# Philosophical Scrutiny of Evidence of Risks: From Bioethics to Bioevidence

Deborah G. Mayo and Aris Spanos†

We argue that a responsible analysis of today's evidence-based risk assessments and risk debates in biology demands a critical or metascientific scrutiny of the uncertainties, assumptions, and threats of error along the manifold steps in risk analysis. Without an accompanying methodological critique, neither sensitivity to social and ethical values, nor conceptual clarification alone, suffices. In this view, restricting the invitation for philosophical involvement to those wearing a "bioethicist" label precludes the vitally important role philosophers of science may be able to play as *bioevidentialists*. The goal of this paper is to give a brief and partial sketch of how a metascientific scrutiny of risk evidence might work.

**1. Introduction.** Risk assessment controversies in biology and other sciences often revolve around disagreements regarding the nature, interpretation, and justification of methods and models used to learn from incomplete and uncertain data. While philosophers of science are ostensibly interested in helping to clarify, if not also to resolve, matters of evidence and inference, they are rarely consulted in practice for this end. Where philosophers are called on to play a role in risk debates, for example, on science panels, their input has largely been focused on the role of ethical and other value judgments in risk policy disputes. As welcome as such participation has been, our position is that issues about values in evidence-based policy call for corresponding attention to methodological issues that enter in collecting, interpreting, communicating, and evaluating the evidence. In Mayo and Hollander (1991), these were dubbed issues of "acceptable evidence" in deliberate contrast to policy questions about "acceptable risk." That risk assessment judgments intertwine with ethical and value judgments demands a greater methodological understanding that allows for a critical or metascientific scrutiny of the uncertainties,

†To contact the authors, write to: Deborah G. Mayo, Department of Philosophy, Virginia Tech, Blacksburg, VA 24061; e-mail: mayod@vt.edu; Aris Spanos, Department of Economics, Virginia Tech, Blacksburg, VA 24061; e-mail: aris@vt.edu.

Philosophy of Science, 73 (December 2006) pp. 803–816. 0031-8248/2006/7305-0030$10.00

assumptions, and threats of error along the manifold steps in risk analysis. Disagreements that might be attributed to diverging policy and ethical values may actually be the result of divergent assumptions guiding the construction and use of models, and disagreements in the foundations of uncertain knowledge and statistical inference. Although these issues are generally intermingled in debates, the philosopher of science's penchant for laying bare presuppositions of claims and arguments would afford real progress in understanding. If this is correct, then restricting the invitation for philosophical involvement to those wearing a "bioethicist" label overlooks the most constructive ways in which philosophers of science can and should contribute to these debates. If a label is needed, perhaps *bioevidentialist* would do.

An important advance represented by our co-symposiasts is the recognition that philosophers of science can serve an important role in developing and conducting a *metascientific analysis* of aspects of risk debates. Even where there is no dispute as to whether a given harm is of concern, they rightly observe, what is often disputed is whether there is evidence of the *existence* of that harm or hazard, and this in turn may revolve around such choices as which end points to measure and how risks are "framed." For example, as Thompson (2006, in this issue) notes, there is a higher estimate of risks of genetically modified (GM) crops if the focus is on the initial stages where uncertainties are high rather than on a later stage after which problematic cases are likely to have been weeded out. Or, again, associating hormesis with mechanisms of natural selection, Elliott (2006, in this issue) observes, renders it of greater scientific respectability than when associated with homeopathic ideas. Such conceptual analysis can help shed light on the nature of risk debates, but more is required to criticize and help to adjudicate competing risk assessments. The question is: Why stop with conceptual analysis? *Why not critique the reliability of the evidence and inferences?* The goal of this paper is to give a brief and partial sketch of the kind of metastatistical scrutiny we have in mind. We will apply these ideas to two examples raised by co-symposiasts Thompson and Elliot, GM crops and hormetic effects.

**2. Metastatistical Critique of Risk Inference Options.** To evaluate how much of a controversy in risk assessment is due to uncertainties in data and how much to conflicting values (social, ethical, economic, religious) requires being able to critically evaluate *what the evidence is*; and as risk evidence invokes probabilistic and statistical methods, an adequate metascientific scrutiny requires coming to grips with these methods. This does not mean that philosophers of science become statisticians, toxicologists, or the like: that would be both too much and too little. Too much because it would be impractical to become experts in all the arenas involved; too

little because there is a great deal of confusion and foundational unclarity among such 'experts'. A sufficient understanding of the inference methods together with a platform for raising questions about fallacies and pitfalls, we argue, could go a long way toward developing a metascientific (and metastatistical) scrutiny with real bite.

In particular, philosophers of science can serve an important role in developing and conducting an analysis of the various judgments and decisions required to determine if data constitute acceptable evidence of a given risk;[1] we might call these *risk inference options*[2] (e.g., choice of statistical significance levels, dose-response models). Because there is latitude for choice among possible inference options, and each choice influences the chance of obtaining evidence for a given risk (or benefit), much risk controversy revolves around these inference options. Notwithstanding the latitude in choosing inference options, we argue, it is possible to determine how different choices influence a method's ability to detect risks, that is, its *risk (or benefit) detecting capacity*. This would be the basis for systematically addressing the following questions:

1. How do various methodological choices made in the generating, modeling, and interpreting of data alter a test's risk detection capacity? (For example, Do data-dependent searches alter risk detection ability? If so, how should 'selection effects' be taken into account?)
2. What uncertainties and errors have been well ruled out? Which have been overlooked and why? (e.g., extrapolations beyond the lab). Are given policy standards met or flouted?
3. What are the statistical and the substantive assumptions in collecting and modeling the data? How well are they satisfied,[3] and what are the consequences for the reliability of inference of their being violated in the analysis at hand?

*2.1. Beyond Dirty Hands.* Failure to have a critical understanding of the (meta-) statistical issues often leads to the position that standards for estimating risks from statistical data are so bound up with subsequent policy decisions that scientists invariably (if unconsciously) introduce policy bias into the interpretation of risk evidence. "While scientists use the 95% rule or confidence limits to the 95% value, they remain loyal to the

1. Risk is usually distinguished from hazard assessment in including estimates of exposure, but our discussion will not turn on this.

2. These may also be called *risk assessment policy options*, as in the NAS-NRC report 1983 (Mayo 1991).

3. For a discussion of testing the model assumptions, see Mayo and Spanos (2004).

conventions of their discipline . . . but they implicitly 'dirty' their hands, . . . because they risk begging important regulatory issues" (Cranor 1993, 42).

The same point is most often put in terms *of type I and type II errors in testing*. In the present context these two errors may be informally summarized as follows:

> **type I error**: the data are taken as evidence of a risk (or benefit), when in fact the risk (or benefit) is absent (false positive),
> **type II error**: the data are *not* taken as evidence of a risk (or benefit), when in fact the risk (or benefit) is present (false negative).

The 95% rule refers to the requirement that a test have a low probability, for example, .05, of inferring a genuine effect (e.g., a genuine risk) when it would be an error to do so (commit a type I error).

The allegation is that choosing the trade-off between type I and II error rates is invariably to "dirty one's hands" with policy. This charge, however, stems from a caricature of statistical tests where a statistical report, based on arbitrary cut-offs, is taken to automatically warrant a policy decision. This would be an abuse of statistical tests. The dirty-hands allegation only underscores the need for a critical assessment of statistical tools; for it is a well-known fallacy to identify statistical significance with substantive importance (albeit still committed), how much worse to go straight from statistical significance to a policy decision. Although what counts as a risk of concern is a policy question, whether a statistically significant/nonsignificant result warrants the presence/absence of a given risk (increase or decrease) is not. The same holds for the various other risk inference options needed to interpret risk data. While, in any particular case, options may be based on 'unthinking conventions' (e.g., the .05 cut-off for statistical significance), on philosophical principles of evidence, or deliberately chosen to further policy preferences, it does not follow that any criticisms of resulting inferences are themselves matters of policy and/or value judgments. For example, one researcher may prefer a less 'protective'[4] extrapolation model for cancer risk on grounds of policy, but such models may be evaluated on grounds of statistical adequacy or predictive reliability, not on policy grounds.

Adding another level of complexity to our bioevidentialist task is the fact that the thorniest risk debates are often intermingled with foundational disagreements regarding methodologies of uncertain inference. Choices about which evidential methods to use may revolve around different philosophies of statistics, quite apart from deliberate policy choices.

4. A risk assessment option that makes it more likely to regard data as evidence of a risk. For a discussion of the protectiveness of RAP options, see Mayo (1991).

For example, the same data that would lead a significance tester to infer evidence of risk may lead a Bayesian statistician to assign a fairly high posterior probability to the no-risk hypothesis (Mayo 2005); or again, a Bayesian (or follower of the likelihood principle) would not regard evidence as altered because of 'optional stopping', whereas a Neyman-Pearson frequentist would (Mayo and Kruse 2001). Without taking sides, the bioevidentialist can compare the standards of protectiveness of the inferences licensed by different schools of inference in particular cases.

*2.2. Acceptable Evidence versus Acceptable Risk Policy Decisions.* In some discussions, the language of type I and II errors is taken out of the formal statistical context and exported into the arena of risk policy management; and unless one is very careful, confusion ensues. Identifying the type I error with "regulating a safe technology as if it has risks" and the type II error as "implementing an unacceptably risky technology," these discussions give ethical arguments for minimizing the type II rather than the type I error probability. It is important to distinguish such discussions of 'acceptable risks' from the current discussion of 'acceptable evidence':

(*a*) **Acceptable evidence.** Given the information and data, what inferences about the extent of risks (or benefits) are evidentially warranted?
(*b*) **Acceptable risk management**. Given the evidence of risk, what policies (or trade-offs) are acceptable?

Although questions under (*a*) and (*b*) are not always neatly distinguished, using the same terms to refer to an error in inference as an error in regulation leads to thinking that because the latter turn on ethical and policy judgments, so do the former. A question under (*b*) might be: should we fail to cut emissions despite the evidence of increased risks, to protect markets (Shrader-Frechette 1991, 2006, in this issue)? To address this, one may invoke general ethical principles that favor protecting the public versus protecting industry in cases with uncertainty where there is some evidence of public hazards. An opposing argument may weigh against such a precautionary stance in the face of likely economic consequences, and the debate may remain until further evidence. A question under (*a*) might be: do the data constitute evidence of greater increased risks than industry risk assessors allege? To suppose that disagreements about (*a*) also rest on ethical-policy differences is to forfeit the essential basis for charging that an assessment misinterprets *the evidence*.

A second consequence of using the same terms to refer to an error in inference as an error in regulation is that what is intended as advocating a protective policy stance is likely to be misunderstood as advocating the

minimization of the type II error *in statistical testing*. Taken seriously, this would allow erroneously construing data as evidence of risk with so high a probability that tests would easily become meaningless tools.

Moreover, even if one's concern is (*b*) (moving from risk estimates to policies), one should begin by scrutinizing the risk evidence, lest one's policy goal be inadvertently thwarted. Ironically, this often happens. For example, it may be argued that a 'precautionary' policy is warranted in cases where the evidence is ambiguous or inconclusive—a concern under (*b*)—but that depends on a distinct assessment of this inconclusiveness. As we will later note, current rules of thumb may allow the data to lead to 'inconclusive' reports when the data actually contain evidence of risk increases.

**3. How to Find Out the Truth with Metastatistics.** In a typical test of risk, we set up a null hypothesis, $H_0$, that asserts that there is no increased risk, and an alternative hypothesis, $H_1$, that says there is:

$H_0$: a risk difference $\delta$ is 0,
$H_1$: a risk difference $\delta$ is nonzero.

Or $H_1$ might be that $\delta$ is greater than 0 (one-sided test). A sample of size $n$ is represented by a set of *random variables* $\boldsymbol{X} = (X_1, \ldots, X_n)$, and the data $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$ is a realization of $\boldsymbol{X}$. The experiment might involve studying and reporting the observed difference in risk among those exposed (or 'treated') and those unexposed (or 'untreated') denoted by $d(\boldsymbol{x}_0)$.[5] One uses the observed $d(\boldsymbol{x}_0)$ to learn about the underlying risk increases that gave rise to the data, as given by the parameter $\delta$ in the statistical hypotheses.

A familiar test rule is: *Reject $H_0$ and infer that data $\boldsymbol{x}_0$ provide evidence of a risk increase if and only if* $d(\boldsymbol{x}_0)$ is statistically significant at the $\alpha$ level.[6]

When we speak of a test 'detecting a risk', we mean 'it reports a statistically significant result' (at the chosen significance level $\alpha$); with an 'insignificant result', by contrast, the null hypothesis is not rejected. Although this is often abbreviated as '$H_0$ is accepted', it is intended to be understood as data $\boldsymbol{x}_0$ do not provide evidence against $H_0$. (Subtleties between Fisherian significance tests and Neyman-Pearson tests will not alter our points, but see Mayo and Spanos 2006 and Mayo and Cox 2006).

5. These hypotheses must actually be stated in terms of parameters of a formal statistical model.

6. Observed difference $d(x_0)$ is statistically significant at level $\alpha$ if $P(d(\mathbf{X}) > d(x_0); H_0) \leq \alpha$.

*3.1. Problems with Statistically Insignificant Results.* A very common worry in risk analysis is that a test fails to detect a risk not because one is absent but, rather, because the test had little chance of detecting risks even if they exist. The concern is with the type II error. Rather than construe statistical risk reports as leading to "dirtying one's hands" with policy values, our metastatistician scrutinizes such claims and avoids taking *no evidence of risk* as *evidence of no risk*! Failure to reject $H_0$ does not license inferring that a risk increase is less than $\delta$, if the test has very little chance of detecting an increased risk of $\delta$, even were it present. The test, we would say, was not a very *stringent* or *severe* scrutiny of possible risks. Since the alternative $H_1$ asserts that there is some (positive) risk $\delta$, it is a *composite* hypothesis containing all the different values that the risk $\delta$ might take. Accordingly, the probability of a type II error will vary for each value of $\delta$, and it is not correct to speak of the type II error without specifying the alternative or discrepancy $\delta$ for which it is being calculated.

Some apparently fear that it would be too difficult for policy makers to understand how to scrutinize insignificant results. In fact, as we will see, the reasoning required is no more complicated than the reasoning that forms the basis of the criticism of a too insensitive test.

*3.2. Problems with Statistically Significant Results.* In other studies a null hypothesis of zero (0) improvement is tested:

$H_0$: an improvement (or benefit) $\delta$ difference is 0,
$H_1$: there is a nonzero improvement.

Data $\boldsymbol{x}_0$ may be considered to provide evidence for inferring that a treatment produces benefits when a null hypothesis of 'no benefit' is rejected at a low significance level. Here a metastatistical scrutiny would center on whether $H_0$ was rejected too readily; high type I error probability. Given that risks of interest are often of low probability, in one sense this is less often a problem than insensitive tests. However, there are aspects of the data and hypothesis generation procedure that can introduce high type I error probabilities into tests purporting to have controlled this error. One way this can occur is if the procedure searches for benefits and reports just those that are found. *In a classic example of "hunting for significance,"* suppose one searches through 20 differences and reports the one that reaches a significance level of .05. The probability of finding at least one, .05 level, statistically significant difference out of 20, even if the null hypotheses are all true, is approximately .64 (i.e., $(1 - .95^{20})$). So the type I error probability would be .64, not .05. Note that here it is the inference to the non-null alternative $H_1$ that lacks sufficient stringency or severity.

*3.3. Two Metastatistical Tools Based on the Severity Criterion.* We can systematize the above reasoning by supplementing standard statistics with metastatistical tools for interpreting (i) insignificant and (ii) significant test results. To allude to the two examples we consider in this paper, (i) is a negative result purporting to have evidence of absence of risk, for example, GM crops do not pose threats to untargeted species; while (ii) is a positive result that purports to have evidence of improvements, for example, low doses of toxins provide beneficial hormetic effects. That is,

(i) **Negative result**: A statistically insignificant departure from the null hypothesis of no risk is taken as evidence for $H$: risks do not exceed $\delta$. (Here $H$ corresponds to failing to reject the null hypothesis.)

A metastatistical rule must say when this is *unwarranted*: If there is a high probability that a test yields a statistically insignificant result, even though risk $\delta$ is present, then $\boldsymbol{x}$ fails to provide acceptable evidence that risks $\delta$ are absent.[7]

(ii) **Positive result**: A statistically significant departure from the null hypothesis of no risk (or benefit) is taken as evidence for $H$: risk (or benefits) exist. (Here $H$ corresponds to rejecting the null hypothesis.)

Again a metastatistical rule must, at the very minimum, say when this is *unwarranted:* If there is a high probability that a test yields a statistically significant result, even though improvements $\delta$ are absent, then data $\boldsymbol{x}_0$ fail to provide acceptable evidence that benefits $\delta$ are present.

*Severity Criterion* (*SC*). Regardless of whether we have statistically significant or insignificant results, and despite the fact that there are two types of statistical errors, we are able to identify a single principle for scrutinizing the acceptability of the evidence for any given claim $H$. To have a unified way of speaking let us adopt testing language wherein hypothesis $H$ 'passes a test' with $\boldsymbol{x}_0$ covers various ways in which $\boldsymbol{x}_0$ 'fits', 'accords with', or otherwise purports to provide evidence for $H$. The severity criterion states:

(*SC*) If there is a high probability a test passes hypothesis $H$ even though $H$ is false, then a passing result $\boldsymbol{x}_0$ fails to provide acceptable evidence for $H$.

*SC* captures our intuition that data $\boldsymbol{x}_0$ fail to provide good evidence for the truth of $H$ if the test had little chance of providing evidence against $H$, even when $H$ is false. Such a test, we would say, is insufficiently strin-

7. Initial discussions of this kind of metastatistical rule are in Mayo (1985, 1988, 1996, 2004, 2005).

gent or *lacks severity*.[8] The onus is on the person claiming to have evidence for $H$ to show that the claim is not guilty at least of egregious lack of severity, and metastatistical scrutiny can provide systematic ways to determine if they have succeeded.

*Anticipated Objection*. Given the controversies between frequentist and nonfrequentist (e.g., Bayesian) statistical approaches, do we bias things by assuming a frequentist error statistical paradigm? No, we limit ourselves here to egregious construals of evidence: any approach, frequentist or Bayesian, that is not able to mount the above criticism of inferences with high error probabilities should be seriously called into question.[9] The severity criterion applies also to the use of other statistical methods aside from significance tests, whether a confidence interval, Bayesian inference, or other. The standard statistical tests do not directly supply severity assessments; severity is a 'metastatistical' concept. However, error probabilities can be used to supplement methods of inference with a severity assessment that is sensitive to the actual outcome $d(\boldsymbol{x}_0)$ from whatever procedure has been used to infer the claim in question.

**4. How to Tell the Truth (about Insignificant Results) with Metastatistics: GM Crops.** Consider the case of GM crops, in particular, plants genetically modified to have pesticidal traits (now called 'plant incorporated pesticides'), such as genes to cause crops to produce Bt (*Bacillus thuringiensis*) toxin. A concern is the possible danger to nontarget species such as earthworms or monarch butterflies. Successful EPA petitions to deregulate a Bt crop are based on evidence of acceptable risks to nontargeted species. This evidence in turn is based on finding that the results of lab exposures are not statistically significant in tests of null hypotheses such as:

$H_0$: *Bt crops do not adversely effect untargeted species.*

The concern is that failure to reject the null may be due, not to absence of effects, but to the experiment and statistical tests not being stringent or powerful enough to detect them. The metastatistical rule for interpreting insignificant results comes into play. For example, as discussed in

8. This notion is developed in much more detail elsewhere (Mayo 1996; Mayo and Spanos 2006). The severity function SEV(.) has three arguments: a test $T$, an outcome or result $\boldsymbol{x}$, and an inference or a claim $H$. SEV(Test $T$, outcome $\boldsymbol{x}$, claim $H$), is to be read "the severity with which claim $H$ passes test $T$ with outcome $\boldsymbol{x}$." If there is a high probability that the test would purport to have evidence for $H$ even though $H$ is false, then SEV ($T$, $\boldsymbol{x}$, $H$) = low.

9. In Bayesian testing, small significance levels with large samples can lead to null hypotheses of 'no risk' receiving high posterior probabilities. In those cases, use of Bayesian posteriors to judge acceptable evidence is problematic (see Mayo 2005).

Marvier (2002), an experiment on four replicate batches of earthworms, 10 to a batch, were exposed to soil that included leaves from either transgenic Bt cotton or nontransgenic cotton, and after 2 weeks (too short a time to expect differences in survival rates) the exposed worms gained 29.5% less weight on average than the others. Because this difference is not statistically significant the study concluded that this particular Bt toxin did not impair weight gain in earthworms.

However, due to the low sample size (four replicates), and the large variability among replicates, the test has low capacity to detect adverse weight effects. Given that the number of replicates the EPA requires fails to take into account within sample variation, Marvier reports, very few of the experiments that resulted in statistically insignificant results had a high (90%) probability of detecting even a 50% change (either in survival or weight decrease). Nearly all had little power to detect risks of concern; abbreviate it as $\delta^*$.

> Data $\boldsymbol{x}_0$ 'fit' the no-risk hypothesis, but the probability is high that no increased risk is detected, even if risks as high as $\delta^*$ were present.

Although $H_0$ 'passed' test $T$, the test it passed was *not severe*—it is highly probable that $H_0$ would pass this test, even if the increased risk is actually as large as $\delta^*$. From (SC), a failure to reject $H_0$ with test $T$ does *not* license inferring that the increased risk is less than $\delta^*$. What counts as a "risk of concern" reflects policy values, but this critique does not.

The severity criterion also directs us to find specific values of $\delta$ that are large enough to be ruled out by dint of the insignificant result. The EPA test had a fairly high probability of detecting a weight change of 56.37% or more; thus the insignificant result may warrant ruling out a 56% decreased weight (between Bt-treated and control worm groups). Generally the reports in the literature provide what is needed for a metastatistical analysis; if not, that alone is grounds for questioning.[10]

There are lessons both for planning and interpretation. *Pre-data*: one should specify the effect size of interest (decreased survival, weight, offspring) and calculate the sample size for reasonable power to detect it. *Post-data*: one should scrutinize the severity attained—based on the actual outcome, variability, and so on. For good discussions, see Marvier (2002) and Burgman (2005, Chapter 11), as well as earlier references.

*Rules of thumb using confidence intervals are not immune*. Appealing to

10. Reports are typically in terms of the power of a test. Although a high power to detect $\delta$ is not necessary for a high severity that $H$: risk increase $< \delta$, it is sufficient. Thus, by selecting a test with high power for detecting $\delta$ one is assured of this much protection: a nonstatistically significant result warrants with severity a risk increase $< \delta$.

confidence interval (CI) estimates of risk is often thought to avoid misinterpreting insignificant results, but more care is needed. For example, a common rule of thumb is that if both the 0 effect and the risk of concern $\delta^*$ are included in the interval estimate formed from data $\boldsymbol{x}_0$, then the results are 'inconclusive' (see Burgman 2005, 341). However, the data may provide reasonably severe evidence of the presence of risk $\delta^*$ (using our criterion); and thus a report of 'inconclusive' may not be warranted.

**5. A Bioevidentialist Critique of Significant Effects: Hormesis.** Hormesis refers to a phenomenon in which a substance that is deleterious at high doses causes a response in the opposite direction at low doses (we can call such low dose reversals 'improvements' to steer clear from calling them 'benefits'). Attention to "framing effects" in risk controversies, as usefully delineated by Elliott (2006, in this issue) for this example, reveals that the way risks are characterized can have a psychological impact in risk controversies. But applying our "bioevidential" scrutiny lets us go much further in waging an effective yet nontechnical critique.

Calabrese (2005), a leading proponent of the hormetic hypothesis, has argued that hormesis is a widespread adaptive, stimulatory response. For example, while high doses of dioxin cause increases in tumors, Calabrese cites data showing a suppression of tumors at low dose exposure to dioxin. Here we have a case where rejecting one or more null hypotheses:

$H_0$: *no benefit (or even harms) at low doses,*

is the basis for inferring evidence of improvements or decreased risk at low doses. So right away the metastatistical question directed at a positive or statistically significant result (3.3 (ii)) comes to mind: *Have they properly controlled type I error probabilities (false positives)?* We know from the metastatistical rule for interpreting statistically significant results that if data $\boldsymbol{x}_0$ are to provide acceptable evidence for the presence of an effect then high severity demands that it not be highly probable to have reported such evidence erroneously. The onus is on the proponents of hormesis to supply convincing evidence that they are *not* open to misconstruing random effects as genuine.

Before discussing this case we want to emphasize that we are not purporting to decide one way or another about the controversial theory of hormesis—for starters, this short discussion could not do justice to so complex an issue (Mayo and Spanos 2007). Our goal is to illustrate how a metastatistical critique can provide standard ways for nonspecialists to *raise questions* even in dealing with complex evidence-based risks before arguing about what policies might be warranted assuming some risk evidence. Considering this case also helps to illustrate the point raised in Section 3.2: it does not suffice that a low type I error is *reported*—the

*actual* type I error probability may be a lot higher or it may be uncontrolled altogether, due to certain features of data-dependent selections. Finally, this case would seem to be of interest to philosophers of science both because of the relevance to evidence-based policy and the fact that it is regarded as a possibly revolutionary change in the standard models used in toxicology (Calabrese 2005).

*Evidential warrant for a paradigm change?* Although some hormetic effects are apparently uncontroversial, existing use of the linear threshold model in toxicology already allows taking these into account (via U or J shaped models) on a case by case basis. Calabrese and Baldwin (2003) want to go much further: they claim to have provided sufficient evidence to actually change the default assumption in toxicology: "These findings challenge the long-standing belief in the primacy of the threshold model in toxicology (and other areas of biology involving dose-response relationships) and provide *strong support* for the hormetic-like biphasic dose-response model characterized by a low-dose stimulation and a high-dose inhibition" (ibid., 246; emphasis added). As Crump (2001) points out, however, this would demand evidence of a near universal prevalence of hormesis. So the evidential hurdle for the bioevidentialist to consider is whether there is evidence of a sufficiently general hormetic effect. Given the difficulty of detecting the low dose effects of interest, Calabrese, Baldwin, and Holland (1999) decide to obtain their evidence of hormesis through a literature search of $n = 10{,}000$ studies. By putting together those that show apparently hormetic-looking risk assessments they make a case for this "strong support." But is there acceptable evidence for this?

Among various methodological questions to which these studies give rise, we limit ourselves to a question about the effect of 'hunting for statistical significance' (Mayo 1996; Mayo and Kruse 2001; Mayo and Cox 2006). Already aware of how type I error rates increase with hunting procedures, our bioevidentialist quickly grasps the gist of Crump's concern: "In order to properly control for the false-positive rate one would need to know how extensive the search was that located the data set. If the data set was the most hormetic looking out of 100 examined, then to conduct a statistical test for hormesis at the standard 0.05 level one should use $p = 0.0005$ (the solution to $1 - (1 - p)^{100} = 0.05$) rather than $p = 0.05$" (ibid., 672). In other words, the researchers would have needed a vastly smaller significance level for each case examined in order for the overall type I error probability to be small. Notice that the task for the bioevidentialist is not to figure out precise significance levels or other error probabilities, it is to point out the kinds of fallacies that must be put to rest. It might next be noted that the data on which they base their inferences are not themselves a random selection from all studies but, rather, are based on a point system they devise, which itself merits scrutiny. On

this point system, data are taken as evidence of hormesis simply because a study could have shown evidence of hormesis, whether or not it actually did. ("A data set could achieve a score as high as 6 (high end of the low evidence region for hormesis) even if there was no evidence for hormesis" [Crump 2001, 675].)

An effective strategy to demonstrate lack of control of the type I error probability is to apply the test to data deliberately generated to have the null hypothesis true (no hormesis). Such a simulation allows determining the expected distribution of scores from studies in which a hormetic effect is *not* present (i.e., false-positive rate.) Our bioevidentialist could make use of Crump's report: "Results of this simulation . . . demonstrates the scoring system does not control the false positive rate (indication that a hormetic effect is falsely identified when none is present). . . . Using the same scoring system, between 94.9% and 99.7% of the simulated data sets showed some evidence of hormesis (score > 2), even though no hormetic effect was present" (Crump 2001, 675).

Unless the results of this simulation are themselves faulty, it appears that Calabrese et al. (1999) have not put the hormesis hypothesis to a stringent or severe test: their data collection and analysis makes it far too easy to produce apparently supporting evidence even where we know the hormetic hypothesis is false. Although philosophers of science would not be expected to run such simulations, using critical information that exists or even asking whether such a challenge could be answered are important first steps.

**6. Concluding Comments.** We have argued that neither sensitivity to social and ethical values, nor conceptual clarification alone, suffices for the responsible analysis and understanding of today's evidence-based risk assessments and risk debates. Although issues of acceptable evidence are generally intermingled with those of acceptable risk management, the philosopher of science's penchant for laying bare presuppositions of claims and arguments would afford real progress in understanding. If this is correct, then restricting the invitation for philosophical involvement to those wearing a "bioethicist" label precludes the vitally important role philosophers of science may be able to play as *bioevidentialists*. We hope to encourage a move in that direction.

## REFERENCES

Burgman, M. (2005), *Risks and Decisions for Conservation and Environmental Management*. Cambridge: Cambridge University Press.

Calabrese, E. J. (2005), "Hormetic Dose-Response Relationships in Immunology: Occurrence, Qualitative Features of the Dose-Response, Mechanistic Foundations, and Clinical Implications," *Critical Reviews in Toxicology* 35: 89–295.

Calabrese, E. J., and L. A. Baldwin (2003), "The Hormetic Dose-Response Model Is More Common than the Threshold Model in Toxicology," *Toxicological Sciences* 71: 246–250.

Calabrese, E. J., L. A. Baldwin, and C. D. Holland (1999), "Hormesis: A Highly Generalizable and Reproducible Phenomenon with Important Implications for Risk Assessment," *Risk Analysis* 19: 261–281.

Cranor, C. F. (1993), *Regulating Toxic Substances: A Philosophy of Science and the Law*. Oxford: Oxford University Press.

Crump, K. (2001), "Evaluating the Evidence for Hormesis: A Statistical Perspective," *Critical Reviews in Toxicology* 31: 669–679.

Elliott, K. C. (2006), "A Novel Account of Scientific Anomaly: Help for the Dispute Over Low-Dose Biochemical," *Philosophy of Science* 73 (5), in this issue.

Marvier, M. (2002), "Improving Risk Assessment for Nontarget Safety of Transgenic Crops," *Ecological Applications* 12: 1119–1124.

Mayo, D. G. (1985), "Increasing Public Participation in Controversies Involving Hazards: The Values of Metastatistical Rules," *Science, Technology, and Human Values* 10: 55–68.

——— (1988), "Toward a More Objective Understanding of the Evidence of Carcinogenic Risk," in Arthur Fine and Jarrett Leplin (eds.), *PSA 1988: Proceedings of the 1988 Biennial Meeting of the Philosophy of Science Association*, vol. 2. East Lansing, MI: Philosophy of Science Association, 489–503.

——— (1991), "Sociological vs. Metascientific Views of Risk Assessment," in Mayo and Hollander 1991, 249–279.

——— (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

——— (2004), "An Error-Statistical Philosophy of Evidence," in M. Taper and S. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Consideration*. Chicago: University of Chicago Press.

——— (2005), "Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved," in P. Achinstein (ed.), *Scientific Evidence*. Baltimore: Johns Hopkins University Press.

Mayo, D. G., and D. R. Cox (2006), "Frequentist Statistics as a Theory of Inductive Inference," in *Optimality: The Second Erich L. Lehmann Symposium*, vol. 49, Lecture Notes-Monograph Series. Beachwood, OH: Institute of Mathematical Statistics.

Mayo, D. G., and R. D. Hollander, eds. (1991), *Acceptable Evidence: Science and Values in Risk Management*. Oxford: Oxford University Press.

Mayo, D. G., and M. Kruse (2001), "Principles of Inference and Their Consequences," in D. Cornfield and J. Williamson (eds.), *Foundations of Bayesianism*. Dordrecht: Kluwer Academic Publishers, 381–403.

Mayo, D. G., and A. Spanos (2004), "Methodology in Practice: Statistical Misspecification Testing," *Philosophy of Science* 71: 1007–1025.

——— (2006), "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction," *British Journal for the Philosophy of Science* 57: 323–357.

——— (2007), "Risks to Health and Risks to Science: The Need for a Responsible 'Bioevidential' Scrutiny," *Human and Experimental Toxicology*, forthcoming.

NRC (National Research Council) (1983), *Risk Assessment in the Federal Government*. Washington, DC: National Academy Press.

Shrader-Frechette, K. (1991), *Risk and Rationality.* Berkeley: University of California Press.

——— (2006), "Comparativist Philosophy of Science and Population Viability Assessment in Biology: Helping Resolve Scientific Controversy," *Philosophy of Science* 73 (5), in this issue.

Thompson, P. B. (2006), "How Risky Are Genetically Engineered Crops? How Philosophers Can Help Answer the Question," *Philosophy of Science* 73 (5), in this issue.