The handle http://hdl.handle.net/1887/22932 holds various files of this Leiden University dissertation.

**Author:** Plischke, Andreas  
**Title:** Non-target effects of GM potato : an eco-metabolomics approach  
**Issue Date:** 2013-12-18
Effect measurement and limits of concern in non-target risk assessment of GM crops

Andreas Plischke¹, Maaike Bruinsma¹,², Peter G.L. Klinkhamer¹, Frans J.A. Jacobs²

¹Ecology and Phytochemistry, Institute of Biology Leiden, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands
²Current address: Naktuinbouw, Team R&D, Sotaweg 22, 2371 GD Roelofarendsveen, The Netherlands

Abstract

A central requirement in assessing the risks of genetically modified (GM) plants towards non-target organisms (NTO) is the definition of threshold values for GM/non-GM differences, the so-called “limits of concern”. Due to a lack of ecological information, these limits of concern are often chosen arbitrarily and set to the same value for all tested species. Here, we show some of the consequences of this current practice for experimental field design as guided by power analysis. Arbitrary limits of concern result in large differences in statistical power between species, although the biological relevance of the assumed thresholds is unclear. Observed variation in field trials may indicate biological baseline variation that could guide the definition of thresholds. Here, we discuss standardized effect sizes as a way of incorporating information about variability in the measurement of effects. Furthermore, we suggest the use of multivariate analyses in non-target risk assessment, because the current approach of measuring effects separately per non-target group lacks a systems biology perspective.

Keywords: insect communities, transgenic plants, power analysis, environmental effects
The safety of genetically modified (GM) crops towards non-target organisms (NTO) is usually assessed by comparing a GM line to its near-isogenic, non-modified counterpart. A central requirement in NTO risk assessment is the definition of threshold values for the difference between the GM and its counterpart in any given measurement endpoint, e.g. the abundance of non-target insect species. These so-called ‘limits of concern’ are defined as the minimum differences of sufficient magnitude to cause ecological harm (EFSA 2010). However, little guidance is currently provided on how limits of concern for non-target organisms should be determined. Ideally, limits of concern should be based on species-specific information about population dynamics and ecological interactions. This approach of establishing ecologically justified limits of concern has been followed for individual species using existing knowledge and mathematical modeling (e.g. O’Callaghan, Soboleva & Barratt 2010). However, such detailed ecological information is rarely available for the wide range of non-target organisms that often is considered in field trials. Therefore, it has become common practice to work with fixed, arbitrary limits of concern for all of the tested species. Typical limits of concern across species are between 30-50% abundance difference between the GM plant and its counterpart (e.g. Prasifka et al. 2008; EFSA 2010; Albajes et al. 2012).

In this study, we examine some of the problems and consequences of the current practice. First, we show that arbitrary limits of concern have implicit consequences for experimental field designs as guided by power analysis. Second, we discuss standardized effect sizes as an alternative way of quantifying effects in relation to biological variation. Finally, we point out the limitation of the current per-endpoint-analysis when effects at the community level may be ecologically as relevant as changes in individual species.

Arbitrary limits of concern and power analysis

Power analysis in non-target risk assessments ensures that experiments generate data that can inform the decision-making process, rather than being interesting from a purely scientific point of view (Romeis, Lawo & Raybould 2009). Statistical power is the probability of correctly rejecting a false null hypothesis of no difference in a statistical test. Statistical power is defined as 1-β, where β is the Type II error rate, or the chance of accepting the null hypothesis while in fact it was false. Calculating the power of a statistical test requires an estimate of the expected variation and the definition of a desired effect size that one wishes to detect. Estimating statistical power before the start of an experiment (prospective power analysis) aids the design of experiments that are capable of detecting differences when they are present (Marvier 2002; Perry et al. 2003). Power analysis deserves particular attention in risk assessment, where avoiding Type II errors (concluding no difference when there actually is a difference) is more important than avoiding Type I errors (concluding that there is a difference when there actually is no difference), because the former may suggest a false sense of safety to regulators and decision makers with potentially hazardous consequences to the environment (Perry et al. 2009). While the Type I error rate (α) is typically set to 5% in scientific practice, β is usually uncontrolled.

In NTO risk assessment, the desired effect size that is defined in power analysis will equal the limits of concern. For example, one may want to be able to detect a difference of 30% difference between GM and counterpart, because this effect size constitutes a biologically relevant effect. When an effect size is defined in power analysis, it is assumed that variation around
means (i.e. variation between replicate plots in a field experiment) indicates sampling quality. Consequently, power analysis will indicate low statistical power for species with high abundance variation between replicates. When limits of concern are set arbitrarily at the same value for all non-target species, this results in large differences between species in statistical power and required sample sizes, even though it is unclear whether the originally assumed limit of concern constitutes a biologically relevant effect in all species.

We performed a power analysis using variability estimates from a field trial conducted in the Netherlands where a number of non-target insect groups were counted on the GM potato “Modena” (grant no. NRR 30805, AVEBE UA, Foxhol, The Netherlands/BASF Plant Science Co. GmbH) and its near-isogenic comparator “Karnico” on six replicate plots (chapter 2, this thesis). Insect sampling was done by harvesting the complete above-ground material of one plant per plot and counting all insects on it. For this example, we used insect data obtained by counting insects on flowering plants at one of two locations (location A) in one of two years (2010). Assuming a two-sample t-test for the difference between GM and non-GM, we performed a power analysis to calculate the number of replicates that would be required to a range of desired effect sizes with a power of 0.8 and two-sided α = 0.05. Figure 1 shows the resulting power curves for different taxa, indicating the amount of replicate plots that would be necessary in order to detect effect sizes between 0% and 100% (percent difference on GM plant compared to counterpart). This analysis shows that when limits of concern are set to 30%, a minimum of 50 replicates would be necessary to detect such an effect with a power of 0.8 in the least variable species, and a multiple of 50 for more variable species. Already for the rather small plot size in this experiment (10 m² per replicate plot), the minimum required surface area would be 500 m² per cultivar, assuming that only the species with the highest power are tested. Although larger replicate plot sizes are likely to reduce variabilities, and thus the number of replicates needed, it is clear that experimental fields in non-target risk assessment require large surface areas, and that large differences in variability between species will make many species practically untestable. Some authors have therefore suggested using statistical power as a criterion for selecting focal species for field trials, thus omitting taxa with low statistical power from further analysis (Prasifka et al. 2008; Albajes et al. 2012). However, when the causes for the differences in variation are unclear, selecting taxa by power can create an undesirable bias in the selection.

The above-mentioned problems of large differences in power and sample size requirements between species are a result of the assumptions that 1) the size of biologically relevant effects is the same across species and 2) that variability between replicate plots indicates sampling precision. However, when differences in variability between species have biological reasons, it may be useful to incorporate a measure of variability in the definition of effects. This is the idea behind standardized effect sizes, which will be discussed in the following paragraphs.
Figure 1: Power curves showing the sample size that is required to test for a given difference (%) between a GM potato plant and its comparator with a power of 0.8 in a two sample t-test. Data were taken from a field field experiment conducted in the Netherlands (chapter 2, this thesis: insect abundances on flowering plants at location A, 2010). Due to the large differences between species in variability, the required sample sizes to test for commonly used limits of concern of 30-50 % strongly differ between species.

Standardized effect sizes

Standardized effect sizes quantify the size of a difference in relation to the variation in the measured variable. Since the parameter of interest in risk assessment is the difference between a GM plant and its non-modified counterpart, we focus here on Cohen’s (1988) standardized difference ($d$). Cohen’s $d$ is calculated by dividing the difference between two means ($m_1 - m_2$) by a standard deviation ($s$) derived from the data. Cohen (1988) suggested using either of the standard deviations of the two compared groups for the calculation of $d$. Other authors have suggested using pooled standard deviations ($s_{pooled}$) for the calculation, where the standard deviations of the two groups ($s_1$ and $s_2$) are weighted by their sample sizes ($n_1$ and $n_2$) (Hedges 1981):

$$ s_{pooled} = \sqrt{\frac{(n_2-1)s_2^2 + (n_1-1)s_1^2}{n_1 + n_2 - 2}} $$

For a more detailed description of SES calculations in different situations the reader is referred to Nakagawa & Cuthill (2007).

Standardized effect sizes are increasingly reported in many disciplines as a supplement to statistical hypothesis tests indicating the magnitude and relevance of an effect, and they are routinely used for comparing effects across studies in meta-analyses. For example, Marvier et al. (2007) performed a meta-analysis of 42 field studies reporting non-target invertebrate abundances on Bt cotton and maize carrying different bacterial cry proteins, using Hedges $g$ (=Cohen’s $d$ with pooled standard deviations).

Standardized effect sizes quantify the size of an effect in relation to the variability in the measured parameter. In contrast, percentages quantify the size of an effect in relation to the size of the mean of the measured parameter. Thus, SES represent a different measurement scale for
effects. The choice of a measurement scale is also a decision on how risks are characterized (see also Andow 2003). In risk assessment, the relevance of a change in the abundance of a non-target organism may depend more on whether or not that change exceeds a certain level of biological baseline variation rather than on how large the change is compared to the mean abundance. When differences are quantified (and limits of concern are set) as SES, highly variable species are allowed to exhibit larger absolute differences in means in a comparative field trial than less variable species. High variation between replicate plots may be caused by high local reproduction rates or spatial aggregation behavior. In such cases, larger standard deviations would indicate higher biological baseline variation. It would therefore be reasonable to assume higher thresholds for biologically relevant effects (i.e. higher limits of concern). Thus, SES may be a way to account for biological baseline variation in the definition of effects on non-target organisms.

There has also been some criticism of SES, especially in toxicological literature, because “combined” (difference + standard deviation) effect size metrics may obscure absolute changes in variables (e.g. Lenth 2001; EFSA 2011). In toxicology, SES may indeed be less useful, because there is no reason to assume that high variability in the concentration of a toxic compound would affect the relevance of a concentration change for human health. The usefulness of SES for NTO risk assessment should also be confirmed by more ecological research on the relationship between abundance variability and population resilience on a larger spatial scale.

Standardized effect sizes can be used to perform power analyses in a similar way as other effect metrics. As shown in Figure 2, the power analysis becomes uniform across species with a single power curve for all taxa as a consequence of using Cohen’s $d$ to quantify effects. The number of replicates that is necessary to detect a given effect has now become independent of the particular species, because differences in variability between species are incorporated in the $d$ statistic. From this analysis more generic conclusions can be drawn with respect to experimental design. For example, detecting an effect of size $d = 0.8$ with a power of 0.8 requires 25 replicates.

![Figure 2: Power analysis of a two-sample t-test using standardized effect size $d$ as a measure of difference instead of percentages, which results in uniform power curves across species. For example, in order to detect an effect size of $d = 0.8$ with a power of 0.8, 25 replicates would be necessary.](image-url)
Standardized effects sizes can also be used to perform so-called equivalence tests. The idea behind equivalence testing is that the “onus is placed back on to those who wish to demonstrate the safety of GMOs to do high quality, well-replicated experiments” (Perry et al. 2009). In equivalence testing, a null hypothesis of ‘difference’ is assumed instead of the traditional null hypothesis of ‘no difference’. Equivalence is concluded when the difference between the GM plant and its comparator is shown to be significantly smaller than a given threshold value, which is defined *a priori* by the limits of concern. Equivalence tests can be illustrated by presenting the size of a difference and its confidence interval (CI), along with the zero line of no difference and the limits of concern (Cohen 1994; Nakagawa & Cuthill 2007). The outcomes of both difference tests and equivalence tests can be conveniently depicted in a single graph using CIs. Equivalence may only be concluded when the CI of the difference falls entirely within the specified limits of concern (for examples see Perry et al. 2009; EFSA 2010b). Confidence intervals can be calculated for standardized effect sizes by using the MBESS package (Kelley & Lai 2011) of the statistical software R (R Development Core Team 2010). Figure 3 shows, for different values of Cohen’s $d$, how the width of the CI for $d$ depends on the sample size. Confidence intervals for $d$ become narrower with larger sample sizes, thus increasing the chance that they fall within the limits of concern. On the other hand, the chance of concluding equivalence decreases with increased $d$.

![Figure 3](image)

**Figure 3:** Upper and lower bounds of 95% confidence intervals of standard effect size $d$, as a function of sample size for three values of $d$ (0.2, 0.5 and 0.8). For example, an effect size of $d = 0.8$ can only be concluded to be significantly different from $d = 0$ (i.e. confidence interval does not include 0) with a minimum sample size of $n = 15$ in a test of difference. In equivalence testing, when the limits of concern are set to $d = 1.0$, a same sample size of $n = 15$ only allows the conclusion of equivalence for effect sizes smaller than $d = 0.2$ (confidence limits fall within $d = ±1.0$).

**Univariate versus multivariate analysis**

Measures of effect sizes and limits of concern should reflect relevant changes in ecosystems. However, system-level changes are rarely considered in NTO risk assessment studies. GM effects are usually quantified separately per measurement endpoint, and limits of concern are set for each endpoint independently. Considering GM effects independently per species in univariate analyses lacks an ecological community perspective and could potentially miss important
changes at the system level. Consider the hypothetical example in Figure 4: abundances of three species (1, 2 and 3) are estimated for each of three genotypes (A, B and C). The abundances of all three species are above the limit of concern (LC) in all three genotypes, and would thus not be considered at risk. However, when all three species are considered simultaneously, it becomes clear that the three genotypes have (quantitatively) distinct NTO communities. These distinct communities may potentially exhibit distinct functionalities in terms of ecological services. Such system-level changes cannot be detected with univariate analyses, but instead require multivariate approaches. Some studies have used diversity metrics (e.g. Whitehouse, Wilson & Constable 2007; Farinos et al. 2008) or quantitative food webs (von Burg et al. 2011) for studying non-target effects of GM plants. However, community changes have to be linked to ecological service functionality, and standards for defining multivariate limits of concern need to be defined. These are important future challenges in NTO risk assessment.

![Figure 4: Hypothetical example of three species (1, 2 and 3) sampled on three plant genotypes (A, B and C). All three species are above the limit of concern (LC) in all three genotypes. However, from a multivariate perspective, all three genotypes host three (quantitatively) distinct communities. Bars represent 95 % confidence intervals.](image)

**Figure 4:** Hypothetical example of three species (1, 2 and 3) sampled on three plant genotypes (A, B and C). All three species are above the limit of concern (LC) in all three genotypes. However, from a multivariate perspective, all three genotypes host three (quantitatively) distinct communities. Bars represent 95 % confidence intervals.

**Concluding remarks**

The current practice of setting limits of concern in NTO risk assessment often lacks an ecological justification. Arbitrary limits of concern may be a starting point for detecting effects, but have important consequences for statistical power and experimental design. Using standardized effect sizes as an alternative way of quantifying effects may offer a biologically more meaningful way of quantifying effects in NTO risk assessments. However, the main challenge is the lack of knowledge about the biological causes of variability and whether or not this variability is related to the resilience of a species at larger spatial scales.

Multivariate analysis techniques offer a systems biology perspective to non-target risk assessment that is lacking in the current way of consideration endpoints separately. However, defining standards for multivariate analyses and for the setting of multivariate limits of concern are important future challenges.
Acknowledgements

We thank AVEBE UA, Foxhol, The Netherlands/BASF Plant Science Co. GmbH for allowing field trials with the GM potato “Modena”, and Kees Koops (Evolutionary Biology, Leiden University) for help with insect identification.

Funding

This study was funded within the ERGO program (Ecology Regarding Genetically Modified Organisms) of the Dutch Research Foundation (NWO), project no. 838.06.07
References


EFSA (2011) Guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed EFSA Journal, 9, 2438.


