Project Number 289706
Start date of the project: 01/12/2011, duration: 54 months

Deliverable 9.6
(including Milestone 20: Proposals to guide further development of ERA guidance)

Environmental Risk Assessment of Genetically Modified Organisms: Comments and draft text for statistical aspects in ERA guidance

Hilko van der Voet[1], Paul W. Goedhart[1], Antoine Messéan[2], Salvatore Arpaia[3]

[1] DLO: Wageningen University and Research centre,
Plant Research International, Biometris, Wageningen, Netherlands
http://www.biometris.nl

[2] INRA: Institut National de la Recherche Agronomique, Unité Eco-Innov, Thiverval-Grignon, France
http://www.inra.fr

[3] ENEA: National Agency for New Technologies, Energy and Sustainable Economic Development, Centro Ricerche Trisaia, Rotondella, Italy
http://www.trisaia.enea.it

June 2016

Dissemination Level: Public

Contents

Abstract

One of the objectives of the EU-funded project Assessing and Monitoring the Impacts of Genetically modified (GM) plants on Agro-ecosystems (AMIGA) is to provide protocols with a set of evaluated, standardized methods for environmental risk assessment (ERA). This report provides such protocols for the statistical aspects of field trials to assess GM crop effects on non-target organisms (NTOs). Part of these protocols may also apply to other ERA field trials.

The protocols have been tried out on the design and analysis of the AMIGA field trials with maize and potato in six countries. Comments and draft text are provided for a possible update of the EFSA Guidance on the ERA of GM plants.

It is suggested to prepare NTO field trials in a more standardised way by following a checklist for the experimental design. Crucial elements are a clear formulation of research questions and the listing of endpoints to be evaluated together with pragmatic and if needed possibly provisional trigger values (limits of concern) for further evaluation. If feasible, limits of concern may be chosen specific for different receiving environments.

Following the checklist enough information about the endpoints is gathered to allow a prospective power analysis to evaluate the balance between costs of the intended field trial and the benefit of sufficient statistical power for the equivalence assessment. AMIGA has provided a software tool for the power analysis of field trials. The AMIGA power analysis tool provides data templates and example analysis scripts.

When the ERA trials will cover multiple sites, multiple time-points and/or multiple taxonomical or functional groups, it is advisable to define already at the planning stage a logical tree how assessments of endpoints should be integrated and according to which criteria.

A separate protocol is provided for the statistical equivalence analysis of non-target effects. For the statistical analysis of count data an appropriate generalized linear model with over-dispersed Poisson (OP) variation and a logarithmic link function has been found to be a pragmatic choice to assess the equivalence of the GM and comparator varieties.

It is suggested to present results of statistical analyses as graphical representations of confidence intervals for the ratio of GM and comparator means, together with indications of the limits of concern to allow the interpretation of equivalence.

1 Introduction

1.1 Background

Statistical aspects of environmental risk assessment (ERA) of genetically modified (GM) plants for effects on non-target organisms (NTOs) have been described previously in broad terms by Perry et al. (2009). The EFSA Guidance document on the ERA of GM plants (EFSA 2010) has incorporated this guidance on the design and analysis of field experiments. The AMIGA research project aims at providing more detailed guidance in the form of ***protocols for design and analysis*** (Arpaia et al. 2014). This report provides suggestions for the statistical elements of ERA guidance.

1.2 Relation to work program and overview

This report (D9.6) describes the final results from the research in AMIGA Work Package 9. Preliminary work for single-environment data has been reported in Deliverables D9.1 (Goedhart et al. 2013), D9.2a (van der Voet and Goedhart 2014), D9.2b (Goedhart and van der Voet 2014) and D9.3 (first version of software tool), and in published papers Goedhart et al. 2014 and van der Voet and Goedhart 2015. Final AMIGA proposals for the statistical elements of protocols for design and analysis of NTO field trials have been reported in Deliverable D9.4 (van der Voet et al. 2016), accompanied by software for power analysis (Deliverable D9.5, Kruisselbrink et al. 2016). In the current report these results are linked to the EFSA Guidance for environmental risk assessment of genetically modified plants (EFSA 2010).

The structure of this report is as follows. In Chapter 2 comments are given mainly with regard to two sections in EFSA (2010):

- section 3.4 on Interactions of the GM plant with non-target organisms,
- section 2.3.3 on General statistical principles .

Chapter 3 provides suggested text for a protocol for experimental design of NTO field trials, and Chapter 4 provides suggested text for a protocol for statistical analysis.

2 Comments and suggestions regarding NTO field trials

In this document GD refers to the EFSA Guidance Document (EFSA 2010). AMIGA comments and suggestions regarding statistical aspects are mainly directed at two sections, GD 3.4 on GM crop effects on NTOs, and GD 2.3.3 on general statistical principles.

2.1 GM crop effects on NTOs (GD 3.4)

GD section 3.4 provides guidance for interactions of the GM plant with non-target organisms (NTOs). According to the general strategy outlined in the GD, section 3.4 describes the six general steps of ERA: 1) Problem formulation including hazard identification; 2) Hazard characterisation; 3) Exposure characterisation; 4) Risk characterisation; 5) Risk management strategies; 6) Overall risk evaluation and conclusions. General aspects of these six steps are described in the GD, sections 2.2.1-6. Statistical aspects are mainly relevant for steps 1 and 2, as is described below.

2.1.1 NTOs: Problem formulation and hazard identification (GD 2.2.1 and 3.4.1)

GD 3.4.1.1 identifies adverse effects on NTOs as one of the environmental concerns, and sustainable land use as a relevant environmental protection goal. According to GD 2.2.1 environmental protection goals, which are general concepts as set out by EU legislation, are to be translated into measurable endpoints. To derive testable hypotheses the assessment should be performed on measurement endpoints such as measurements of mortality, reproduction or abundance. In a typical NTO field study the list of measurement endpoints may be very long.

Comment:

- Part of the problem formulation should be to organise the measurement endpoints hierarchically in a tree-like structure. This has the advantage that the assessment can be subdivided in more manageable parts. Examples have been given in D9.4.

For the measurement endpoints, GD 2.2.1 asks for defining limits of concern that identify minimal levels of difference between the GM plant and the conventional counterpart that *may* lead to harm: 'For field studies, the limits of concern will reflect [...] the minimum effect that is considered to potentially lead to harm'.

Comment:

- It is noted that limits of concern are defined differently in the Glossary of the GD, namely as 'the minimum ecological effects [...] that are deemed of sufficient magnitude to cause harm'. For a workable procedure the definition in the Glossary should be adapted to conform to the definition of GD 2.2.1 (and other places in the GD).
- One of the issues with 'minimum ecological effects that are deemed to cause harm' is that their value is likely to depend on the specific receiving environment (RE) under consideration. The GD has not discussed this as such. Limits of concern should preferably be chosen specifically for each type of RE. When this is difficult to do a priori because of lack of knowledge, it might be easier to set a general value for limits of concern defining 'potential harm' than for limits of concern defining 'actual harm'. In such cases, these general limits of concern are indicative values for the interpretation of a set of NTO field trials as triggers for

further investigation. It then remains necessary to put the observed results in the context of the specific agroecosystems for which the ERA is performed.

GD 3.4.1.2 provides a scheme for selecting focal NTO species to be practically tested. It consists of four steps: 1) identification of functional groups exposed to the GMO; 2) categorization of NTO species; 3) ranking of species per functional group; 4) final selection based on practical criteria.

Comments:

- In practice, NTOs are not always identified at the species level, but for example only at the genus, family or order level. Therefore, for cases where a taxonomic endpoint is to be defined it is suggested to replace 'species', with 'taxon' to allow case-by-case selection of the most appropriate level.
- The classification of taxa in functional groups is a good example of hierarchical ordering of measurement endpoints.
- The selection of focal NTO taxa is a necessity in laboratory and semi-field studies, where animals are manipulated. However, it should be better described in the text that the situation is different in field studies, where the NTOs are just observed. In field studies where the whole NTO community is considered relevant, it may therefore be an option to bypass the selection of focal taxa, thus avoiding the inevitable uncertainties of a selection.

GD 3.4.1.5 makes a distinction between specific hypotheses, related to intentional alterations in the compounds produced by the GM plant, and generic hypotheses, when only the GM plant composition is altered.

Comments:

- Hypotheses have to be specified in both cases, so are specific. It may be preferable to rephrase the terminology to 'hypotheses about specific endpoints' and 'hypotheses about a generic class of endpoints'.
- In the case of NTOs it may be confusing to speak of intended effects. Probably what is meant here is 'intended effect on a target endpoint'. However, the hypotheses to be tested in GD 3.4 is about unintended effect on a selection of NTO taxa. The main difference between 'hypotheses about specific endpoints' and 'hypotheses about a generic class of endpoints' may then be that in the former case some (or many) NTOs do not have to be considered because effects are have been already excluded based on previous tests or knowledge from the literature. The statistical methodology to be applied to whatever is selected as endpoints is the same.

2.1.2 NTOs: Hazard characterization (GD 2.2.2 and 3.4.2)

GD 3.4.2.1 discusses laboratory studies and GD 3.4.2.2 gives a discussion of field trials. The advantage of field trials is that they allow to observe potentially different effects in different environments (here termed trait x environment interaction).

For design and analysis field trials reference is made to GD 2.3.3, so we will further discuss these issues there.

Six objectives for field trials are mentioned: 1) a. to identify exposure routes and b. to confirm observed effects in laboratory or semi-field studies; 2) to discover unintended effects; 3) to provide feedback for further testing; 4) to study food chain effects; 5) to determine effects on multiple generations and other spatio/temporal interactions; 6) to study interactions between NTOs.

Comments:

- Objective 2 seems the common case, where the assessment of field trial data is requested without prior information. For this equivalence tests are advised as the primary method of assessment.
- The first mentioned objective consists of two, not necessarily related sub-objectives, say 1a and 1b. Objective 1b leads to hypotheses for specific effects to be tested. The existence of such effects can be tested in a difference test, the potential biological relevance in an equivalence test.
- For the other objectives (1a, 3, 4, 5, 6) detailed modelling would be necessary to formulate the hypotheses that have to be tested with the field trial data. This remains open for further work.

2.2 General statistical principles (GD 2.3.3)

The introductory section of GD 2.3.3 mentions the need to list the research questions and re-state them in formal terms in the form of precise null hypotheses for testing

Comments:

- We suggest an approach that starts with calculating confidence intervals for the effects of the endpoints, and presenting these graphically.
- Testing is then easily performed by comparing the confidence intervals to effect sizes representing no effect (for difference tests) and the limits of concern (for equivalence tests).

It is stated that ‘for field trials, applicants shall provide a clear and explicit statement concerning the minimum levels of abundance acceptable for each taxa sampled, below which results would lack credibility’.

Comment:

- At low levels of abundance the coefficient of variation becomes very large, and estimates of abundance ratios, such as are used in a comparative approach, will be imprecise. Whereas restricting the data to taxa with a minimum level of abundance is one possibility, the AMIGA project is proposing an alternative approach in which no taxa are omitted from the results, but limits of concern are adjusted for taxa with low abundances. See 2.2.2 for more details.

In this section equivalence testing is introduced. Limits of concern are introduced as follows: ‘For lower-tier studies [...] the limits of concern will usually be trigger values which, if exceeded, will usually lead to further studies at higher tiers. […] For higher tier studies, especially field studies, the limits of concern shall reflect more directly the minimum ecological effects (in positive and negative directions) that are deemed biologically relevant. For field studies, at least one of the limits of concern shall represent the minimum effect that is considered by applicants potentially to lead to environmental harm’

Comment:

- In the second quoted sentence it remain unclear what 'deemed biologically relevant' should mean. It might be misinterpreted to mean that effects outside the limits of concern would be necessarily harmful. However, the third quoted sentence clarifies that such effects is only indicating *potential* harm, and that therefore the limits of concern in field studies are only trigger values for further investigation. In this respect there is no difference with the meaning of the term in lower-tier and higher-tier studies, although of course different values may be used.
- It can be noted again that the definition of limit of concern in the Glossary of the GD should be amended.
- In this context the work of an ongoing EFSA working group for developing guidance for the identification of biological relevance of adverse positive health effects from experimental & human studies (EFSA 2011) is of interest. Whereas the importance of distinguishing between statistical significance and biological reference is recognised, the published document admits 'that it may be difficult to define the size of effect that will be considered biologically relevant or important in every situation'. No solution for such situation is provided. Here we suggest that defining effect sizes that will be *potentially* biologically relevant may be easier to do.

2.2.1 Difference and equivalence testing (GD 2.3.3.1)

In the GD difference and equivalence testing are discussed in a rather symmetrical way, and it is stated that the two approaches are complementary.

Comments:

- More specific guidance can be given here:
    - For confirmation of specific effects previously found in lower tiers or suggested by modelling, the first relevant test is the difference test, testing a classical null hypothesis of equality against an alternative hypothesis of non-equality. In addition, the equivalence may be used to assess biological relevance.
    - For generic testing whether non-target effects of genetic modifications are small enough to be accepted unless further arguments are made, the most relevant test is the equivalence test which seeks to assert that differences or ratios are within given limits.
- A graphical representation using confidence intervals allows both tests to be performed, yet visual elements such as colouring of points, lines and/or background can be used to emphasise either the differences or the non-equivalences. Examples can be found in D9.4.

2.2.2 Effect size and limit of concern (GD 2.3.3.2)

GD 2.3.3.2 re-iterates that risk characterization cannot be done without relating effects to potential harm. Therefore effect sizes should be specified which are considered to potentially have a relevant impact on the receiving environment.

Comments:

- The word 'potential' is crucial here, there need not be any evidence that harm really occurs at the specified effect size.
- Effect sizes for count data are usually expressed as ratios between abundances for the GMO and the comparator. A problem is that these estimates become highly variable when the

abundances are low. The possibility to obtain very low or very high ratios for taxa with a low abundance is a basic statistical fact, and therefore cannot be a reason for concern. To allow for this fact in a flexible manner, AMIGA has proposed a default scaling of limits for concern for endpoints with a low abundance (see 2.2.5 and D9.4 for details). Effectively this will make limits wider at abundances below a given level. Note, that ecological arguments may always be used to override default scaling choices and request other limits.

The GD distinguishes effect sizes (mainly seen as a prerequisite for power analysis) from limits of concern (considered as null hypothesis values for equivalence tests), but admits that 'usually, these quantities will be identical'.

Comment:

- The GD does not give any example where it would be logical to use different values for effect size and limit of concern. Therefore, for the purpose of equivalence testing as described in this document, there is no need to distinguish the two concepts. Future work may elaborate on models for context-specific limits of ecological harm where the variation of limits across receiving environments would be modelled. It is suggested that future guidance should then be adapted with practical procedures how to work differently with effect sizes and limits of concern in such context-specific approaches.

2.2.3 Power analysis (GD 2.3.3.3)

GD 2.3.3.3 asks for each study to perform a prospective power analysis to ensure that the design is such that the difference test has sufficient statistical power to provide reasonable evidence. The power analysis should relate to a single site, not the entire set of trials. However, when many species are sampled, the power analysis is required only for species of prime importance and those expected to be the most abundant.

Comments:

- Commonly used methods for power analysis consider difference tests and are based on the normal distribution.
- The AMIGA project has produced a software tool for power analysis of field trials (D9.5, Kruisselbrink et al. 2016), both for difference tests and equivalence tests, and allowing different measurement types, notably count data and nonnegative continuous measurements.
- Power analysis for a difference test considers the statistical power when in reality the effect size is equal to a specified value, for which we may take the limit of concern. Power analysis for an equivalence test considers the statistical power when in reality the effect size is equal to a specified value, for which we may take zero if the absence of the effect is a realistic option.
- The focus in the GD on power analysis for the difference test may have been caused by the novelty and non-availability of software for power analysis for the equivalence test. However, for generic testing whether non-target effects of genetic modifications are small enough to be accepted unless further arguments are made, equivalence testing is the most relevant method, and therefore the corresponding type of power analysis should be preferred.
- It is recognised that power analysis for a design at a single site is useful if the results are to be interpreted for each site separately. However, when results of an ERA are to be interpreted across sites or across multi-environments, there may be an additional need to

consider the question whether sufficient sites have been included in a multi-environment study. D9.4 provides an example of a simplified power analysis for this question.

- When many taxa are sampled, it may be difficult to demonstrate enough power for taxa with a low expected abundance if fixed limits of concern are used. This explains why the GD allows to restrict the power analysis to taxa of prime importance and taxa with a high abundance. An alternative approach is to allow for wider limits of concern for taxa with a low abundance (see 2.2.5 and D9.4 for details).

2.2.4 Experimental design (GD 2.3.3.5)

GD 2.3.3.5 discusses the relative merits of laboratory and field studies, and gives detailed prescriptions for the latter, also in relation to genotype by environment (GxE) interaction. It is stated that 'for field trials, the principle shall be followed that each field trial at a site on a particular occasion shall have sufficient replication to be able to yield a stand-alone analysis if required, although the main analysis shall derive inferences from averages over the complete set of field trials at all sites and years.'

- The GD does not choose between a per-site analysis and an overall analysis, but requires both types to be possible. This may be related to the unclear position of risk management in Europe, which also seems undecided between requiring a risk assessment for each country or on average. Further input from the EU would be needed to make progress. Anyhow, from an ecological perspective, for cultivation dossiers there is no point to consider the EU as a single environment.
- In the AMIGA project it has been considered to quantify GxE. However, the number of available site/year combinations was too low for statistically relevant analyses. Therefore it was decided to analyse the data site by site (possibly integrating over the years at each site within each analysis), and to integrate equivalence results over sites (see 2.2.5). This hierarchical analysis may be the most practical way to address GxE across different biogeographical regions in Europe.

2.2.5 Statistical analysis and reporting (GD 2.3.3.6 and 2.3.3.7)

GD 2.3.3.6 asks applicants to prepare protocols for the experimental design and the statistical analysis of proposed studies. For many measurement endpoint response variables a logarithmic transformation or a generalized linear model (GLM) with a logarithmic link function is recommended, allowing differences between GM plants and comparators to be interpreted as ratios on the natural scale. A graphical representation of 90% confidence intervals for effects should be used as a basis for difference and equivalence testing.

Comments:

- AMIGA has elaborated on the suggested protocol elements in the GD, and the resulting protocols are presented in Chapters 3 and 4 of this report. Motivations can be read in D9.4.
- AMIGA has performed a simulation study regarding statistical methods for count data (see D9.2b). A general conclusion is that the Overdispersed Poisson (OP) model, which is a GLM with logarithmic link function, is the preferable analysis method for confidence intervals and judging these against limits of concern in equivalence tests.
- AMIGA has elaborated a procedure for graphical representation, see examples in D9.4.

GD 2.3.3.7 requires a main analysis addressing all field trials simultaneously based on the full dataset from all sites. Sites and years are not necessarily entered as random factors (such as would be the logical choice in statistical mixed models), but may also be entered as fixed factors. For each endpoint a statement about presence or absence of GxE interactions should be made.

Comments:

- The approach proposed in the GD (simultaneous analysis based on all sites) may be impossible in practice, because the set of endpoints may not be the same at all sites. This may be a result from ecological differences or from experimental conditions and measurement possibilities at each site.
- AMIGA proposes more flexibility, and suggests the applicant to prepare a logical tree for analysis already at the time of planning the experiments. Such a tree connects the endpoints at different hierarchical levels by operations that can belong to three categories:
    - Pre-processing the data, e.g. by summing all counts over time points within a year, or even over multiple years at the same experimental units,
    - statistical analysis, thus estimating effects at the same level as the data or at a higher level (e.g. statistical hierarchical analysis or meta-analysis), and
    - equivalence analysis, which integrates estimated effects in relation to their limits of concern to higher levels.
- Examples of the proposed approaches are given in D9.4.
- Based on the limited number of sites and years for NTO field studies, it may be unjustified to derive firm statements on the presence or absence of GxE interaction. Instead of this, the logical tree for analysis will specify whether equivalence analysis steps will average equivalence conclusions over e.g. environments, or will require equivalence to be met at each environment individually.

3 Protocol and software for the statistical aspects of designing experimental studies of non-target effects

Attention is required before a field trial is performed to ensure that the experiment will be meaningful to answer research questions. We present relevant points from a statistical viewpoint as a checklist.

3.1 Checklist for experimental design and power analysis

1. Describe the **questions** the experiment is meant to answer, in words.
2. Decide on the **site(s)** and the **year(s)/season(s)** for the experiment based on the questions to be answered and feasibility.
3. Decide on the **spatio-temporal and/or taxonomic integration levels** at which conclusions regarding the research questions have to be formulated.
4. Prepare the **list of endpoints**. This may typically be organised in a hierarchy of endpoints. At the lower levels endpoints may be open to the possibilities of observation in the field (e.g. it may be impossible to predict if individual arthropod taxa will or will not be present under the conditions of the experiment).
5. Construct a **logical tree for the analysis** of all observed endpoints, showing how *data* may be pre-processed (**data pre-processing steps**), how *effects* will be estimated from the data by statistical analysis (**statistical analysis steps**), and how conclusions on *equivalence* will

follow from the collection of effects and the limits of concern (**equivalence analysis steps**). The branches of the trees may have equal or different schemes for the subtrees. In general, many different trees will be possible; the chosen tree should therefore be motivated.

6. For count or fraction data, a typical way of **pre-processing** the data is to **sum** over primary levels, e.g. over individual time points to obtain year totals, or over individual taxa to obtain totals for functional groups.
7. Indicate the **nature of the statistical analysis steps** in the logical tree as being a statistical analysis (**SA**, where the effects are calculated at the same level as the data), a statistical hierarchical analysis (**SHA**, where the analysed data are at a lower level of integration than the estimated effects) or a statistical meta-analysis (**SMA**, where effect estimates of a previous analysis are integrated to a higher level). More guidance on SA is provided in section 4.2, more guidance on SHA and SMA in section 4.3.
8. Indicate the **nature of the equivalence analysis integration steps** in the logical tree as requiring equivalence conclusion to be valid for all members (**EA_{all}**) or as allowing members to compensate for each other by averaging concern quotients (**EA_{av}**). More guidance is given in section 4.5.
9. For each endpoint to be used in the statistical analysis classify the **measurement type**, e.g. non-negative continuous data, count data or fractions (percentage) data.
10. For each endpoint to be tested formulate the **Limits of Concern (LOCs)**. For each endpoint one lower and/or one upper LOCs can be set. For non-negative continuous and count data these will typically be ratios of GMO divided by CMP true values. The LOCs define the **null hypotheses for equivalence tests**.
11. For endpoints which are counts, decide **how to address low abundance data**. Either define a criterion to omit uninformative low counts (e.g. abundance below 5, or CV larger than 100%), or adopt a rule that LOCs will be set as a function of the expected or observed abundance in the field. A proposed scaling factor for the log(LOC) expressed as a factor is $\sqrt{\mu_0/m}$, where m is the expected or observed abundance in the experiment, and μ_0 is a chosen threshold abundance below which concern limits will be widened (e.g. $\mu_0 = 10$).
12. Set the **significance levels (**α) for statistical testing. Conventionally the level (size) will be 0.05. In the TOST approach to equivalence testing (Schuirmann 1987) the significance level of a two-sided confidence interval is twice the significance level for the equivalence test, therefore employing a two-sided (1-2α) confidence interval corresponds with (1-2α) confidence equivalence tests.
13. Set the **required power** of the equivalence tests to detect a zero differences. Under the null hypothesis the effect sizes will be equal to the LoCs. Conventional values for power are between 70 and 90%.
14. Describe the structure of the proposed **experimental design**, e.g. completely randomized, randomized blocks, split-plot, balanced incomplete blocks.
15. Describe the **experimental units** (typically plots or sub-plots), and give details of the **blocking structure** (e.g. 4 main plots per randomized block, each main plot split into 3 sub-plots) and the **treatment structure** (e.g. three types of spraying and four crop varieties). Also describe if interactions should be included.
16. Describe whether **repeated measurements** will be taken from the same experimental unit.
17. Provide a **model formula** specifying how the data will be analysed, using the syntax of one of the common software tools for statistical analysis (e.g. SAS, GenStat, R), for example *block/plot/subplot + treatment + variety*. Include terms and a correlation structure for repeated measurements if used. Indicate which factors are random rather than fixed.
18. For each primary endpoint provide **prior estimates of central value and variation** for a measurement on a single experimental unit. For non-negative continuous and count data the prior estimates for central values will typically be expected values or geometric means, and the prior estimates for variation will typically be coefficients of variation. Such values can be derived from previous experiments or based on expert knowledge.
19. For each endpoint specify the simplest **statistical analysis method** that will be used (the analysis method may need to be adapted if there are unexpected deviations in the execution of the field study or unexpected data). See the statistical analysis protocol for details.

20. Based on the prior estimates **estimate the power of the proposed design as a function of replication,** for the equivalence test at the chosen integration level(s), see point 3. For this the AMIGA Power Analysis software may be used (Kruisselbrink et al. 2016).
21. From the power curves derive the **replication** of the comparison of GMO to CMP in the proposed design.
22. If the calculated minimal replication cannot be realized in practice, the **power is insufficient**. In such case adapt the design or reformulate the research questions.
23. **Randomise** the treatments over the experimental units taking proper account of the design.

4 Protocol for statistical equivalence analysis of non-target effects

In this chapter we present a protocol for the statistical analysis of data from ERA field trials. In principle, the methods of statistical analysis have already been decided at the time of planning the experiment, but it may be needed to update the methods based on the context or unexpected findings.

4.1 General

1. Check and if necessary update the **list of endpoints** that was established in the design phase. Motivate any change.
2. Check and if necessary update the **logical tree for the analysis** of all observed endpoints. Motivate any change.
   The logical tree for analysis shows how *data* may be pre-processed (**data pre-processing steps**), how *effects* will be estimated from the data by statistical analysis (**statistical analysis steps**), and how conclusions on *equivalence* will follow from the set of all estimated effects and the limits of concern (**equivalence analysis steps**). The branches of the trees may have equal or different schemes for the subtrees. In general, many different trees will be possible; therefore the chosen tree should be motivated.
   a. For count or fraction data, a typical way of **pre-processing** the data is to **sum** over primary levels, e.g. over individual time points to obtain year totals, or over individual taxa to obtain totals for functional groups.
   b. Indicate the **nature of the statistical analysis steps** in the logical tree as being a statistical analysis (**SA**, where the effects are calculated at the same level as the data), a statistical hierarchical analysis (**SHA**, where the analysed data are at a lower level of integration than the estimated effects) or a statistical meta-analysis (**SMA**, where effect estimates of a previous analysis are integrated to a higher level). More guidance on SA is provided in section 4.2, more guidance on SHA and SMA in section 4.3.
   c. Indicate the **nature of the equivalence analysis integration steps** in the logical tree as requiring equivalence conclusion to be valid for all members (**EA_{all}**) or as allowing members to compensate for each other by averaging of concern quotients (**EA_{av}**). More guidance is given in section 4.5.
3. Graphical summaries of results are to be prepared for estimated effects (section 4.4) and, if deemed useful, for LoC-scaled differences (section 4.5).

4.2 Statistical analysis of single endpoints

The basic approach is to calculate estimates and 90% confidence intervals for effects (GMO vs. CMP differences, expressed on an appropriate scale), and then compare these to the (possibly provisional) limits of concern which were specified during the design of the experiment (see Figure 1 for a flowchart for the statistical analysis of a single endpoint).

1. The method of statistical analysis depends on the type of endpoint. For continuous endpoints with necessarily positive values it is recommended to perform an analysis on the log transformed data. For discrete endpoints such as count data and fraction data it is recommended to perform an analysis on the original scale using an appropriate statistical distribution and link function.
2. Analyse the transformed data by linear models: ANOVA if the design is balanced, or by a mixed model (REML) if they are not.
3. Analyse the untransformed data by generalized linear models (GLM), or by a generalized linear mixed models (GLMM) in case there are additional random effects in the model. Allow for over-dispersion in counts and fractions.
4. Check whether statistical assumptions are reasonable, e.g. as follows:

a. Outliers: check data points with large standardised residuals. Compare analyses with and without such data points in a sensitivity analysis.
b. A QQ plot of the residuals should show approximately a straight line
c. A plot of residuals vs. fitted values can be used to check if there is heteroscedasticity.

5. If statistical assumptions are unreasonable, then an ad-hoc strategy will have to be followed. For example, another variance function might be more appropriate or non-parametric tests may be used. This protocol further assumes that the model fits sufficiently well.
6. Extract the estimated difference between the GMO and CMP from the statistical model, e.g. the log-ratio for count data, along with the standard error of the estimate. Employs these to calculate a two-sided 90% interval taking account of the degrees of freedom for residual. Display the confidence interval in a graph along with the LoCs. Note: for visual display it is recommended to calculate and display both limits, even if there is only one LoC, for either a decrease or an increase.

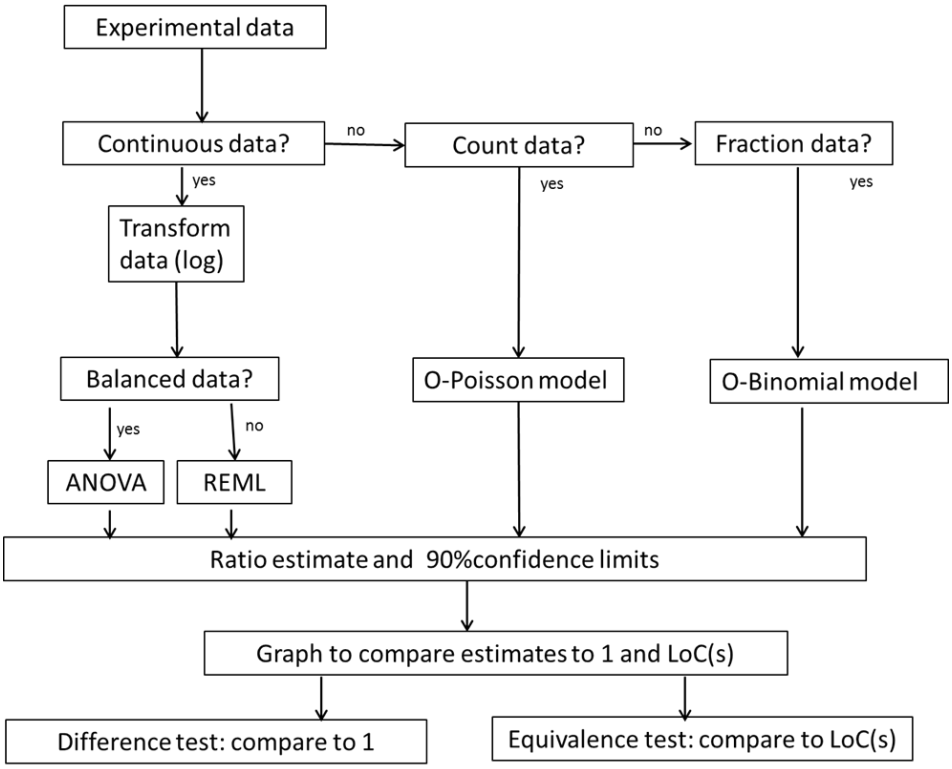

Figure 1. Flow chart for the statistical analysis of single endpoints. ANOVA = analysis of variance, REML = residual maximum likelihood. O-Poisson = over-dispersed Poisson model. O-Binomial = over-dispersed binomial model. LoC = Limit of Concern.

4.3 Statistical analysis integrating multiple endpoints

1. Integration over multiple endpoints may be automatically performed in a **statistical hierarchical analysis** (SHA) model as described in section 4.2. Perform a **statistical meta-analysis** (SMA) if described in the logical tree for analysis. For this, consider the estimated effects with their standard errors (at an appropriate scale, e.g. the log scale) as input for the meta-analysis. Consider the level over which an average is taken as a random factor.
2. From the output, construct an **estimate and a 90% confidence interval for the overall effect**.
3. The use of SHA or SMA is only logical if limits of concern are defined for the integrated output or if they are equal for all individual endpoints.

4.4 Graphical representation of effects

1. For each endpoint, **plot point estimates and 90% confidence intervals of estimated effects**, together with lines for the equality ratio 1, and the LoCs. In most cases plots on a logarithmic scale are advised. The 90% limits of the interval represent a 5% significance level for equivalence testing in a two one-sided tests (TOST) approach.
2. Prepare **one or more graphs**, depending on the number of endpoints, and possible groupings in the hierarchy which are of interest as specified in the logical tree for analysis.
3. Compare the intervals to the LoCs to obtain **conclusions regarding equivalence** of the GMO and the CMP.
4. If of interest, compare the intervals to zero to obtain conclusions regarding the **statistical significance of the difference** between the GMO and the comparator. Note that this implicitly employs a significance level of 10% for a two-sided difference test.
5. Optionally, **confidence intervals can be displayed on the LoC Scaled Difference (LoCsDIF) scale**. This possibly allows an easier comparison in case (scaled) limits of concern are not the same for various endpoints.

4.5 Integration of endpoints, overall equivalence analysis

1. For each estimated effect and its corresponding limit(s) of Concern (LoC) calculate the **concern quotient** (*CQ*):

$$\text{Two-sided:} \quad CQ = \max\left[\frac{\log(Q)}{\log(LoC_{low})}, \frac{\log(Q)}{\log(LoC_{upp})}\right]$$

$$\text{One-sided left:} \quad CQ = \max\left[\frac{\log(Q)}{\log(LoC_{low})}, \quad 0\right]$$

$$\text{One-sided right:} \quad CQ = \max\left[0, \frac{\log(Q)}{\log(LoC_{upp})}\right]$$

   Use the same formulae to convert the confidence limits for log(Q) to the *CQ* scale, using the lower confidence limit Q_{low} in combination with LoC_{low}, and the upper confidence limit Q_{upp} in combination with LoC_{upp}.
2. For each **equivalence analysis (EA) step in the logical tree for the analysis**, check whether the intended equivalence criterion is that 1) all member endpoints should comply to $CQ \leq 1$ (EA_{all}) or 2) *CQ*s can be averaged (EA_{av}).
3. For each **EA_{all} step**, integrate over individual endpoints *i* by $CQ = \max_i(CQ_i)$ and $CQ_{upp} = \max_i(CQ_{upp,i})$.
4. For each **EA_{av} step**, integrate over individual endpoints *i* by $CQ = \text{mean}(CQ_i)$ and $CQ_{upp} = \text{mean}(CQ_{upp,i})$.

References

Arpaia S, Messéan A, Birch NA, Hokkanen H, Härtel S, van Loon J, Lövei GL, Park J, Spreafico H, Squire GR, Steffan-Dewenter I, Tebbe C, van der Voet H (2014). Assessing and monitoring impacts of genetically modified plants on agro-ecosystems: the approach of AMIGA project. Entomologia, 2: 154. http://dx.doi.org/10.4081/entomologia.2014.154.

EFSA (2010). EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. EFSA Journal, 8(11): 1879. [111 pp.], doi:10.2903/j.efsa.2010.1879. https://www.efsa.europa.eu/en/efsajournal/pub/1879.

EFSA (2011). EFSA Scientific Committee. Statistical Significance and Biological Relevance. EFSA Journal 2011;9(9):2372 [17 pp.]. doi:10.2903/j.efsa.2011.2372 https://www.efsa.europa.eu/en/efsajournal/pub/2372.

Goedhart PW, van der Voet H (2014). Environmental Risk Assessment of Genetically Modified Organisms: Simulation study to investigate properties of difference and equivalence tests. Deliverable 9.2b, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.

Goedhart PW, van der Voet H, Baldacchino F, Arpaia S (2013). Environmental Risk Assessment of Genetically Modified Organisms: Overview of field studies, examples of datasets, statistical models and a simulation tool. Deliverable 9.1, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.

Goedhart PW, van der Voet H, Baldacchino F, Arpaia S (2014). A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. Ecology and Evolution, 4: 1267–1283. http://dx.doi.org/10.1002/ece3.1019.

Kruisselbrink JW, Goedhart PW, van der Voet H (2016). Environmental Risk Assessment of Genetically Modified Organisms: Software for power analysis and analysis of data from field studies. Deliverable 9.5, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.

Perry JN, ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F & van der Voet, H (2009). Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environmental Biosafety Research, 8: 65-78. http://dx.doi.org/10.1051/ebr/2009009.

Schuirmann DJ (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6): 657-680. http://dx.doi.org/10.1007/BF01068419.

van der Voet H, Goedhart PW (2014). Environmental Risk Assessment of Genetically Modified Organisms: Statistical aspects of a protocol for single-environment GMO field studies. Deliverable 9.2a, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.

van der Voet H, Goedhart PW (2015). The power of statistical tests using field trial count data of non-target organisms in environmental risk assessment of genetically modified plants. Agricultural and Forest Entomology 17: 164–172. http://dx.doi.org/10.1111/afe.12092.

van der Voet H, Goedhart PW, Kruisselbrink JW (2016). Environmental Risk Assessment of Genetically Modified Organisms: Protocols for statistical aspects of non-target effects field studies. Deliverable 9.4, AMIGA project, project number 289706. Available at http://www.amigaproject.eu/documents/deliverables.