RISKS TO HEALTH AND RISKS TO SCIENCE: THE NEED FOR A RESPONSIBLE "BIOEVIDENTIAL" SCRUTINY

Deborah G. Mayo
Department of Philosophy
Virginia Tech
Blacksburg, VA 24061, USA
Phone 540-231-8488
E-mail: mayod@vt.edu

Aris Spanos
Department of Economics
Virginia Tech
Blacksburg, VA 24061, USA
E-mail: aris@vt.edu

INTRODUCTION AND AIMS

We are pleased to take part in this forum on ethical issues of hormesis risk assessment and policy. In our view, ethical issues surrounding evidence-based risk policy in general are not properly addressed if divorced from issues of the responsible interpretation of the associated risk evidence. The former, bioethical issues, are adequately addressed only along with an accompanying methodological critique that may be dubbed "bioevidential". Just as bioethics requires developing and applying knowledge of ethical theory and principles to the assessment of controversial risk policies, bioevidentialism calls for applying a critical understanding of theories of data, statistical modeling, and inference to the evaluation and assessment of controversial risk evidence.

We do not present ourselves as medical or toxicological experts. However, our combined areas of expertise—philosophical foundations of science, statistical inference and modeling—enables the critical evaluation of the uncertainties, assumptions, and errors along the manifold steps in arriving at inductive/statistical inferences underlying risk assessments. The focus here is evidence for hormetic hypotheses concerning carcinogenic risks. Our goal is not to pass judgment on the truth or falsity of hormetic theory, but to evaluate the epistemological warrant of the evidence given in support of hormetic hypotheses by some of their main advocates.

It is laudable that leading hormesis proponents are opening the evidential and policy-laden issues to widespread critical appraisal, as represented by this and other forums. We aim not to provide ammunition to those who take issue with the likely policy implications of accepting hormesis, but to constructively suggest how hormesis proponents may strengthen existing efforts at responsible self-criticism, and in so doing demonstrate the ethical soundness of the evidence on which recommended policies are based. We examine both the evidential sources themselves and critical overviews: Crump (2001), Zapponi and Marcello (2006), Thayer, et al (2005) and Kitchin and Drane (2005). Our remarks are also informed by the American Statistical Association's "Ethical Guidelines for Statistical Practice" which lists such rules as "Report the limits of statistical inference of the study and possible sources of error". As we proceed we will offer constructive suggestions for reporting if not ameliorating such errors in inference. We conclude that the consequences of deciding risk management policy with the current knowledge gaps poses risks not only to health but to science; see Mayo (1991).

A MINIMAL STANDARD FOR EVIDENCE

Hormesis refers to a phenomenon where 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 avoid calling them benefits.) Although some hormetic effects are 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 (C&B) well-known supporters of hormetic theory want to go much further: they claim to have provided sufficient evidence to change the default assumption in toxicology in general. We assume the main claims of Calabrese and Baldwin (e.g., 1998, 2003), Calabrese (2005) are well known to readers of this forum.

Evidence for hormetic hypotheses are based on data that disagree with one or more 'no effect' or null hypotheses asserting:

H_0: there is 0 risk decrease, or 0 improvement, at low doses. (H_0 might also include risk increases.) Although an observed risk decrease in low-dose compared to untreated (controls) does not logically contradict H_0, it may be regarded as statistical grounds for inferring:

H_1: there is evidence of improvements or decreased risk at low doses, which may then be the basis for a *hormesis hypothesis:*

H: observed improvements are due to a hormetic effect.

Data **x** purporting to provide evidence for hormesis, at minimum, accords with H_1 but more is required to have genuine evidence for *H.* Mere accordance with the data is too easy whether for statistical hypothesis H_1 or a substantive hormesis hypothesis *H*.

We focus here on the least stringent standard for evidence: if it can be shown that the observed accordance between **x** and *H* would very probably have occurred even if *H* is false, or if the test turns out to have very poor ability to discriminate between cases where *H* is genuinely indicated by **x** and those where it would be clearly fallacious to infer *H*, then there are grounds to question the scientific credentials of the particular inference to *H*. We can abbreviate this:

> ***Severity principle (Weak):*** If data **x** 'accords with' *H* but the test very probably would have erroneously inferred *H* even if false, then *H* is *not well warranted by* **x**.

To run afoul of this weak severity principle would seem to abrogate the very basis for using empirical data to appraise hypotheses, and is scarcely a source of controversy.

Far murkier are questions about what is required to show that seriously insevere tests are avoided. How does one succeed in inferring only reasonably severely warranted hypotheses? The bioevidentialist program approaches these questions by identifying classic examples of flaws and foibles of general types that are found across the landscape of uncertain inferences, whether formal or informal. If one deliberately considers circumstances that would, with high probability, have told against an observed accordance between data and *H*, and yet no flaw or error is detected, then the severity with which *H* passes is fortified; see Mayo & Spanos (2006a). It is therefore highly advantageous, if not obligatory, for those claiming to have evidence for *H* to show at least that egregious lack of severity is avoided. Bioevidentialist scrutiny can provide systematic ways to check this.

HUNTING FOR SIGNIFICANT HORMETIC EFFECTS IN THE LITERATURE

Calabrese and Baldwin (1998, 2003) obtain their evidence of hormesis through an extensive literature search of existing studies, carried out for different reasons, rather than through controlled trials testing a null hypothesis of no improvement. Since this may well be the only reasonable evidence available at present, it is important to address issues of evidence regarding these literature searches and the uses hormetic proponents make of them.

Among various methodological questions to which these studies give rise, the most notable are questions arising out of the effect of 'hunting for statistical significance'[i]. Although insisting on a low significance level before rejecting H_0 in favor of H_1 ensures a low probability of erroneously inferring evidence of improvement H_1 (low type I error probability), this error probability guarantee breaks down in the case of searching. In the hormetic case, the searching would be for low-doses, or for risk factors, that are prima facie consistent with hormesis. We may refer again to the Ethical Guidelines of the ASA (1999) which stipulates the need to:

> "Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading."

Let us put the issue as non-technically as possible: In order to avoid insevere inferences to H_1, standard statistical tests direct one to reject H_0 and infer data **x** provide evidence of a risk decrease if and only if the observed risk decrease is statistically significant at a small level α (e.g., .01 or .05). Suppose, however, that one searches through twenty differences and reports just the one that reaches a significance level of .05. The probability of finding at least one, .05 level, nominally statistically significant difference out of 20, even if all the null hypotheses are true, is approximately .64 [i.e., $(1 - .95^{20})$]. So the type I error probability would be .64, not .05. The inference to the non-null alternative H_1 has passed an insevere test. This concern is behind Crump's (2001) remarks:

"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." (Crump 2001, p. 672).

In other words, one would need to insist on a much smaller significance level for each case examined in order for the overall type I error probability to remain small. The task for the bioevidentialist is not to fix precise significance levels or other error probabilities, but to raise the kinds of problems that can prevent controlling error rates.

The data from the literature search may be all that is reasonably available, but it is important to recognize that they are not a random selection from all relevant studies. C&B have developed a specially designed point system to ferret them out. We discuss some problems with this point system elsewhere (Mayo and Spanos, 2006b). Crump demonstrates a lack of control of the type I error probability by applying their scoring rules to data deliberately generated so that the null hypothesis is true (no hormesis). Such a simulation allows determining what distribution of scores would be expected from studies in which a hormetic effect is not present (i.e., false-positive rate.) Crump finds, based on his simulation, that "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, p. 675). However, Crump's charge may be mitigated if this scoring system is merely to pinpoint cases worth following up. Even if many are actually not hormetic, C&B may escape the charge of high type I error rate so long as the cases identified as potentially supplying hormetic evidence are properly treated. We now turn to this.

ARE CRITICISMS MITIGATED?

The relevant criticisms could be mitigated in a number of ways. First, one may insist that the observed improvement picked out for closer scrutiny (by their scoring algorithm) show, in the original study, *a statistically significant* improvement. Second, one can help mitigate selection bias by a deliberate consideration of as much as possible of the available risk evidence, including factors with both increased and decreased risks as well as other studies on the same risks. Third, even failing to mitigate these threats to validity (by the first two means), clearly revealing this, and taking steps to scrupulously avoid misleading claims, would disarm criticisms. However, thus far, the hormetic proponents appear not to have mitigated and rarely fully expose such noteworthy shortcomings.

Improvements are Statistically Insignificant. Questions arise from the fact that the cases with the most impressive hormetic-looking effects have been picked out for close scrutiny precisely because they show a high incidence among controls. By chance alone, from time to time, a control group may show a higher than normal incidence of an effect, and a thorough literature search is bound to find them[ii]. The obvious danger is that the most impressive hormetic looking effects may simply be aberrations. Zapponi and Marcello (2006) point out a number of cases where the apparent evidence for hormesis is explainable by such high controls (despite the pattern reversing in other trials). Moreover, even where the incidence rate among low-treated subjects is lower than controls (else they would not have been picked out), the observed decrease is virtually never statistically significant.

To understand the implications of this, consider what is being asked in probing the relevant null hypothesis: can the hormetic dose group be considered to have come from the same population as the controls (with respect to the incidence of the effect in question)? Evidence for hormesis would correspond to a 'no' answer, and in particular, a no answer that results because the incidence rate in the low-dose group is statistically significantly lower than in the controls. That observed differences are insignificant means they fail to supply evidence against the null hypothesis:

$$H_0: (p_C - p_T) = 0 \text{ versus } H_1: (p_C - p_T) > 0$$

p_C and p_T being the population relative frequencies of the risk effect in the control vs. low-dose treated groups respectively. That the observed differences fail to reach statistical significance says, in effect, that the low dose group may be considered to have come from the same population as the control group. This is evidence against the hormetic effect in question[iii]. This underscores the danger of relying on a point estimate for dose-response without supplying an associated estimate of its reliability (e.g., via a standard error)

Problems also arise as regards generalizability. The many agents or substances that have an incidence rate of zero (0) or close to zero in the control group are omitted from the literature analysis of hormetic effects; see Zapponi and Marcello (2006). C&B (1998) are searching for cases where a low-dose treated group (of rats) show less cases with the risk effect than controls: there would be no room for observed improvement if controls are already 0. Since many substances associated with risk increases have 0 or near 0 risk rates among controls, it may be of concern that positive support for hormesis from the literature search does not extend to them.

Incompleteness of Evidence and Selective Reporting. Unlike deductive inference, where if a set of premises entails a conclusion *H*, then so do these premises in addition to others, in inductive inference, the addition of other premises can easily turn an impressive looking inference into an illicit one. In particular, to assess overall improvement, it must be recognized that substances are often linked with several risk effects. Selectively reporting on improvements, say, a decreased incidence of testicular cancer, when at the same low dose the data show an increased incidence of some other cancer, would be to omit important information; see Thayer et al (2005). Yet the study of the effects of cadmium chloride on the incidence of *testicular* tumors in male rats is taken as a striking example of hormesis while overlooking relevant evidence reported in the same study that cadmium injections at low doses (hormetic effect region) increased significantly the incidence of *prostate* tumors. Waalkes (2003) makes a good case that prostate tumors constitute the more serious effect on health because the testicular tumors are usually benign. When these results are viewed in conjunction with the relevant significance levels, the evidence for beneficial hormetic effects are called into question.

These seem reasonable questions many of which critics have asked. Scientific responsibility would seem to call for direct responses. Acknowledging them up front, will be the best way to disarm critics and strengthen the evidential credentials of the hormetic research program.

WHAT KINDS OF INFORMATION WOULD BE USEFUL?

(1) Reliable estimates of control incidence rates would enable determining if the high incidence among controls that form the most impressive evidence for hormesis are likely to be due to chance, to background exposure, or to unusually high susceptibility in the animals observed.

(2) Rather than ignore cases with 0 incidence in the control, it would be good to check that no increased incidence is seen even at the very low doses being examined. If none is seen, it would fortify the cases purporting to show evidence of hormesis, because it would increase the severity of the analysis. Were it a mere aberration we might expect increased risk incidence with low doses, so to the extent that none are seen, the cases picked out for study are strengthened.

(3) Now that hormetic hypotheses are achieving fairly widespread attention, we think that attempts to carry out genuinely controlled studies, with several gradations in the hormetic range, for at least some of the more impressive looking cases, should be considered. This will enable the researcher to assess the validity of the underlying statistical model in order to ensure the reliability of inductive inferences; see Mayo and Spanos (2004).

REFERENCES

American Statistical Association (1999), "Ethical Guidelines for Statistical Practice", http://www.amstat.org/profession/index.cfm?fuseaction=ethicalstatistics.

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 (1998), "Can the Concept of Hormesis Be Generalized to Carcinogenesis?" *Regulatory Toxicology and Pharmacology*, 28: 230-241.

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.

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*, 73:790-802.

Kitchin, K. T. and J. W. Drane (2005), "A critique of the use of hormesis in risk assessment," *Human and Experimental Toxicology*, 24: 249-253.

Mayo, D. G. (1991), "Sociological vs. Metascientific Views of Risk Assessment," pp. 249-279 in Mayo and Hollander (1991).

Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

Mayo, D. G. (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.

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

Mayo, D. G. and M. Kruse (2001), "Principles of Inference and their Consequences," pp. 381-403 in *Foundations of Bayesianism*, edited by D. Cornfield and J. Williamson, Kluwer Academic Publishers, Netherlands.

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

Mayo, D. G. and A. Spanos (2006a), "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction," *The British Journal for the Philosophy of Science*, 57: 323-357.

Mayo, D. G. and A. Spanos (2006b), "Philosophical Scrutiny of Evidence of Risks: From Bioethics to Bioevidence," *Philosophy of Science*, 73: 803-816.

Mayo, D. G. and D. R. Cox (2006), "Frequentist Statistics as a Theory of Inductive Inference," pp. 77-97 in *Optimality: The Second Erich L. Lehmann Symposium*, vol. 49, Institute of Mathematical Statistics. Lecture Notes-Monograph Series, Ohio.

Thayer, K. A., R. Melnick, K. Burns, D. Davis and J. Huff (2005), "Fundamental Flaws of Hormesis for Public Health Decisions," *Environmental Health Perspectives*, 113: 1271-1276.

Waalkes, M. P. (2003), "Cadmium Carcinogenesis," *Mutation Research*, 533: 107-120.

Zapponi, G. A. and I. Marcello (2006), "Low-Dose Risk, Hormesis, Analogical and Logical Thinking, *Annals of the N.Y. Academy of Sciences*, 1076: 839-857.

FOOTNOTES

[i] For some general discussion see Mayo, 1996, Mayo and Kruse, 2001, Mayo and Cox, 2006.

[ii] Likewise, however, one can find apparent improvements (observed risk decreases) in the highest dosed groups.

[iii] For instance, on the basis of table 1 in C&B (1998), the test statistic comparing the difference between the proportions of the control and treated groups at low dose (.01) in male rats is:

$$\frac{\frac{10}{73}-\frac{6}{71}}{\sqrt{\frac{\frac{10}{73}\left(1-\frac{10}{73}\right)}{73}+\frac{\frac{6}{71}\left(1-\frac{6}{71}\right)}{71}}}=1.008[.157]$$

with a p-value in square brackets. Similar lack of significance can be shown for each entry.