GM Crops & Food, 10:63–76, 2019

ISSN: 2164-5698 print / 2164-5701 online
DOI: 10.1080/21645698.2019.1621615

REVIEW

Using problem formulation to clarify the meaning of weight of evidence and biological relevance in environmental risk assessments for genetically modified crops

Alan Raybould,[a] Karen Holt,[b] and Ian Kimber[c]

[a]Syngenta Crop Protection AG, Basel, Switzerland; [b]Syngenta Ltd., Jealott's Hill International Research Centre, Bracknell, UK; [c]Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

ABSTRACT. Weight of evidence and biological relevance are important concepts for risk assessment and decision-making over the use of GM crops; however, their meanings are not well defined. We use problem formulation to clarify the definition of these concepts and thereby identify data that are relevant for risk assessment. Problem formulation defines criteria for the acceptability of risk and devises rigorous tests of the hypothesis that the criteria are met. Corroboration or falsification of such hypotheses characterize risk and enable predictable and transparent decisions about whether certain risks from using a particular GM crop are acceptable. Decisions based on a weight of evidence approach use a synthesis of several lines of evidence, whereas a "definitive" approach to risk assessment enables some decisions to be based on the results of a single test. Data are biologically relevant for risk assessment only if they test a hypothesis that is useful for decision-making.

KEYWORDS. biological relevance; decision-making; hypothesis testing; risk assessment; weight of evidence

Correspondence to: Alan Raybould alan.raybould@syngenta.com Syngenta Crop Protection AG, Rosentalstrasse 67, Basel 4002, SwitzerlandSwitzerland
Received February 05, 2019; Revised May 15, 2019; Accepted 17 May 2019.

INTRODUCTION

Risk assessment makes predictions about the likelihood and potential severity of harmful effects that may occur following a course of action. The predictions contribute to decision-making about whether the action ought to be taken.[1] Sometimes, a risk may be characterized sufficiently for decision-making based on the result of a single test; such tests are often said to be "definitive".[2,3] In contrast, characterizing a risk using a synthesis of data from many sources to help reach a decision is called a weight-of-evidence (WoE) approach.[4] Reliability and relevance of data are central to discussions about how syntheses of data should be used in risk assessments to inform decision-making about uses of chemicals, including pesticides, and genetically modified (GM) crops,[5,6] and are also crucial to activities such as meta-analysis and systematic reviews.[7]

Discussion of WoE approaches to risk assessment has concentrated on the reliability of the data, which depends, among other things, on the validity of the methods used to obtain the data and the detail in which the data are reported.[8–11] Here we discuss the relevance of data. We propose that relevant data are those that rigorously test a hypothesis that the activity under assessment does not pose an unacceptable risk. This view of relevance leads to the idea that a WoE approach to risk assessment is not necessarily characterized by greater amounts or diversity of data than other methods, but by difficulties in organizing hypotheses and data to devise a single, clear decision-making criterion.

The ability of data to test hypotheses of no unacceptable risk also clarifies the concept of biological relevance, which is central to many regulatory risk assessments of GM crops, but which has proved difficult to define.[12] A biologically relevant difference shows a hypothesis of no unacceptable risk is false. It follows that if no conceivable value of a variable could falsify such a hypothesis, the variable is not biologically relevant for the purposes of the risk assessment. As we explain below, defining biological relevance in these terms targets risk assessment towards testing whether specific variables take particular preset values; for example, a human health risk assessment for a GM crop might test whether the concentration of an endogenous plant toxin exceeds a certain value that decision-makers have previously defined as unacceptable. The corollary is that risk assessment avoids profiling, which in this case would be testing the null hypothesis of no difference between the GM crop and a comparator for numerous compounds of unassigned importance. In other words, biological relevance should be defined at the beginning of a risk assessment, not thought about only once some statistically significant differences have been discovered.

We use hypothetical environmental examples to illustrate how hypothesis testing can clarify the concepts of weight of evidence and biological relevance for GM crop risk assessment and decision-making. Such clarification will help to identify data that are essential for decision-making and thereby reduce the likelihood that resources will be misallocated to the assessment of negligible risks or to the measurement of variables that have no relevance for decision-making.[13,14]

PROBLEM FORMULATION: RISK ASSESSMENT AS HYPOTHESIS TESTING

Risk is a combination of the probability and severity of a harmful effect. Risk is low if a harmful effect is unlikely, and its consequences would be minimal. Risk is high when a harmful effect is likely and would have serious consequences were it to occur. Ascribing a level of risk is more complicated when the probability is low and severity is high or vice versa. In these circumstances, the risk tends to follow the severity of the harmful effect: a severe effect that is unlikely is usually regarded as posing a higher risk than a minor effect that is almost certain to happen. Risk is unavoidable: every decision we make, including to continue as we are, may lead to a harmful effect. Improving decision-making by estimating the probability and severity of harmful effects is the purpose of risk assessment.

Effective risk assessment is a structured analysis of the risks posed by undertaking a certain

activity, such as cultivating a particular GM crop. Risk assessment should not seek to catalog all possible effects of an activity. Instead, it should define what would be regarded as its harmful effects, and then estimate the likelihood and severity of their occurrence (i.e., the risk) if the activity were to go ahead. Similar analyses could be undertaken for risks associated with not undertaking the activity, which may include the loss of potential benefits. Risk may not need to be quantified precisely; instead determining whether risk exceeds a threshold of acceptability is sufficient.[15] Agreeing the definitions of harm and the amount of risk that is acceptable, and devising a plan to test whether a decision poses an acceptable risk is called problem formulation.

In essence, all risk assessments test the same general hypothesis: there are no unacceptable risks associated with the activity that is being proposed. Acceptability of risk may be measured against an absolute standard; that is, there is some level of risk that is unacceptable in all circumstances. Alternatively, acceptability of risk may be judged relative to the potential benefits of the proposed activity.[16] In situations where the probability and value of benefits are sufficiently high, we may accept serious risk. On the other hand, if an activity is unlikely to produce benefit, or the benefits are likely to be small or uncertain, even minor risk may be unacceptable.

PATHWAYS TO HARM

The general hypothesis that a proposed activity poses no unacceptable risks needs to be translated into one or more specific testable scientific hypotheses. A conceptual model called a pathway to harm is a useful tool for producing such hypotheses. Such pathways set out the events that must occur if the intended activity is to produce harm, and are conceptually similar to adverse outcome pathways (AOPs) used in toxicology and ecotoxicology.[17–19]

The term AOP tends to be used when determining whether exposure to a chemical causes a particular adverse effect on individual organisms (hazard characterization), whereas a pathway to harm aims to determine the probability and severity of harm, which may be defined by the likelihood of damage to populations or the functions they provide (risk characterization). Nevertheless, the terms overlap in their meaning. Pathway to harm is used in this paper as this term is more common than AOP in the GM crop risk assessment literature.

A generic pathway to harm is given below:

Event A

(A proposed course of action)

↓

Event B

↓

Event C

↓

Event D

(A harmful effect)

Each step in the pathway can lead to testable hypotheses of the form "Event A does not lead to Event B", "Event B does not lead to Event C", and so on. The exact form of these "risk hypotheses" will depend on the definition of the harmful event and the amount of risk that is acceptable, and could be that Event B never occurs, Event B occurs only at a certain time or place (e.g., when or where the conditions for Event C to occur are not met) or Event B occurs below a certain frequency or magnitude (e.g., at a level unlikely to trigger Event C).

To illustrate how the hypotheses are used, we will consider initially only the first type – the absence of an event establishes that risk is acceptable.

Event A

↓ → Hypothesis 1: Event A does not lead to Event B

Event B

↓ → Hypothesis 2: Event B does not lead to Event C

Event C

↓ → Hypothesis 3: Event C does lead to Event D

Event D

(A harmful effect)

Proving that Event A will never lead to Event D is not possible. However, rigorous testing of one or more of the hypotheses 1, 2 and 3 may lead to the conclusion that Event D is highly unlikely and that the risk via this pathway is acceptable. Corroboration of a hypothesis under rigorous testing would suggest that the pathway is blocked at the respective point. Consequently, the risk from carrying out Event A would be negligible via this pathway, although it may be higher, and possibly unacceptable, via others.

Testing does not necessarily imply new studies as hypotheses may be tested with existing data. Formal testing with existing data may occur within the risk assessment, perhaps when it is not immediately clear why certain data are relevant or what they show. In such circumstances, the pathway is regarded as plausible and requiring evaluation and as a useful way to organize and communicate the relevance of existing data. Informal testing with existing data may also occur outside the risk assessment. This happens when a certain harmful outcome is deemed to be implausible and not requiring assessment; it is "obvious" that the proposed activity will not cause the stated outcome. "Obvious" means that if we were to develop a pathway to harm, it is easy to see how existing data would corroborate one or more hypotheses that steps in the pathway are highly unlikely to occur.[20]

A complete risk assessment is likely to evaluate several pathways to harm. The use of a particular GM crop may plausibly lead to more than one harmful effect. Also, there may be more than one pathway leading from the use of the crop to a specific harmful effect. Conclusions about the overall acceptability of risk posed by a proposed activity, and decisions based on those conclusions should consider all plausible pathways.

HYPOTHESIS TESTING

A rigorous test has the power to reveal that an untrue hypothesis is false. Although rigor and confidence are difficult to quantify, greater rigor in testing risk hypotheses leads to higher confidence in conclusions about risk.

Consider a hypothetical pathway to harm associated with maize producing a *Bacillus thuringiensis* (Bt) toxin for insect control:

Cultivation of Bt maize

↓

Dispersal of Bt maize pollen

↓

Pollen settles on the food plant of an endangered butterfly

↓

Ingestion of the pollen by larvae of the butterfly

↓

Toxicity of the pollen to the butterfly

↓

Fewer rare butterflies (harm)

We may decide to test the hypothesis that ingestion of pollen has no toxic effects on the butterfly. If this hypothesis is tested rigorously and corroborated, we will have confidence that the pathway to harm will not be realized and that risk to the butterfly is low, at least via this pathway. One such test would be to feed pollen of the Bt maize, or the Bt protein produced by the maize, to larvae of the butterfly in the laboratory. Rigorous testing would ensure that the concentration of protein or the amount of pollen ingested is exaggerated and in excess of worst-case predictions of exposure levels in the field.[21] If no increased toxicity were observed in groups of larvae exposed to Bt maize pollen or to the relevant Bt protein compared with groups fed control pollen or protein, there is strong corroboration of the "no toxicity" hypothesis and the pathway is blocked at that point. As a result, we might conclude that risks to the butterfly from exposure to the Bt maize pollen are negligible and acceptable regardless of the probability of earlier steps in the pathway occurring. If toxicity is observed, we may decide that we are unable to conclude that the risk is acceptable. In these circumstances, further characterizing the risk, perhaps by testing for toxicity at lower concentrations that more closely resemble likely exposures in the field.[22] In the remainder of the paper, we will call decision-making[a] ostensibly based on the result of a single study the "definitive approach" to risk assessment. We use the term because the studies on which the approach is based are called definitive studies (refs 2 and 3) and because we are unable to find a simple term that is commonly used to mean "opposite of a weight of evidence approach". Definitive is not meant to imply that decisions have greater reliability than those based on a weight of evidence, and certainly not that such decisions should never be reviewed, revised or revoked.

Now consider a situation in which it is not possible to culture the butterfly in the laboratory and test the absence of toxicity hypothesis in this way. Also, perhaps the Bt protein is toxic to some butterflies, but we have no data about its toxicity to butterflies that are closely related taxonomically to the butterfly of interest. Hence, we cannot run a study that will give us reliable predictions about the toxicity of the pollen to the butterfly, nor can we confidently exclude toxicity using existing data. Hence, the hypothesis that the pollen is not toxic to the butterfly is untested and is perhaps untestable for the immediate purposes of the risk assessment.

In these circumstances, we may choose to look at the probability of the pathway as a whole rather than focusing on a single step. Perhaps we have data on the distribution of the food plant which shows that it grows mainly in woodland where maize pollen rarely penetrates. We may also have observations on the ecology of the butterfly: it tends not to feed on leaves that are covered in pollen and many larvae die through overcrowding on the food plant; that is, there is strong density-dependent mortality.[23]

Putting these data together, we may make the following argument: even if the butterfly is as sensitive to the protein as is the target pest species, pollen is unlikely to collect on the leaves of its food plant in sufficient quantities to be toxic. If the pollen were to collect in sufficient quantities, it is unlikely that the butterfly would eat it. In addition, even if there were toxic effects on some larvae, the population size of the butterfly is unlikely to be reduced owing to strong density-dependent mortality; that is, exposure to the pollen would kill fewer larvae than would otherwise die through overcrowding.[b]

This is an example of a WoE approach to risk assessment. While none of the data we have provides rigorous ("definitive") corroboration of a single hypothesis in the pathway, we have reasonable confidence that the pathway will be realised rarely, if ever. We have shown that several steps in the pathway are sufficiently unlikely that their immediate consecutive occurrence, which is necessary for the pathway to be completed, will be very rare. We may conclude, therefore, that the risk via this pathway is acceptable.

DIFFERENCES BETWEEN THE DEFINITIVE AND WEIGHT OF EVIDENCE APPROACHES

Weed[24] states that one interpretation of weight of evidence is a simple premise: "all available evidence should be examined and interpreted" in a risk assessment. One may infer from this statement that the difference between the definitive and WoE approaches is that the latter considers all available evidence while the former does not. This inference would be incorrect. First, no method should consider all available evidence; it should consider only evidence that tests a hypothesis useful for decision-making (relevant evidence). Second, even though decision-making in the definitive approach may depend directly on the result of a single study, it still considers several lines of relevant evidence. The key difference between the definitive and WoE approaches is the way relevant evidence is organized.

Table 1 lists important stages in risk assessment and compares how they are dealt with in the WoE and definitive approaches. To be effective, a risk assessment needs a clear purpose, which comprises a reason for conducting the assessment, and a clear idea of how the results of the assessment will be used. In short, the risk assessment should be tailored to achieve policy objectives; it is not open-ended scientific research.[25] Risk assessments for GM crops are usually conducted to characterize the potential for harm to human health or

TABLE 1. Comparison of activities in definitive and WoE approaches to risk assessment.

Task	Definitive	Weight of evidence	Comments
Define protection goals	Identify the reason for conducting the risk assessment and the decision that the risk assessment will inform	Identify the reason for conducting the risk assessment and the decision that the risk assessment will inform	An unclear purpose may result in unfocussed data gathering that appears similar to a weight of evidence approach to risk assessment
Operationalise protection goals and set assessment endpoints	Define exactly what the risk assessment aims to protect and what the risk assessment will predict	Define exactly what the risk assessment aims to protect and what the risk assessment will predict	Difficulty in defining operational protection goals or assessment endpoints may result in unfocussed data gathering that appears similar to a weight of evidence approach to risk assessment
Define risk hypotheses	Formulates at least one hypothesis that is quantitative and incorporates a clear decision-making criterion	Formulates hypotheses that tend to be qualitative or semi-quantitative Not possible to formulate a hypothesis such that its corroboration or falsification is possible by a single test	An example of a quantitative hypothesis incorporating a decision-making criterion is $LC_{50}/EEC \geq 1$ An example of a semi-quantitative hypothesis is that the rate of hybridisation between a GM crop and a wild species will be no greater than the rate between similar non-GM crops and the wild species
Test risk hypothesis	The results are widely applicable; they are largely independent of local conditions The test is powerful and repeatable	The results apply only to the specific times or locations at which the test was performed The test is weak and its results are difficult to reproduce	Data from laboratory ecotoxicology studies are widely applicable Data on crop hybridisation rates may depend on local conditions (occurrence of sexually compatible species, weather etc.)

the environment or both, to inform decisions about regulatory approval or commercial development of a product.

Avoiding damage to human health and the environment are examples of policy protection goals. To be useful for risk assessment, these general aims must be made more specific. The policy protection goal of protecting biodiversity might be stated as preventing the cultivation of a GM crop from adversely affecting species of conservation value on non-agricultural land; this narrower aim is an operational protection goal. The aim might be made still more specific to make it tractable for scientific analysis; for example, relative to the cultivation of a similar non-GM crop, there should be no decrease in the population size of dormice in off-field areas during cultivation of the GM crop or in the following season. The abundance of dormice at a particular place and time is an assessment endpoint – an expression of a policy protection goal that is sufficiently clear for scientific analysis in a risk assessment. Garcia-Alonso and Raybould[26] discuss the practicalities of operationalizing protection goals and defining assessment endpoints for GM crop risk assessments.

Clearly defined assessment endpoints do not guarantee that a definitive approach to risk assessment is feasible or that a WoE approach will not be needed. However, unclear operational protection goals and assessment endpoints may reduce the ability to formulate decision-making criteria in terms of corroboration or falsification of a single hypothesis (see below). Also, unclear operational protection goals and assessment endpoints may indicate a flawed approach to risk assessment that begins by collecting many data and then tries to deduce policy objectives from those data. Such "science-led" risk assessment tends to include large amounts of data of unknown relevance. In contrast, "policy-led" risk assessment, which tests hypotheses clearly related to policy objectives, tends to use far fewer data.[14] Science-led risk assessment may appear to follow a WoE approach because it considers many data. However, it is really untargeted collecting of data, not testing hypotheses, and hence is the opposite of the WoE approach that we describe here.

The crucial difference between definitive and WoE approaches lies in their formulation and use of hypotheses. In a definitive approach, a decision is based on corroboration or falsification of a single hypothesis by a single study (or study type); whereas in a WoE approach, decisions are based on the results of testing several hypotheses, and each test may comprise several types of study or other sources of data. However, decision-making under the definitive approach is based on far more than the result of a single study.

Risk assessments that protect biological control functions from the side-effects of insect-resistant GM crops non-target arthropods (NTAs) are the archetype of a definitive approach. Risk is characterized by estimating the concentration of the insecticidal substance (e.g., a Bt protein) that has a certain effect on a test organism in the laboratory. The effect may be the LC_{50}, the concentration of the substance that kills 50% of a test population or the no observed adverse effect concentration (NOAEC), the highest concentration of the substance that has no adverse effect on the test organism.

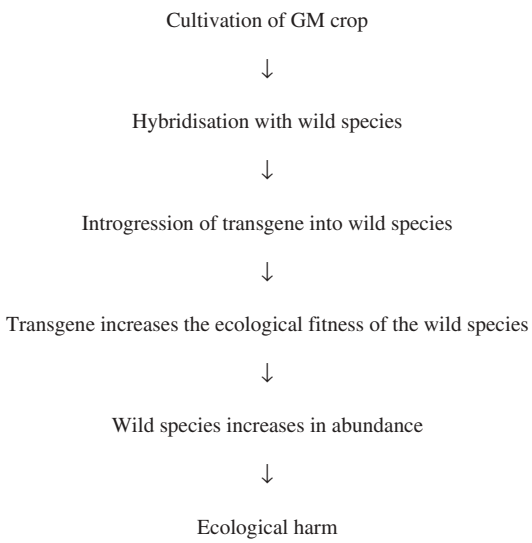
The effect concentration is divided by the concentration of the substance to which the test species or the species that it represents, will be exposed to in the field when the GM crop is grown – this is the estimated environmental concentration or EEC. The risk is acceptable if the effect/EEC ratio is greater than a specified "level of concern" (LOC). The LOC is set by policy; for example, the United States Environmental Protection Agency (US EPA) sets its LOC as $LC_{50}/EEC = 5$.[27] Another common LOC is $NOAEC/EEC = 1$.[28] The definition of the LOC represents a judgment about how conservative to make the decision-making criterion: if the LOC is too low, unsafe product uses may be judged to have acceptable risk; if it is too high, safe uses of useful substances may be judged unacceptable.[29]

If we know the EEC, a laboratory study to estimate the LC_{50} or NOAEC becomes a test of the hypothesis that the effect/exposure ratio is greater than the LOC. Sufficiently robust corroboration of that hypothesis would lead

to two conclusions: the risk to NTAs from exposure to the insecticidal substance is acceptable; and no further testing of NTAs is required. Acceptability of the risk to NTAs would contribute to a final decision about whether the GM crop should be cultivated. Falsification of the hypothesis may lead to further work to characterize the risk to NTAs or to a conclusion of high risk to NTAs.[22,30] In practice, several NTA species may be used to test the hypothesis about exceedance of the LOC because no single species is sufficiently representative of all potentially exposed NTAs. Nevertheless, each test would be definitive for all the NTAs represented by the species concerned.

The example above shows that a definitive approach to risk assessment relies on far more data than the results of a single study or the results of a set of similar studies that test the same hypothesis in the same way. Corroboration or falsification of the LOC hypothesis by the laboratory tests requires the calculation of the EEC, which uses data on the concentration of the insecticidal substance in various tissues of the crop, knowledge of the diet of NTAs, and models that predict the dispersal of the insecticidal substance in the environment.[31] Also, the choice of test organism relies on the concept of representativeness; that is, an organism that we test in the laboratory acts as a surrogate for a group of organisms in the field that we do not or cannot test. Deciding whether one organism is a suitable surrogate for others may use knowledge of the species' taxonomy and ecology and the mode of action of the insecticidal substance.[32]

In a WoE approach, existing knowledge is not organized to produce a single hypothesis that leads unambiguously to different actions when it is corroborated or falsified under rigorous testing. Instead, several hypotheses may be tested and the accumulated knowledge is used to make a decision. In environmental risk assessment for the cultivation of GM crops, a weight of evidence is often used to assess the ecological risks of gene flow from the crop to a related wild species.[33-35] A typical pathway to harm is as follows:



In principle, there is no problem in defining hypotheses that, if convincingly corroborated, would indicate acceptable risk; "no hybridisation" or "no increase in fitness" are perfectly clear hypotheses. Nevertheless, there are serious barriers to devising a hypothesis that could be tested with sufficient rigor by a single study and lead directly to decision-making.

There are two related problems. First, devising decision-making criteria similar to an LOC is difficult. For all of the steps above, any value above zero is likely to lead to our requiring more information because we are unable to say with confidence that a non-zero result is acceptable without further analysis: a single hybrid may be enough for the transgene to begin spreading; a slight increase in fitness conferred by the transgene may give an increase in abundance over several generations; and even a small increase in abundance may be considered harmful, especially if we are unclear about exactly what should constitute harm (i.e., our operational protection goals and assessment endpoints are vague, see above).

The second problem is that tests showing that one or more of the steps will not be completed may not be convincing. The result of the ecotoxicology study that tests the LOC hypothesis is accepted as applying in all situations that

the risk assessment covers. Suppose our risk assessment is for the cultivation of an insect-resistant GM cotton in the United States. We tested the toxicity of the insecticidal substance to *Coleomegilla maculata* (12-spotted ladybird), calculated the EEC, and found that the LC_{50}/EEC >5; therefore, we decided to ask for no more information to assess the risk to all beetles. Implicit in this decision is that the result of the study applies through the United States and assumes, among other things, that the test adequately predicts the sensitivity to the insecticidal substance of beetles in Arizona, Louisiana and all other states where the GM cotton may be grown.

In the gene flow example, we are likely to have less confidence in the general applicability of a study that shows that a step in the pathway will not occur. Say we introgressed the transgene into the wild species and grew the GM and similar non-GM plants in a trial simulating a natural habitat.[36,37] We then measured seed production and found no difference between the GM and non-GM plants, even under conditions of heavy insect infestation; thus, we concluded that the transgene is unlikely to increase the fitness of the wild relative. Would we accept this test as definitive and require no further information?

If the trial had been conducted in Arizona, we may be unsure that the result applies in Louisiana or indeed anywhere other than at the trial site under the conditions of the experiment. Perhaps variation in insect abundance and diversity, an interaction between herbivory and rainfall, or different values of any number of other environmental variables may reveal an increase in fitness conferred by the transgenes. Similar concerns could be raised about the predictive power of ecotoxicology tests – the sensitivity of insects to a toxin may vary from place to place, for example – but have not been considered important when using the results of ecotoxicology studies for risk assessment.[28]

In general, at each stage in the pathway to harm in the gene flow example we are likely to be put in the position of requiring more data. Existing data may show that a particular step can be completed, but we are unable to say what "non-zero value" of that step would not cause concern without knowledge of the probability of the other steps. Alternatively, the data may show that the step in question will not be completed, but we have insufficient confidence in the general applicability of the data for us to conclude no concern on this basis alone. Hence, conclusions about risk will need to draw on evidence about each step in the pathway.

This discussion shows that definitive and WoE approaches do not necessarily differ in the amount or variety of evidence used to assess risk. The main difference is that the definitive approach is able to use policy objectives to derive clear decision-making criteria and organize existing evidence such that a decision depends on the results of a single test. A WoE approach tends to be used when a single decision-making criterion cannot be set, owing to uncertainty about the implications of low probability events taken in isolation, and when there is low confidence that the tests of hypotheses are generally applicable.

Both definitive and WoE approaches use expert judgment[38] but in different ways. The definitive approach uses judgment to set clear decision-making criteria prior to a test of a hypothesis, whereas a WoE approach uses judgment to assess risk once the hypotheses have been tested. This difference may lead one to underestimate the role of expert judgment in the definitive approach and assume that judgment plays a significant role only in WoE assessments.

BIOLOGICAL RELEVANCE

As we have seen, problem formulation defines criteria for judging whether a risk is acceptable and devises tests of the hypothesis that those criteria are met. Studies or other sources of data that are not capable of testing such hypotheses are not useful for risk assessment. Being "capable of testing" means that the study can produce observations that show such a hypothesis is false; that is, observations can show the risk is unacceptable or at least that it cannot be shown to be acceptable without further work. Biological relevance is the property of being able to test hypotheses about the acceptability of risk for the risk assessment in hand.

Consider the Bt pollen toxicity study described above. The difference in percent mortality between treatment and control groups after several days' exposure Bt and near-isogenic non-Bt pollen, respectively, is clearly biologically relevant because it can falsify the hypothesis that LC_{50} /TER >5, which is defined as the acceptability criterion. Differences between other properties of the treatment and control groups, say their metabolic profiles at a given time, are not biologically relevant unless we have previously defined specific differences in the profiles as indicators of unacceptable risk.

This discussion reveals a crucial point: biological relevance is defined during problem formulation. In setting acceptability criteria based on mortality, we decide that mortality is a biologically relevant property and that a certain percent increase in mortality at a given concentration is a biologically relevant difference. The corollary is that properties other than mortality are not biologically relevant for this part of the risk assessment – testing for statistically significant differences in these properties would, at best, add nothing to the risk assessment.[14]

Attempts to define biological relevance usually start after the collection and analysis of data[12] owing to a failure to apply problem formulation. Instead of testing whether preset acceptability criteria are met, many studies carried out for risk assessment simply compare numerous properties of a GM crop and a non-GM comparator, an approach called profiling.[14] There are many techniques for profiling, including phenotypic characterization, compositional analysis, molecular characterization, various omics methods, and mass spectrometry.[3945]In essence, profiling searches for statistically significant differences and then tries to decide whether any are biologically relevant. This is an inefficient and ineffective method of risk assessment because it wastes resources measuring properties that are unrelated to decision-making criteria.[14]

The profiling approach to biological relevance is seen clearly in the phenotypic characterization studies that are usually mandatory for regulatory environmental risk assessments for the cultivation of GM crops.[12,39] Typically, many properties of the gross phenotype of a GM crop are compared with those of a near-isogenic non-GM line in multi-location field trials.[39] Such trials provide breeders with data on the performance of the crop and its likely suitability for commercialization. The trials could also provide data to test hypotheses about the acceptability of environmental risks from cultivating the crop. In practice, however, the trials do not test such hypotheses. Instead, they test null hypotheses of no difference and consequently, properties measured are not selected based on predetermined biological relevance.

Consider the hypothetical pathway from cultivating Bt maize to harm to a rare butterfly, above. Perhaps for a particular butterfly species, Bt maize pollen dispersal is the key element in demonstrating acceptable risk. Let's assume for the sake of argument that our knowledge of Bt protein concentrations in the pollen and the sensitivity of the butterfly to the protein lead us to conclude that if pollen densities on the food plant were to double over those measured for existing maize varieties then the risk posed by cultivating the Bt maize would be unacceptable. We may have field trial data for the Bt maize that show statistically significant phenotypic differences between it and a near-isogenic line. For the purposes of the butterfly risk assessment, only differences in properties that could plausibly lead to an increase in pollen deposition on the food would are potentially biologically relevant.

Suppose that only two properties measured in the maize phenotypic characterization study show statistically significant differences between the Bt crop and its near-isogenic line: germination rate and flowering time. We predict with high confidence ("we know") that the deposition of maize pollen on the butterfly's food plant cannot increase through a change in germination rate of the crop. Hence, for the purposes of this part of the risk assessment, germination rate is not biologically relevant.

If we cannot envisage an increase in risk through changes in germination rate via any pathway, then germination rate has no biological relevance at all for the risk assessment. In these circumstances, we do not need to evaluate germination data to complete the risk assessment, regardless of there being a statistically significant difference between the Bt crop and near-isoline. In addition, if we had no data on germination rate, we would not need to acquire any.

On the other hand, perhaps a change in flowering time of maize could increase the exposure of the butterfly to deposited pollen. Knowledge of the butterfly's ecology may lead us to conclude that if the GM maize flowers one month earlier then there is a low probability of a doubling of exposure to pollen in some populations of the butterfly, and that further work would be needed to characterize the risk from this change in flowering time. However, if flowering is brought forward by less than one month, there is a minimal probability of the exposure doubling and no further risk characterization is required. In these circumstances, flowering time is a biologically relevant property, but all changes that brought forward flowering by less than one month would not be biologically relevant for the purposes of the butterfly risk assessment.

This hypothetical example makes explicit that problem formulation can define biological relevance prospectively and retrospectively. If we had no phenotypic characterization data, problem formulation might lead us to design an experiment or orgainse existing to test whether the Bt crop can flower one month earlier than relevant comparator varieties. Measurement of other properties, including germinate rate, would not be necessary. Where we have phenotypic characterization data collected for other purposes, problem formulation enables us to quickly focus on the relevant data for risk assessment. We should test the hypothesis that the Bt crop does not flower one month earlier than the comparator. No other properties need to be evaluated, regardless of whether there are statistically significant differences. We use the property's ability to show that risk is unacceptable, not its statistically significant difference in a comparison of the GM and non-GM lines as the criterion for determining its biological relevance.

A consequence of the points above is that developers of GM crops should avoid using data from biologically irrelevant traits to make claims about safety. If a trait is not biologically relevant – such as germination rate, above – then lack of a statistically significant difference in that trait between a GM crop and a comparator should not be used to support a conclusion of negligible risk from using the GM crop. This would amount to a "free hit" because if there were a statistically significant difference, the developer could simply claim that it had no biological relevance. Nevertheless, even if developers wanted to refrain from collecting data that are not biologically relevant in risk assessment studies or avoid submitting such data if they have been collected for a purpose other than risk assessment, regulatory requirements may prevent their so doing.

To reduce the collection and submission of irrelevant data, regulators should therefore concentrate on defining traits they regard as biologically relevant. Phenotypic profiling of a GM crop – that is, describing all the differences between a GM crop and a non-GM crop – should not be an aim of regulatory risk assessment.[14] The aim should be to identify potentially unacceptable changes, which can be done efficiently and effectively only if harm, plausible means of its realization, and a criterion for acceptability of risk are defined before experiments are designed. Also, there is little point in creating detailed guidelines to make sure experiments have adequate statistical power-[46]) if the experiments do not measure properties that are biologically relevant. Although statistical power is important, ensuring that it is adequate cannot substitute for decisions about what changes to what properties would indicate that risk is potentially unacceptable.

CONCLUSIONS

Risk assessment is ineffective when risk managers (policy- or decision-makers) fail to define clear protection goals and decision-making criteria. Risk assessors are then placed in the invidious position of inferring protection goals and decision-making criteria – in effect, assessors have to set policy, which is not their job – or of

predicting the effects of an activity with no direction about what effects should be considered harmful. Evans, Wood, and Miller[47] call this situation the "risk assessment – policy gap".

Problem formulation closes the risk assessment – policy gap by defining protection goals and a series of events – a pathway – by which the activity being assessed could damage them. Problem formulation then derives a set of testable hypotheses about the acceptability of risk from the pathways. Testing these hypotheses characterises risk and helps decision-making. Biologically relevant properties are capable of providing data that test such hypotheses – there is some value of the property that falsifies a hypothesis about the likelihood of harm occurring. If no value of the property could falsify a relevant hypothesis, then the trait has no biological relevance for the risk assessment. Not all statistically significant differences between biologically relevant properties in a GM crop and a suitable non-GM comparator are biologically relevant – only those differences that falsify a hypothesis about the acceptability of risk.

Problem formulation reveals fundamental similarities between WoE and definitive approaches to decision-making. To be effective, both approaches should test hypotheses about the acceptability of risk from the activity being assessed. They should also identify biologically relevant properties that should be measured to test the hypotheses. In addition, both approaches require judgment to either devise a single hypothesis that leads directly to different decisions depending on whether it is corroborated or falsified by a single test or interpret the results of several tests of several interrelated hypotheses. Focusing on hypothesis testing protects the WoE approach from becoming simply data collecting. Consequently, resources can be allocated more effectively by ensuring that data are only collected when they strengthen testing of the hypothesis that is critical for decision-making.

NOTES

[a] We emphasize that decision-making does not necessarily refer to approval or non-approval of the proposed use of a GM crop; it may simply refer to deciding whether or not risk has been characterized sufficiently.

[b] The arguments about exposure and population dynamics may also be useful for evaluating risk if toxicity were observed in the toxicity study.

REFERENCE

1. Finizio A, Villa S. Environmental risk assessment for pesticides. Environ Impact Assess. 2002;22:235–48. doi:10.1016/S0195-9255(02)00002-1.
2. Villeneuve DL, Garcia-Reyero N. Predictive ecotoxicology in the 21st century. Environ Toxicol Chem. 2011;30:1–8. doi:10.1002/etc.415.
3. Manibusan MK, Touart LW. A comprehensive review of regulatory test methods for endocrine adverse health effects. Crit Rev Toxicol. 2017;47:440–88. doi:10.1080/10408444.2016.1272095.
4. EFSA Scientific Committee. Guidance on the use of the weight of evidence approach in scientific assessments. Efsa J. 2017;15(8):4971. doi:10.2903/j.efsa.2017.4971.
5. Krimsky S. The weight of scientific evidence in policy and law. Am J Public Health. 2005;95 (suppl 1):S129–136. doi:10.2105/AJPH.2004.045799.
6. Constable A, Jonas D, Cockburn A, Davi A, Edwards G, Hepburn P, Herouet-Guicheney C, Knowles M, Moseley B, Oberdörfer R, et al. History of safe use as applied to the safety assessment of novel foods and foods derived from genetically modified organisms. Food Chem Toxicol. 2007;45:2513–25. doi:10.1016/j.fct.2007.05.028.
7. Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q. 2016;94(3):485–514. doi:10.1111/1468-0009.12211.
8. Klimisch K, Andreae M, Tillman U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol. 1997;25:1–5. doi:10.1006/rtph.1996.1076.
9. Fenner-Crisp PA, Dellarco VL. Key elements for judging the quality of a risk assessment. Environ Health Perspect. 2016;124:1127–35. doi:10.1289/ehp.1409567.
10. Roth N, Ciffroy P. A critical review of frameworks used for evaluating reliability and relevance of (eco)toxicity data: perspectives for an integrated eco-human decision-making framework. Environ Int. 2016;95 (1):16–29. doi:10.1016/j.envint.2016.07.011.

11. Kaltenhäuser J, Kneuer C, Marx-Stoelting P, Niemann L, Schubert J, Stein B, Solecki R. Relevance and reliability of experimental data in human health risk assessment of pesticides. Regul Toxicol Pharmacol. 2017;88:227–37. doi:10.1016/j.yrtph.2017.06.010.
12. EFSA Scientific Committee. Statistical significance and biological relevance. Efsa J. 2011;9(9):2372. doi:10.2903/j.efsa.2011.2372.
13. Durham T, Doucet J, Snyder LU. Risk of regulation or regulation of risk? A de minimus framework for genetically modified crops. AgBioForum. 2011;14:61–70.
14. Raybould A, Macdonald P. Policy-led comparative environmental risk assessment of genetically modified crops: testing for increased risk rather than profiling phenotypes leads to predictable and transparent decision-making. Front Bioeng Biotechnol. 2018;6:43. doi:10.3389/fbioe.2018.00065.
15. Raybould A. Ecological versus ecotoxicological methods for assessing the environmental risks of transgenic crops. Plant Sci. 2007;173:589–602. doi:10.1016/j.plantsci.2007.09.003.
16. Sanvido O, Romeis J, Gathmann A, Gielkens M, Raybould A, Bigler F. Evaluating environmental risks of genetically modified crops – ecological harm criteria for regulatory decision-making. Environ Sci Policy. 2012;15:82–91. doi:10.1016/j.envsci.2011.08.006.
17. Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem. 2010;29:730–41. doi:10.1002/etc.34.
18. Tollefsen KE, Scholz S, Cronin MT, Edwards SW, de Knecht J, Crofton K, Garcia-Reyero N, Hartung T, Worth A, Patlewicz G. Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). Regul Toxicol Pharmacol. 2014;70:629–40. doi:10.1016/j.yrtph.2014.09.004.
19. Rhomberg L. Hypothesis-based weight of evidence: an approach to assessing causation and its application in regulatory toxicology. Risk Anal. 2015;35:1114–24. doi:10.1111/risa.12206.
20. Raybould A. The bucket and the searchlight: formulating and testing risk hypotheses about the weediness and invasiveness potential of transgenic crops. Environ Biosafety Res. 2010;9:123–33. doi:10.1051/ebr/2011101.
21. Romeis J, Hellmich RL, Candolfi M, Carstens K, De Schrijver A, Gatehouse AMR, Herman RA, Huesing JE, McLean M, Raybould A, et al. Recommendations for the design of laboratory studies on non-target arthropods for risk assessment of genetically engineered plants. Transgenic Res. 2011;20:1–22. doi:10.1007/s11248-010-9446-x.
22. Garcia-Alonso M, Jacobs E, Raybould A, Nickson TE, Sowig P, Willekens H, Van der Kouwe P, Layton R, Amijee F, Fuentes A, et al. A tiered system for assessing the risk of genetically modified plants to non-target organisms. Environ Biosafety Res. 2006;5:57–65. doi:10.1051/ebr:2006018.
23. Stiling P. Density-dependent processes and key factors in insect populations. J Anim Ecol. 1988;57:581–93. doi:10.2307/4926.
24. Weed DL weight of evidence: a review of concept and methods. Risk Anal. 2007;25:1545–57.
25. Hill RA, Sendashonga C. General principles for risk assessment of living modified organisms: lessons from chemical risk assessment. Environ Biosafety Res. 2003;2:81–88. doi:10.1051/ebr:2003004.
26. Garcia-Alonso M, Raybould A. Protection goals in environmental risk assessment: a practical approach. Transgenic Res. 2014;23:945–56. doi:10.1007/s11248-013-9760-1.
27. United States Environmental Protection Agency (US EPA) Biopesticides registration action document. Modified Cry3A protein and the genetic material necessary for its production (via elements of pZM26) in Event MIR604 corn SYN-IR604-8; 2007 Mar [accessed 2019 Jan 30]. https://www3.epa.gov/pesticides/chem_search/reg_actions/registration/decision_PC-006509_1-Mar-07.pdf.
28. Raybould A, Carron-Lormier G, Bohan DA. Derivation and interpretation of hazard quotients to assess ecological risks from the cultivation of insect-resistant transgenic crops. J Agric Food Chem. 2011;59:5877–85. doi:10.1021/jf1042079.
29. Chapman PM, Fairbrother A, Brown D. A critical evaluation of safety (uncertainty) factors for ecological risk assessment. Environ Toxicol Chem. 1998;17:99–108. doi:10.1002/etc.v17:1.
30. Romeis J, Bartsch D, Bigler F, Candolfi MP, Gielkens MMC, Hartley SE, Hellmich RL, Huesing JE, Jepson PC, Layton R, et al. Assessment of risk of insect-resistant transgenic crops to nontarget arthropods. Nat Biotechnol. 2008;26:203–08. doi:10.1038/nbt1381.
31. Head G, Brown CR, Groth ME, Duan JJ. Cry1Ab protein levels in phytophagous insects feeding on transgenic corn: implications for secondary exposure risk assessment. Entomol Exp Appl. 2001;99:37–45. doi:10.1046/j.1570-7458.2001.00799.x.
32. Romeis J, Raybould A, Bigler F, Candolfi MP, Hellmich RL, Huesing J, Shelton A. Deriving criteria to select arthropod species for laboratory tests to assess the ecological risks from cultivating arthropod-resistant transgenic crops. Chemosphere. 2012;90:901–09. doi:10.1016/j.chemosphere.2012.09.035.
33. Hokanson KE, Ellstrand N, Dixon AGO, Kulembeka HP, Olsen KM, Raybould A. Risk

assessment of gene flow from genetically engineered virus resistant cassava to wild relatives in Africa: an expert panel report. Transgenic Res. 2016;25:71–81. doi:10.1007/s11248-015-9911-7.
34. Hokanson KE, Ellstrand NC, Ouedraogo JT, Olweny PA, Schaal BA, Raybould AF. Biofortified sorghum in Africa: using problem formulation to inform risk assessment. Nat Biotechnol. 2010;28:900–03. doi:10.1038/nbt.1665.
35. Devos Y, Ortiz-García S, Hokanson KE, Raybould A. Teosinte and maize × teosinte hybrid plants in Europe – environmental risk assessment and management implications for genetically modified maize. Agric Ecosyst Environ. 2018;259:19–27. doi:10.1016/j.agee.2018.02.032.
36. Snow AA, Pilson D, Rieseberg LH, Paulsen MJ, Pleskac N, Reagon MR, Wolf DE, Selbo SM. A Bt transgene reduces herbivory and enhances fitness in wild sunflowers. Ecol Appl. 2003;13:279–86. doi:10.1890/1051-0761(2003)013[0279:ABTRHA]2.0.CO;2.
37. Halfhill MD, Sutherland JP, Moon HS, Poppy GM, Warwick SI, Weissinger AK, Rufty TW, Raymer PL, Stewart JCN. Growth, productivity and competitiveness of introgressed weedy Brassica rapa hybrids selected for the presence of Bt cry1Ac and gfp transgenes. Mol Ecol. 2005;14:3177–89. doi:10.1111/mec.2005.14.issue-10.
38. Mumpower JLStewart TR. Expert judgement and expert disagreement. Think Reasoning. 1996;2:191–211. doi:10.1080/135467896394500
39. Horak MJ, Rosenbaum EW, Kendrick DL, Sammons B, Phillips SL, Nickson TE, Dobert RCPerez T. Plant characterization of roundup ready 2 yield® soybean, mon 89788, for use in ecological risk assessment. Transgenic Res. 2015; 24: 213–225. doi:10.1007/s11248-014-9839-3
40. Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. Plant phenomics, from sensors to knowledge. Curr Biol. 2017;27:R770–R783. doi:10.1016/j.cub.2017.05.055.
41. Rayan AM, Abbott LC. Compositional analysis of genetically modified corn events (NK603, MON88017 x MON810 and MON89034 x MON88017) compared to conventional corn. Food Chem. 2015;176:99–105. doi:10.1016/j.foodchem.2014.12.044.
42. Davies H. A role for "omics" technologies in food safety assessment. Food Control. 2010;21:1601–10. doi:10.1016/j.foodcont.2009.03.002.
43. Ricroch A, Bergé JB, Kuntz M. Evaluation of genetically engineered crops using transcriptomic, proteomic and metabolic profiling techniques. Plant Physiol. 2011;155:1752–61. doi:10.1104/pp.111.173609.
44. Li R, Quan S, Yan X, Biswas S, Zhang D, Shi J. Molecular characterization of genetically-modified crops: challenges and strategies. Biotechnol Adv. 2017;35:302–09. doi:10.1016/j.biotechadv.2017.01.005.
45. García-Cañas V, Simó C, León C, Ibáñez E, Cifuentes A. MS-based analytical methodologies to characterize genetically modified crops. Mass Spectrom Rev. 2011;30:396–416. doi:10.1002/mas.20291.
46. Perry JN, Ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, et al. Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environ Biosafety Res. 2009;8:65–78. doi:10.1051/ebr/2009009.
47. Evans J, Wood G, Miller A. The risk assessment – policy gap: an example from the UK contaminated land regime. Environ Int. 2006;32:1066–71. doi:10.1016/j.envint.2006.06.002.