generate the randomisation list should be reproducible (e.g. a predefined fixed seed should be used to generate the randomization list) and any associated programs, logs or listing should be provided in the study report. The date of randomisation and unblinding should be documented in the study report.

4. Considerations when designing an experiment

There are five requirements for a well designed experiment which are further detailed below:

1. Absence of bias
2. High power
3. A wide range of applicability
4. Simplicity
5. Being amenable to a statistical analysis.

The following sections (4.1 to 4.5) provides additional information on the five requirements.

4.1. Absence of bias

Experimental bias should be minimised. The term bias is interpreted, slightly modified from the ICH E9 guidelines (ICH, 1998), as "the systematic tendency of any factors associated with the design, conduct, analysis and interpretation of the results of trials to make the estimate of feed/food effect deviate from its true value". It is important to identify and limit the impact of any potential sources of bias as completely as possible. The presence of bias is likely to seriously hamper the ability to draw valid conclusions from the experiment.

Bias can arise as a result of improper design (e.g. putting the cages for the control group at the bottom and the highest dose group at the top), during the conduct of the experiment (e.g. systematically taking measures of animals in some treatment groups in the morning and others in the afternoon) or as a result of the analysis method (e.g. by not including key factors in the statistical models).

Bias may lead to false positive or negative results, so it is important to ensure that any possible bias is minimised. This can be achieved by:

1. Correct identification of the "experimental unit" (ExpU), defined as the smallest division of the experimental material such that any two ExpUs can receive different treatments. This is important in diet studies because, for ethical reasons, rodents should not be housed individually. However, the animals within a cage can't receive different treatments. Assuming that the animals are housed in pairs, then the ExpU is the cage, with two animals in it and the statistical analysis should be based on the mean of the two animals. Inter-individual variability is somewhat reduced by averaging across two animals by regarding the cage as the ExpU.

2. Randomisation of the ExpUs to the treatments using a formal method based on random numbers. This or similar randomisation should continue throughout the experiment, including when the data is collected.

3. Staff should, where possible, be "blinded" to the experimental treatment. Diets should be coded so that staff do not know to which treatment group individual ExpUs (individual animals) belong. This is particularly important if there is any subjective element to assessing experimental outcomes. For example, pathologists should be blind to the treatment group when assessing histological slides.

4.2. High power

Statistical "power" is the ability of the experiment to detect a treatment effect, if it exists. Low powered experiments will have an increased chance of false negative results. For quantitative outcomes investigators should attempt to achieve a high signal/noise ratio, where the signal is the response (difference between means of treated and control groups) and the noise is the variation within

the groups quantified by the standard deviation. For binary or discrete parameters the aim should be to maximise the response to the treatment. In both cases power might be increased by increasing sample size although cost, ethics and the law of diminishing returns set a practical upper limit.

The power of the experiment can be increased by adjusting the variables presented below.

4.2.1. Reducing the variability of the experimental material (the "noise")

Treated and control groups should be as similar as possible at the start of the experiment. As far as possible animals should be the same weight and age, they should be free of pathogens and they should be housed in optimum conditions.

With large experiments it is difficult to ensure that both the animals and the environmental conditions are reasonably homogeneous. Blood samples or behaviour measurements taken in the morning may be different from those taken in the afternoon due to circadian rhythms, and it may not be possible to do all the measurements on many animals in one short time period in one day. There may also be day-to-day fluctuations in the environment. Housing conditions may vary, with the top shelves getting more light and heat than lower shelves, etc. All these environmental factors can increase inter-individual variability and therefore reduce the power of the experiment.

A way to reduce the variability is to split the experiment up into smaller, more easily managed, parts (i.e. blocks) using a randomised block experimental design. Typically each block contains a single ExpU (usually a cage of two animals) on each treatment. For example, a block may consist of three cages of males and three of females each receiving one of the three treatments (control, low dose and high dose) assigned at random within each sex. The animals of each block would then be housed on the same shelf and they would be bled, weighed and measured within a short time period.

4.2.2. Increasing magnitude of the response (difference between treated and control group; "signal")

The larger the treatment effect (the signal), the higher the power of the experiment, other things being equal. When testing small molecules the signal is usually increased by giving high dose levels, up to the maximum tolerated dose. However, this may be difficult with whole foods in view of the limitations in dose levels due to bulkiness and satiation.

The magnitude of the response also depends on the sensitivity of the experimental material. If strains or species of animals are available which are known to be particularly sensitive to the type of treatment being investigated, then consideration should be given to using them. Inbred strains may be intrinsically more sensitive than outbred stocks (Kacew, 1996).

4.2.3. Increasing sample size

Other things being equal, increasing the group size will increase power. However, this also increases costs and ethical concerns. Moreover, the relationship between power and sample size is not linear. Increasing sample size in a small experiment produces a good increase in power, but the same increase in an experiment which is already sufficiently large is not worthwhile (see Figure 1).

Toxicity tests with single compounds (e.g. following OECD TG 408) usually involve three dose levels and a control and considerable importance is attached to obtaining a clear dose-response curve. However, if the effects of the treatment are at the limits of detection, as may be expected in case of whole foods, maximising the number of control and a highest dose level animals by reducing the number of dose levels will aim to maximise power.

4.2.4. Increasing the significance level

By convention most investigators use a 5 % significance level. If it were to be increased to 10 % any real effects which just failed to reach significance at the 5 % level would now be judged "significant". However, this will also increase the number of false positive results (Type I errors). As toxicity tests usually involve many outcomes, there is in any case a problem with an excess of false positive results. For the assessment, strong emphasis should be placed on the biological relevance of any observed differences whether or not they reach the chosen level of statistical significance. This is best done by looking at the point and interval (e.g. confidence) estimates and not by just focussing on the P-value.

4.2.5. Determination of sample size

The sample size in clinical trials is usually determined using "power analysis". There is a mathematical relationship between the variables discussed below, such that if five of them are specified, it is possible to determine the sixth. It is assumed here that there are only two means and that they are to be compared using a two sample t-test. These variables are:

1. The effect (the "signal"). This is the magnitude of any difference between the means of the treated and control groups judged to be of scientific interest.
2. The significance level. This is usually set at 0.05 (5 %), although this is entirely arbitrary.
3. The sidedness of the test (or nature of the alternative hypothesis) A two-sided test is specified if a change in either direction from the control would be of interest. Otherwise a one-sided test would be used.
4. The standard deviation (the "noise"). As the experiment has not yet been done, this has to be estimated from a previous experiment.
5. The power of the experiment. This is the probability of being able to detect the specified effect size (signal). Somewhat arbitrarily this is usually set to 80-90 %. The higher value might be appropriate if failure to detect a biologically important effect could have serious consequences.
6. The sample size. This is what is usually determined when planning clinical trials. However, if resources are limited the power analysis may be used to determine the power of an experiment for a specified sample size, or the size of effect likely to be detected if both power and sample size are specified.

These six factors can be combined as shown in Figure 1, which shows sample size in experimental units (number in each group) as a function of the Standardised effect size (i.e. difference between treatment groups divided by its SD) This is also known as the signal/noise ratio).

In a 90-day toxicity test the number of animals in each treatment group is usually 20 (pooling across the 10 males and 10 females), so from the graphs in Figure 1 there would be about an 80 % chance of detecting an effect of 0.9 standard deviations and a 90 % chance of being able to detect an standardised effect size of 1.1 SD difference between the means, with the assumptions given above and assuming no sex by treatment interaction.

Although power can be increased by increasing sample size, substantially larger numbers of animals are needed to detect signal/noise ratios of much less than one. However, power can also be increased by reducing inter-individual variation more effectively. This can be done by choosing animals which are phenotypically and genetically uniform, and by controlling their environment. Group housing, for example, may reduce stress which often increases inter-individual variability.

The use of a power analysis to determine sample size when there are many outcomes (parameters), as in a toxicity test, presents problems. If the most important outcome could be identified, sample size could be determined for that parameter, but it may be sub-optimal for other parameters. Alternatively an average sample size will need to be determined among a large number of parameters. In many cases it will be challenging to specify how large a difference between the treated and control means for each

parameter measured is likely to be important, and there may be no accurate estimates of the standard deviation of each parameter. An alternative approach, taken here, is to design the experiments so that they have a good chance of detecting a standardised effect size of about 1 SD or slightly less assuming that there are no sex by treatment interactions. Large group sizes would be required to detect standardised effect sizes much smaller than about 0.8 and effect sizes of 1 SD or less may not be of much biological relevance. It is up to the applicant to demonstrate the validity of the testing method, including sample size determination.

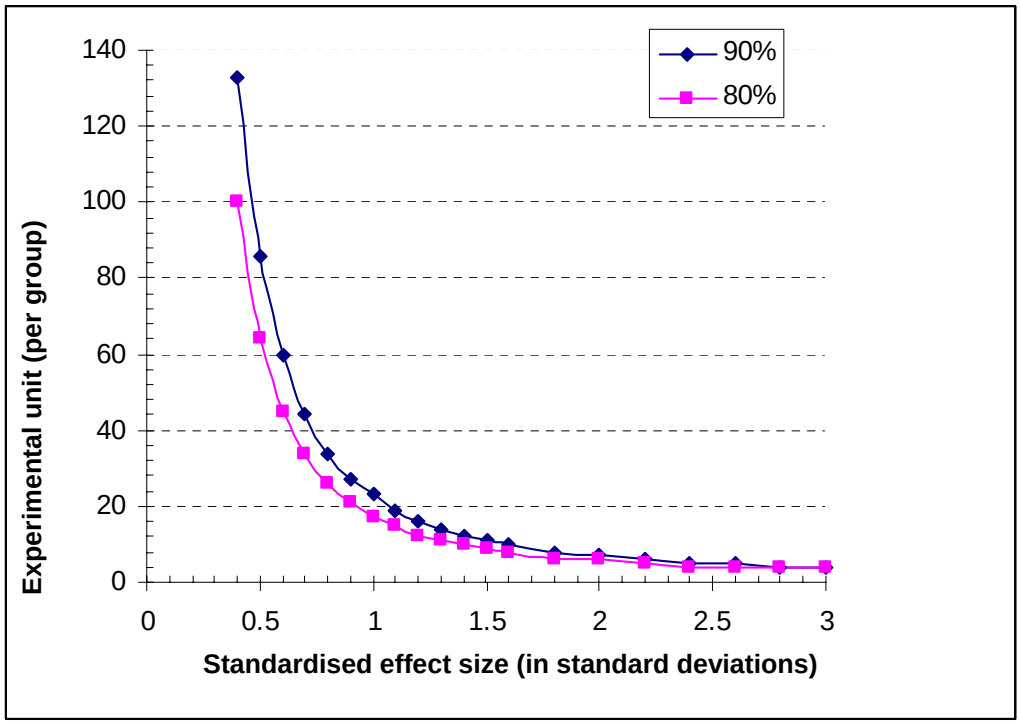

Figure 1: Number of experimental units needed per group as a function of Standardised Effect Size for an 80 % and 90 % power and 5 % significance level using a two-sided t-test. This will approximate the situation in a 2 (treatments) x 2 (sexes) factorial design.

For comparison, rats grow by about 0.8 of a standard deviation per day from 21-35 days of age (data averaged over seven strains of rats and both sexes, 20 rats per group). The suggested experimental designs shown below would be sufficiently sensitive to be able to pick up changes equivalent to slightly more than one day of growth in rats or in any other parameter which changes this amount in terms of standard deviations.

4.2.6. The resource equation (RE)

The "Resource equation" is an alternative way of determining sample size for quantitative (measurement) parameters (Mead, 1988) . It depends on the law of diminishing returns. If one extra ExpU is added to a very small experiment it will provide a useful amount of information. However, if the experiment is already large, then it will make little difference. Mead suggested that E, the error degrees of freedom in an analysis of variance should be between about 10 and 20 although for some more variable outcomes this could be extended to 30 or more. For a completely randomised design E is the total number of ExpU minus the number of groups.

The RE method can be used for complex experiments and those with multiple end points such as toxicity tests. It does not require separate calculations for each endpoint nor an estimate of the standard deviation. It is somewhat more objective than the power analysis because an estimate of the effect size of scientific interest is not required. As OECD TG 408 uses a fixed sample size of 80

animals, it effectively uses the resource equation method. If the animals are housed in pairs, then there are 40 ExpU and if it was regarded as a single experiment (i.e. including both males and females) with four dose levels then E would be 40-8 = 32. Although this is larger than the RE method would suggest, it can be justified on the grounds that it is important not to have to repeat the experiment if the results are equivocal. The method is only appropriate for measurement parameters and OECD TG 408 is also concerned with discrete, often binary, parameters such as presence or absence of a pathological lesion. Such parameters require larger sample sizes so this is an additional justification for the larger sample size.

4.3. A wide range of applicability

There are many factors that can influence the outcome of an experiment. For example, a toxic effect may be seen in one sex but not in the other, a response may only be seen under one set of environmental conditions, or with particular diets or at a certain time. Cox (1958) (and RA Fisher before him) suggests, therefore, that it is important to test the range of applicability of an experiment by incorporating some of these factors using randomised block and factorial designs. This can usually be done without increasing the total number of subjects. Randomised block designs not only increase power, but they also increase generality because each block will sample a slightly different environment.

Factorial designs can be used specifically to increase generality by adding in additional factors which are not them selves of great interest, but which may influence the outcome. Such designs are powerful because they provide extra information at little or no extra cost (Fisher, 1960). For example, OECD TG 408 is a 4 (dose levels) x 2 (sexes) factorial design. In comparing the top dose with the control dose, there are 20 animals in each group. It makes little difference to the power of this comparison whether these 20 are all males, 10 males and 10 females or, say, two animals of 10 strains. If correctly analysed, then it will show whether the response to the treatment depends on the sex, or if different strains were to be used, on the strain.

4.4. Simplicity

Experiments should be simple so as to minimise the chance of making a mistake. They should always be pre-planned and additional groups should not be added during the course of the experiment. Standard operating procedures should be written to cover all the procedures involved such as mixing the diets, administering treatments and collecting data for the analysis.

4.5. Being amenable to a statistical analysis

The statistical analysis should be planned at the same time as the experiment is being designed. It is often a good idea to simulate some of the sort of data which is expected to be used in trial statistical analyses.

5. Experimental designs

5.1. The completely randomised design

By far the majority of experiments involving laboratory animals involve a completely randomised design, i.e. ExpU are assigned to treatment groups at random regardless of any characteristics of the ExpU. These designs are simple and can easily accommodate unequal numbers in each group. Their disadvantage is that if the experiment is relatively large they become difficult to handle without introducing unwanted sources of variability. For example, it may be difficult to obtain 80 rats of

uniform weight and age, house them all under identical conditions and gather data from them all over a short period of time.

5.2. Randomised block designs

Randomised block designs are widely used in agricultural research, but are not always used in research involving laboratory animals. They are quite widely used in in vitro studies, where investigators will often repeat the "experiment" several times. In effect this is a randomised block design with blocks being repeated in time, provided it is analysed correctly.

If an experiment has been done as a randomised block it is possible to calculate its relative efficiency compared with a completely randomised design. Unfortunately these designs are rare in toxicological research and testing so there is no data available to do such calculations.

The use of randomised block designs is recommended, particularly if, for convenience, the experiment needs to be split among different animal rooms or spread over a period of time.

5.3. Split plot designs

These are like a combination of factorial and randomised block designs. Formally, they are a randomised block factorial design in which a main effect is confounded with the block.

The design can best be described by an example. Suppose a factorial design was planned using rodents with four dose levels (control, low, medium, high), and both sexes with animals housed two per cage (the cage being the experimental unit). This would be a 4 (dose levels) x 2 (sexes) factorial design with 8 treatment combinations exactly like an OECD 408 design. If there were to be five-fold replication this would mean the experiment would involve 40 cages. This is a large experiment which might be difficult to manage efficiently. Any uncontrolled time and space-associated variables would increase the inter-individual variation and reduce the power of the experiment. This variation could be reduced by using a randomised block design with five blocks. Each block would then consist of eight cages (4 males, 4 females), one for each treatment combination.

In some cases it would be more convenient to deal separately with the males and females, so an alternative design would be to have five blocks only of females (each with four cages, one for each dose) and another five blocks only with males. This would be a split-plot design. The difference between the sexes will be "confounded" (i.e. mix with) differences between the blocks. However, there will still be a good estimate of whether the two sexes respond differently to the treatments. The advantage of this design would be in convenience and the small block size. The disadvantage is that differences in the means between the two sexes may not be estimates with very high precision. But possibly this does matter because it is already known that males and females differ in many ways.

6. Statistical analysis

Data will normally be accumulated in a spread sheet such as EXCEL. From there it should be read into a suitable high-level statistical package such as SAS, SPSS, MINITAB, R, S+, etc.

The first step is to screen the data for obvious inaccuracies arising from transcription errors. Graphical methods showing individual points such as strip charts are normally used. Box and whisker plots where outliers are shown at the ends of the whiskers are also a convenient preliminary screening tool. Residuals diagnostic plots can also be used.

Outliers which are not transcription errors should not be removed at this stage. Some may disappear if the data needs to be transformed. If not, one approach is to analyse the data with and without the outlier to see if it changes the conclusions. In most cases it will be found to have little effect on the

over-all conclusions. However, if the conclusions depend only on an outlier then further investigation is necessary.

The method of statistical analysis depends on the type of data. Most parameters involve measurements of haematology, clinical chemistry and organ weights. Where possible, these parameters would be analysed using parametric statistical methods such as the analysis of variance and t-tests. Counts and proportions, say of histological data, will need to be analysed using methods appropriate for contingency tables. Growth curves and feed consumption need separate consideration as they involve a series of correlated measurements on each animal, assuming these are measured at weekly intervals.

There are three assumptions underlying a parametric statistical analysis.

1. The observations are independent. This will normally be met by correct identification of the ExpU with appropriate randomisation.
2. The variance is the same in each group (homoskedasticity).
3. The residuals (deviation of each observation from its group mean) have a normal distribution.

These last two assumptions can be checked in a number of ways. One widely used method is to carry out a trial analysis of variance and produce residual model diagnostic plots. These can be used to identify outliers, which should then be checked for accuracy. A plot of fits versus residuals will give a visual indication of whether there is serious heteroskedasticity and a plot of the normal scores will give an indication of whether the residuals have a normal distribution. Most modern statistical textbooks show examples of these plots, with explanations (e.g. Crawley, 2005).

In general, the ANOVA is quite robust against deviations from these assumptions. However, in some cases it is advisable to transform the data. A logarithmic transformation will often correct heteroskedasticity and in many cases outliers will disappear. Other transformations are available. On rare occasions a non-parametric test may be necessary, although where possible this should be avoided as such tests lack power compared with parametric methods, particularly when analysing factorial experiments. A possible non-parametric approach in such cases is to do an analysis of variance using individual rankings.

Once the data is judged suitable, a final analysis of variance is used to assess over-all statistical significance for each trait. This should take account of blocks, gender, treatment and any other factors which are represented in the design. The structure of the analysis of variance for the suggested plans is in Appendix 2.

Most interest will be on the differences between genotypes or doses. Means, standard deviations and 95 % confidence intervals using the pooled estimate of the standard deviation should be presented for each parameter. The number of subjects in each group should be clearly indicated.

If reference groups have been included, then equivalence/non-inferiority testing should be carried out.

Differences between the treated and control groups can be shown graphically for all parameters using standardised effect sizes with confidence intervals. If all responses are expressed in the same standard deviation units then the pattern of response across different parameters is easier to see.

Sex-limited traits (i.e. ones such as testis and uterus weights) which can only be measured in one sex should be analysed using an appropriately reduced analysis of variance. Any statistically significant interactions, particularly those involving treatment and gender should be fully explored using sub-group analyses.

The separate analysis of many parameters, most of which are not expected to differ between treatment groups, may result in a large number of statistical tests. In order to control the number of false positive

1361 results the use of false discovery rate (FDR) methods have been suggested (Kall, 2008), although their
1362 use in the analysis of toxicity tests is not well established. The FDR is the estimated proportion of
1363 false positives among all the significant hypotheses tested. However, this technique is not applicable in
1364 experiments where there are no strongly positive responses. Should there be no real differences
1365 between the groups being compared across many parameters, then all the positive results will be false
1366 positives and the FDR will be 100 %. Therefore, this method is only recommended when there are
1367 some strong and statistically highly significant differences between the groups.

1368 Body weights of each animal should be recorded weekly. A comparison of body weight for each sex,
1369 genotype and dose using an analysis of variance at a few key time points is a simple method for
1370 analysing the results, but it is weak at testing changes in the shapes of the curves and it increases the
1371 number of statistical tests and resulting false positives. The EFSA (EFSA, 2007a, 2007b) used a linear
1372 mixed model with rat as a random factor and gender, dose, genotype and week as fixed effects. With a
1373 randomised block design, block is also a random factor which should be included in the model.

1374 Where there are groups of parameters which are correlated, such as red blood cell parameters, this
1375 should be recorded in the report. Principle Components Analysis (PCA) can be used to reduce the
1376 dimensionality of the data and provide a graphical method of clustering the data (EFSA, 2007a,
1377 2007b). This can be followed by an analysis of variance of the principle components scores for each
1378 individual (Festing et al, 2001).

1379

REFERENCES FOR APPENDIX 1

Cochran WG and Cox GM, 1957. Experimental designs. New York, London: John Wiley & Sons, Inc., 1-611.

Crawley MJ, 2005. Statistics. An introduction using R. Chichester: John Wiley & Sons, Ltd, 1-327.

Cox DR, 1958. Planning experiments. Ed. John Wiley and Sons: New York.

EFSA (European Food Safety Authority), 2007a. Statement on the Analysis of Data from a 90-Day Rat Feeding Study with MON 863 Maize by the Scientific Panel on Genetically Modified Organisms (GMO). European Food Safety Authority, Parma. http://www.efsa.europa.eu/EFSA/efsa_locale-1178620753812_1178621169104.htm

EFSA (European Food Safety Authority), 2007b. EFSA Review of Statistical Analyses conducted for the Assessment of the MON 863 90-Day Rat Feeding Study. European Food Safety Authority, Parma. http://www.efsa.europa.eu/EFSA/efsa_locale-1178620753812_1178621342614.htm

EFSA GMO Panel Working Group on Animal Feeding Trials, 2008. Safety and nutritional assessment of GM plants and derived food/feed and feed: The role of animal feeding trials. Food/feed and Chemical Toxicology 46:S2-S70.

Festing MFW, Diamanti P and Turton JA, 2001. Strain differences in haematological response to chloramphenicol succinate in mice: implications for toxicological research. Food/feed and Chemical Toxicology 39, 375-383.

Fisher RA, 1960. The design of experiments. New York: Hafner Publishing Company, Inc, 1-248.

Hartnell GF, Cromwell GL, Dana GR, Lewis AJ, Baker DH, Bedford MR, Klasing KC, Owens FN, and Wiseman J, 2007. Best Practices for the Conduct of Animal Studies to Evaluate.

ICH Harmonised Tripartite Guideline, 1998, Statistical principles for clilnical trials, E9. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use, pp. 1-39.

Kacew S and Festing MFW. Role of rat strain in the differential sensitivity to pharmaceutical agents and naturally occurring substances. Journal of Toxicology and Environmental Health 1996;47:1-30.

Kall L, Storey JD, MacCoss MJ and Noble WS, 2008. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 7:40-44.

Mead R, 1988. The design of experiments. Cambridge, New York: Cambridge University Press, pp. 1-620.

Montgomery DC, 1984. Design and Analysis of Experiments. Ed. John Wiley & Sons, Inc. New York.

APPENDIX 2 – EXAMPLES OF EXPERIMENTAL PLANS

1. Novel foods

There should be a control and at least two dose levels, the highest dose being the maximum amount of the food which can be incorporated in the diet or given by gavage without distorting its nutritional balance. If the food, e.g. an novel oil, has a nutritional effect, then the control should have a nutritionally equivalent normal ingredient (e.g. a corn oil).

As an example, a randomised block design involving eight identical blocks is shown in Figure 1. The experiment can be split up by block. Differences between blocks are removed in the statistical analysis. Each block consists of three cages of females and three of males, each sex having control, low or high dose levels of the test novel food. Within each block animals should be matched (stratified) for weight and any other attributes such as age and source. Cages should be housed by block (e.g. all block 1 might go on a top shelf) and measurements should be done one block at a time. Each cage contains two animals, giving a total of 6 x 8 x 2 = 96 animals.

Block 1	M Control	F High	F Low	M Low	F Control	M high
Block 2	F Low	F Control	M high	M Control	M Low	F High
Block 3	F Control	M Low	M high	F High	F Low	M Control
Block 4	F High	M Low	F Control	M Control	F Low	M high
Block 5	F High	M Control	F Low	M high	M Low	F Control
Block 6	M Control	M Low	F Control	F Low	M high	F High
Block 7	F Control	M high	F High	M Control	M Low	F Low
Block 8	M Control	M Low	F Control	F Low	M high	F High

Figure 1: Example of a randomised block design as suggested for testing novel foods. Randomisation has been done within each block, so each block has exactly the same treatments but in random order. There are two animals in each of the 48 cages and each treatment mean is based on 16 cages (32 animals, 16 of each sex).

Assuming that there are no treatment by sex interactions (i.e. these would indicate that there is a statistically significant response to the treatment but it differs between the two sexes), there will be 16 cages (32 animals) of each of the three treatments (control, low dose and high dose). With eight blocks and a total of 48 cages (96 animals), the detectable standardised effect size would be 1.02 standard deviations with 80 % power or 1.18 standard deviations with a 90 % power assuming a 5 % significance level and no sex by treatment interaction. Increasing the size of the experiment to 12 blocks (72 cages, 144 animals) would increase the power to detect an estimated effect of 0.82 standard deviations using an 80 % power or 0.96 using a 90 % power with the same assumptions. The detectable effect sizes (in standard deviations) for different numbers of blocks is shown in Table 1.

Note that better control of variation using randomised blocks and two animals per cage will increase the detectable effect size, as measured in standard deviation units, so it will be easier to detect. The layout of the analysis of variance table for novel foods is shown in Table 2.

Table 1: Novel food (six cages/block). Detectable effect size (in standard deviations) in a comparison of the control and the top dose group (one third of cages are controls and one third have the top dose) for an 80 % and 90 % power and a 5 % significance level.

Blocks	No of cages (No ExpU/treatment group)	No of animals	Detectable Effect Size (SDs)	
			80 % power	90 % power
8	48 (16)	96	1.02	1.18
10	60 (20)	120	0.91	1.05
12	72 (24)	144	0.82	0.96

Table 2: Layout of the analysis of variance for the randomised block design.The table shows the source of variation (Blocks, Sexes, Treatments, etc) and the degrees of freedom (DF) assuming that eight blocks are used. The columns for sums of squares, mean squares, F-ratios and p-values are not shown.

Source	Degrees of Freedom (DF)
Blocks	7
Sexes	1
Treatments	2
Control vs Treated	1
Low vs High	1
Sex by Treatment	2
Error	35
Rats/Cages	48
Total	95

For comparison, in a completely randomised design without using blocks, cages are distributed within the animal house at random and any measurements are done in random order. Should it be necessary to split the experiment up, say in time or space, there is no way in which it can be done without increasing the within-group variation and thereby reducing statistical power. This design is not recommended unless there are compelling scientific.

2. GM foods

The proposed design for GM foods involves both sexes, isogenic control food (or feed) at low and high levels and GM food at low and high levels, or a total of eight groups in a 2 x 2 x 2 (sexes x dose x genotype) factorial design. The suggested plan is shown in Figure 2.

Block 1	F GM low	F Ctrl high	M Ctrl high	M GM low	F Ctrl low	F GM high	M GM high	M Ctrl low
Block 2	M GM low	F GM low	F Ctrl high	M Ctrl high	M GM high	M Ctrl low	F Ctrl low	F GM high
Block 3	F Ctrl low	M Ctrl high	M GM low	M Ctrl low	F GM low	F Ctrl high	F GM high	M GM high
Block 4	M Ctrl high	F Ctrl high	F Ctrl low	M GM low	F GM low	M GM high	F GM high	M Ctrl low
Block 5	M Ctrl high	F Ctrl high	F GM low	F Ctrl low	M GM low	M GM High	F GM high	M Ctrl low
Block 6	M GM high	M Ctrl high	F Ctrl high	F Ctrl low	M GM low	F GM high	M Ctrl low	F GM low

Figure 2: A randomised layout for an experiment involving six blocks of eight cages each containing two animals. With this plan half the cages (i.e. 24) will receive GM feed and half the control feed.

This plan involves six blocks of eight cages, giving 48 cages total and 96 animals. Half of the cages (24) will receive the isogenic control food and half (24) the GM variety (12 cages each of the low and high levels). Note that the layout should be re-randomised for each experiment. The effect of increasing the number of cages is indicated in Table 3. For example, increasing the number of cages

from 48 to 72 would make it possible to detect an effect size of 0.67 standard deviations rather than 0.83 with 48 cages, with a power of 80 %, or a slightly higher effect size for a 90 % power, again assuming no sex by treatment interaction

Table 3: Genetically modified food (eight cages/block). Detectable effect size (standard deviations) for a comparison of the control and the group receiving the GM feed averaged across both dose levels for an 80 % and 90 % power and a 5 % significance level.

Blocks	No of cages (No ExpU/treatment group)	No of animals	Detectable Effect Size (SDs)	
			80 % power	90 % power
6	48 (12)	96	0.83	0.96
7	56 (14)	112	0.76	0.88
8	64 (16)	128	0.71	0.82
9	72 (18)	144	0.67	0.77

The layout of the analysis of variance for the design in Figure 2 is shown in Table 4. An alternative would be to use only two dose levels, as was done with the Monsanto MON863 study (EFSA, 2007a, 2007b). In that case there would be eight instead of twelve treatment combinations and a block size of eight could be used with six blocks, making the same total of 48 cages.

Table 4: Layout of the Analysis of variance for the plan for testing GM foods. This shows the source of variation (Blocks, Sexes, Treatments, etc) and the degrees of freedom (DF) assuming that eight blocks are used. The columns for sums of squares, mean squares, F-ratios and p-values are not shown.

Source	Degrees of Freedom (DF)
Blocks	5
Sexes	1
Genotypes	1
Doses	1
Sex x genotypes	1
Sex x doses	1
Genotypes x doses	1
Sex x genotypes x doses	2
Error	35
Total (cage stratum)	47
Cages	48
Total	95

APPENDIX 3 – STUDY REPORT TEMPLATE

The study report should be a complete and easy to review report which clearly presents the aim of the study, the study design and developed protocol, methods used, results obtained, discussion of results and provide a clear description of the conduct of the study and any deviations from the developed study protocol.

The following titles and appendices should be considered to be included in the study report:

- Title page
- Synopsis
- List of abbreviations and definition of terms
- Ethics
- Investigators and study administrative structure
- Introduction
- Study objectives and hypothesis
- Brief description of any pilot studies (if performed)
- Investigational plan
    - Description of overall study design and plan
    - Discussion of study design, including choice of control groups/reference groups
    - Selection of study population
- Treatments (i.e. diets)
    - Treatments administered
    - Identity of test substance (origin, physical nature, purity, contaminants, nutritional information etc.)
    - Method of assigning animals to treatment groups (randomisation)
    - Selection of doses in the study
    - Administration of dose and justification for choice of administration
    - Actual doses (mg/kg bw/day), conversion factor from diet/drinking water
    - Details of diet and water quality
    - Blinding
    - Treatment compliance
- Test animals
    - Species and strains used
    - Health status, results of microbiological screening
    - Number, age and sex of animals
    - Source, housing conditions, etc.
    - Individual weights of animals at the start of the study
- Data quality assurance
- Statistical methods planned in the protocol and determination of sample size
    - Statistical and analytical plans
    - Determination of sample size
- Changes in the conduct of the study or planned analysis
    - Protocol deviations
- Result evaluation
    - Data sets analysed
    - Measurements of treatment compliance
    - Results and tabulations of individual animal data
    - Analysis of toxicological parameters
    - Statistical/analytical issues
        - Adjustments for covariates
        - Handling of missing data
        - Handling of outliers

- Any data transformations
- Interim analyses and data monitoring
- Multiple comparison/multiplicity
- Examination of subgroups

- Tabulation of individual response data
- Dose, concentration, and relationships to response
- Efficacy conclusions (if relevant)

- Deaths and other notable events
  - Listing and discussion of deaths and other notable events
- Results
  - Body weight and body weight changes
  - Feed consumption, and water consumption
  - Toxic response data by sex and dose level, including signs of toxicity
  - Nature, severity and duration of clinical observations (whether reversible or not);
  - Results of ophthalmological examination;
  - Sensory activity, grip strength and motor activity assessments (when available)
  - Haematological tests;
  - Clinical biochemistry tests;
  - Terminal body weight, organ weights and organ/body weight ratios;
  - Necropsy findings;
  - A detailed description of all histopathological findings;
  - Absorption data if available;
- Discussion and Overall Conclusions

- Tables, figures and graphs referred to but not included in the text
  - Environmental data
  - Response data
- Reference list

- Appendices
  - Study information
    - Protocol and protocol amendments
    - List and description of investigators and other important participants in the study, including brief (1 page) CVs or equivalent summaries of training and experience relevant to the performance of the study
    - Signatures of principal or coordinating investigator(s) or sponsor's responsible officer
    - Listing of animals receiving treatment from specific batches, where more than one batch was used
    - Randomisation scheme and codes
    - Audit certificates (if available)
    - Documentation of statistical methods
    - Documentation of inter-laboratory standardisation methods and quality assurance procedures if used
    - Publications based on the study
  - Animal data listings
    - Early terminated animals
  - Protocol deviations
  - Animals excluded from the analysis
  - Adverse event listings (each animal)
  - Listing of individual laboratory measurements by animal.

APPENDIX 4 – STATISTICAL OUTPUTS

Table 1: Summary Statistics for VAR (units) and the change from baseline (day 0) by treatment group and day

Variable Day		Control N = XX	Feed Group: Low (xx g/day) N = XX		Feed group: High (xx g/day) N = XX	
				Difference from Control		Difference from Control
VAR (units)	n	xx	xx	xx	xx	xx
Day 0	Mean (s.d)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Median	xx	xx	xx	xx	xx
	Q1 – Q3	xx – xx	xx – xx	xx – xx	xx – xx	xx – xx
	Min – Max	xx - xx	xx - xx	xx - xx	xx - xx	xx - xx
Day 15	n	xx	xx	xx	xx	xx
	Mean (s.d)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Median	xx	xx	xx	xx	xx
	Q1 – Q3	xx – xx	xx – xx	xx – xx	xx – xx	xx – xx
	Min – Max	xx - xx	xx - xx	xx - xx	xx - xx	xx - xx
...	...	...	...	...	...	...
Day 90	n	xx	xx	xx	xx	xx
	Mean (s.d)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
	Median	xx	xx	xx	xx	xx
	Q1 – Q3	xx – xx	xx – xx	xx – xx	xx – xx	xx – xx
	Min – Max	xx - xx	xx - xx	xx - xx	xx - xx	xx - xx

Note that "N = xx" is the total number of animals randomised in the respective treatment group and "n" is the number of observations available for that day. Produced on DD MMM YYY at HH:MM by PROGRAME.NAME

Table 2: Point estimate and 95% confidence interval by variable and treatment group in the original units (as standardised effect size)

Variable	Control N = XX		Feed Group: Low N = XX		Feed group: High N = XX	
	Estimate[1]	95% C.I.	Estimate[1]	95% C.I.	Estimate[1]	95% C.I.
Variable 1 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 2 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 3 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 4 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 5 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 6 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 7 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 9 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 10 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 11 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 12 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
...	...	...	...	...	...	...

1 The point estimate and 95% confidence intervals we derived from a linear mixed model with [VARS] as covariates and [VAR] as a random effect. Produced on DD MMM YYY at HH:MM by PROGRAME.NAME

Table 3: Point estimate for the difference from control and 95% confidence interval by variable and treatment group in the original units (as standardised effect size)

Variable	Feed Group: Low N = XX		Feed group: High N = XX	
	Estimate[1]	95% C.I.	Estimate[1]	95% C.I.
Variable 1 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 2 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 3 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 4 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 5 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 6 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 7 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 9 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 10 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 11 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
Variable 12 (units)	xx.x	(xx.x, xx.x)	xx.x	(xx.x, xx.x)
...	...	...	...	...

1 The point estimate and 95% confidence intervals we derived from a linear mixed model with [VARS] as covariates and [VAR] as a random effect. Produced on DD MMM YYY at HH:MM by PROGRAME.NAME

1597 **GLOSSARY AND ABBREVIATIONS**

Term	Description
Dose (OECD)	The amount of test substance administered. Dose is expressed as weight (g, mg) or as weight of test substance per unit body weight of test animal (e.g., mg/kg bw), or as constant dietary concentrations (ppm).
Dosage (OECD)	A general term comprising of dose, its frequency and the duration of dosing.
ExpU	Experimental unit(s). The smallest division of the experimental material such that any two ExpU can receive different treatments.
NOAEL	No observed adverse effect level. The highest dose level where no adverse treatment-related findings are observed.

1598