Perspective

The Pebble Remains in the Master's Hand: Two Careers Spent Learning (Still) from John Evans

Adam M. Finkel[1,*] **and George M. Gray**[2]

In this article, we discuss four vexing problems in risk-based decision making that John Evans has addressed over the last nearly 40 years and has perennially challenged the two of us and others to think about. We tackle the role in decision making of potential thresholds in dose–response functions, how the lack of health reference values for many chemicals may distort risk management, the challenge of model uncertainty for risk characterization, and the yet-untapped potential for value-of-information analysis to enhance public health decision making. Our theme is that work remains to be done on each of these, but that some of that work would merely involve listening to ideas that John has already offered.

KEY WORDS: Decision making; risk management; uncertainty analysis; value-of-information

1. INTRODUCTION

The title of this essay refers to the mostly-forgettable 1970s TV series "Kung Fu," where the student was not allowed to leave the monastery until he could snatch a pebble from the master's hand—as soon as he did, he had to pack up and move on. The two of us left the "monastery" anyway, *sans* pebble, but are still learning from John Evans.

John started many of us on a lifetime of learning about risk assessment and management. In an increasingly irrational world, his mantra was always "analysis is useful" (along with "mice are more like rats than people are… in most cases").

But John has always stood for, and advanced, a *brand* of analysis that is not merely useful because practitioners say it is, or because it has more (deserved) appeal than "qualitative risk assessment" (Cox, 2008) or than the kind of analysis needed to justify "precaution" (Montague & Finkel, 2007; Wiener, 2001). When all one needs to implement a policy is "noun plus adjective," as in "[name of chemical here] BAD" or "These Expenditures BAD," the only "analysis" that is necessary is to claim that exposures to the substance, or the analogous "exposures" to the costs of control, *could be nonzero under some scenario and hence cannot be tolerated*. John's career has stood for the premise that these four more thoughtful (and less reflexive) attributes of analysis, among others, are what *make* analysis useful (Evans, 1986):

- Careful attention to uncertainty and to interindividual variability, keeping the two different phenomena conceptually and mathematically separate (Cullen & Frey, 1999; Morgan & Henrion, 1992), but combining them when enlightening (in particular, the extent to which any citizen can know what risk she faces is limited both by the uncertainty in anyone's risk and by the partial or complete inability of analysis to tell her where she falls on the distribution of interindividual risk; Finkel, 2008);

[1] Department of Environmental Health Sciences, University of Michigan School of Public Health, Ann Arbor, MI, USA.

[2] Department of Environmental and Occupational Health, George Washington University Milken Institute School of Public Health, Washington, DC, USA.

[*] Address correspondence to Adam M. Finkel, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA; tel: (202) 406-0042; adfinkel@umich.edu

0272-4332/21/0100-0678$22.00/1 © 2020 Society for Risk Analysis

- Full quantitation of exposures, risks, valued benefits, and control costs. We cannot pin any of these down precisely, but that's no excuse for reducing quantitative information to yes/no pronouncements, or to "green/yellow/red" bins (Cox, 2008)—instead, we should quantify the uncertainty whenever we quantify the quantities (but see Section 2.3 below for a few caveats).[1]
- "Analysis is useful" because it exists in service of better decisions. Analysis that exists merely to extrude more and more information-in-a-vacuum, divorced from any consideration of how the information can/should affect action, is vain, in both senses of that word.
- John has made us all think hard about the virtues of design/specification rules versus numerical targets (Wagner, 1999). We think he agrees with us that while we can certainly decide to "move the dial" on exposure until the marginal benefit of further reductions equals the marginal cost of more controls (that is, a performance standard dictated by cost-benefit balancing), this alone does not get us anywhere unless we understand how "the dial actually gets moved." The technologies are lumpy/discrete, and so what we really should be doing is using risk and economic information to compare real choices that are available to us. But John has also helped us remember that it's often too facile merely to advocate for "Best Available Technology (BAT)" or "As Low as Reasonably Achievable (ALARA)"—if we have the capability to reduce risks far below de minimis levels, but at ever-increasing costs, we should think hard before insisting that society does so.

2. FOUR VEXING PROBLEMS IN RISK-BASED DECISION MAKING

This essay discusses four vexing problems in risk-based decision making that John has shed massive light on over the last nearly 40 years, and has perennially challenged the two of us and others to think about. Our theme is that work remains to be done on each of these, but that some of that work would merely involve listening to ideas that John has already offered.

2.1. Thresholds are Irrelevant (or Worse) to Decision Making, Unless they Occur at Relevant Exposures

A thriving industry continues to attack the assumption that "low" doses of a substance will pose some risk when "high" doses are clearly risky. Even at a time when many scientists are pointing out problems with the traditional assumption that *noncarcinogens* must always have thresholds (e.g., Tennekes, 2016), dozens of papers annually are making claims about thresholds for *carcinogens*. Some of these articles (e.g., Bogen, 2019; Calabrese, 2004; Clewell, Thompson, & Clewell, 2019; Slikker et al., 2004) make generic claims about the ubiquity of thresholds for many carcinogens, or about levels below which exposures to carcinogens are salutary (via hormesis) rather than benign or harmful. Other articles (e.g., Pecquet, Martinez, Vincent, Erraguntla, & Dourson, 2018; Stelljes, Young, & Weinberg, 2019) claim that one particular carcinogen has, or "must have" a threshold. This controversy is quite fundamental: if a dose–response relationship has a threshold, then it may be *irrelevant* that effects are seen at "high" doses, and therefore any positive epidemiology or toxicology study should be discounted, ignored, or deemed "interesting" but not an indication of human risk. If adopted as science-policy, this stance could, of course, completely upend much of the practical utility of the fields of toxicology and epidemiology as they relate to chemical, radiological, and perhaps biological exposures.

The concept of the threshold has considerable merit, both for very low exposures to carcinogens (e.g., if faithful DNA repair exceeds the rate of new DNA lesions) or noncarcinogens (if, for example, mucociliary clearance can completely remove infrequent trespass by fine particles). *But from the underappreciated but absolutely fundamental point of view of decision theory and risk management, the existence of a "threshold somewhere" is completely unimportant to any decision that effects reductions in exposure from one point that is clearly above the threshold to one "above but less far above" said threshold*. Put another way, we assert that anyone interested in decisions should be unimpressed with a claim of threshold behavior unless it could possibly affect the magnitude of risk at *specific* "low" doses to which we might wish to regulate. Decisionmakers

[1] John Evans and I (A.F.) "naturally" assumed in our 1987 paper on the value of information (Finkel and Evans 1987) that risk was uncertain, but that cost was not. This was naïve of us, and I've written several papers since then arguing that cost uncertainty is often larger, but far more well-hidden, than risk uncertainty (e.g., Finkel, 2014a).

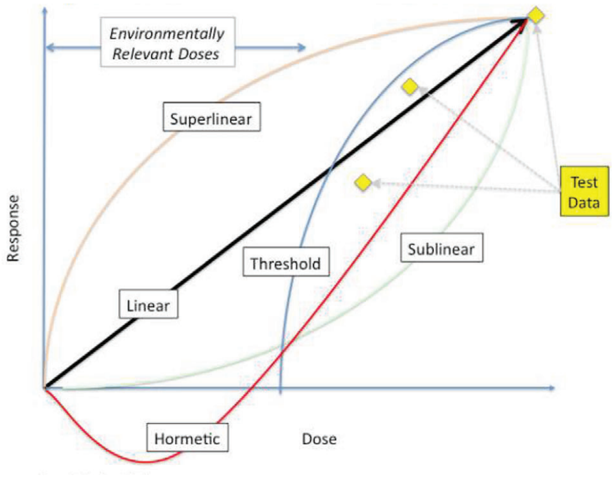

Fig 1. A stylized depiction of a hypothetical set of three exposure levels where adverse effects were seen ("test data") and how various dose-response models might fit the data acceptably well but have different implications for lower-dose risk.

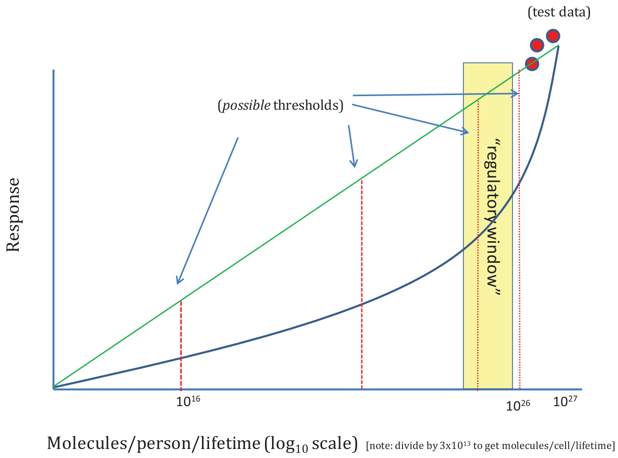

Fig 2. A different depiction of the same "test data," showing two possible locations of a dose-response threshold (the dashed vertical lines) that would have no practical relevance because no decision-maker is contemplating lowering exposures nearly to the point where the threshold occurs.

and the public need to understand that the health benefits of modest exposure reductions being proposed could well be the same *whether or not* the dose–response has a "threshold somewhere."

Figs. 1 and 2 offer two different mental pictures of the relationship between the typical configuration of the observed toxicologic or epidemiologic data and possible thresholds below the observed data. Fig. 1 (reprinted from Belzer, 2012, with permission of the author) is very typical of diagrams that show how various dose–response models can each provide reasonable fits to the observed data; it depicts the data as spanning much of the horizontal distance between the highest administered dose and the origin (zero dose), offering a compelling mental picture. More importantly, Fig. 1 implies that in the typical case, the threshold falls "just below" the observed range, and therefore leads the viewer to conclude that *for many, perhaps nearly all of the situations where extrapolation is necessary, linear extrapolation will grossly overestimate true risk because the exposure of interest in fact confers zero (subthreshold) risk*.

In contrast, Fig. 2 (developed *de novo* for this paper) depicts what may be a more accurate visual representation: the test data are in fact clustered very near each other when the horizontal scale is expanded to units of molecules/person/lifetime (see footnote 4). More importantly, Fig. 2 shows that in many cases, regulatory agencies seek to reduce exposures from somewhere at or near the low end of the range of the observed data to a new level that may only be a factor of 5, or 10, or 100 below that level—and *these* modest risk management reductions will only implicate a threshold if they happen to fall within a rather narrow portion of the complete dose–response relationship for the substance.

We emphasize that there logically are two situations in which a threshold would not be crossed: the one mentioned here (where pre-decision and post-decision exposures are both *above* the threshold), and the obverse case, in which pre-decision exposures are already *below* the threshold. This latter case could occur where the exposures seen in bioassays are never experienced by humans, or where the exposures seen in epidemiologic studies are no longer encountered anywhere.[2]

We learned from John Evans early on that what matters most in "real life" are the *consequences of error*—of making a choice that is inferior to another available option, especially if this squandering of benefit follows from misunderstanding the science or economics (decision theorists call this "regret"; Bell, 1982). So the continuing debate about the presence or absence of thresholds for carcinogens must begin

[2] Of course, a threshold is decision-relevant no matter how low it is, *when* that decision might involve a complete ban on a substance such that postregulatory exposures to it would also be zero. But these interventions are vanishingly rare; the Environmental Protection Agency (EPA) has only explicitly banned a handful of substances in its 50 years of existence, and the Occupational Safety and Health Administration (OSHA) never has done so.

by acknowledging the two basic ways in which misunderstanding this issue can lead to regret: (1) we can incorrectly *overestimate* the benefit (risk reduction) of any decision, by analyzing a change in exposure that crosses a threshold *as if* the dose–response was in fact linear (or nonlinear but nonthreshold); or (2) we can incorrectly *underestimate* the benefit if we analyze the change in exposure *as if* a threshold is crossed when in fact it is not.

A proper analysis of decision regret requires consideration of the probability of, and consequences resulting from, errors of either type. But it is impossible to even realize that these errors exist unless the concept of the threshold is *grounded* in an understanding, however imperfect or fragmentary, of *where* on the exposure–response continuum the threshold falls and where on that continuum any pending decision seeks to influence exposure. To foreshadow the conclusion of this discussion, we also suggest that grounding the debate in this way will lead to erasing the false black/white distinction between "no-threshold versus threshold responses." Instead, we urge analysts and decision-makers to consider that insisting that there *cannot* be a threshold for a particular dose–response may be unrealistic and unnecessary—infinitesimal amounts of a substance may well be harmless, but reductions to these levels almost certainly cannot be attained by regulation anyway. On the other hand, *insisting that there is a* ***relevant*** *threshold is, and should be, a difficult and evidence-based task*. We could help reconcile these positions by suggesting that if one considers the entire dose–response range from absolute zero to the LD_{50} and above, the most generic form of that function is arguably **"continuous with threshold"**—a relationship between exposure and risk that **changes** from threshold to nonthreshold when viewed as having two regions within which (and across which) exposure reductions have two very different implications.

So as EPA and other agencies continue to receive more and more pressure to "admit that there are thresholds for carcinogens," the question we have is how likely it is, and how likely it will be, that *poor decisions* will result from analyzing specific exposure reductions using the wrong model (either too precautionary or too naïve). We present here some brief observations about theory, practice, and policies that are of concern, because *they suggest a small but growing tendency to let vague claims of "thresholds somewhere" affect how we perceive and regulate exposures that may not implicate thresholds at all*.

2.1.1. Theory Regarding Thresholds

Many articles that claim threshold behavior is the general case for carcinogens as well as noncarcinogens begin (and often end) by asserting that when the exposure is sufficiently low, it is impossible for harm to manifest. Some authors merely claim that if the carcinogenic stimulus does not involve genotoxicity, "an ineffective threshold dose can be assumed" (Schrenk, 2018; p. 509). Others try to explain this "assumption" a bit more, by using the "proof by contradiction" logic (also known pejoratively as *reductio ad absurdum)*. See, for example, this representative quote from Schnell, 2016:

> [t]he existence of thresholds for carcinogens becomes inescapable when one simply converts the dose to number of molecules and plots it on a logarithmic scale, beginning with one molecule.[3]

Even if this logic is sound, it is in practice irrelevant: when regulators seek to reduce lifetime exposure to a substance from (say) 10^{25} molecules to 10^{24}, the behavior of the dose–response function at 1, 10, or 10 quintillion molecules is not worth a moment's thought (or a decade's delay in taking action while advocates investigate this behavior).[4]

Other papers focus on the extreme upper end of the exposure–response function and assert from theory or observation that as exposures are lowered, a threshold can or "must" be crossed. The two most common theoretical arguments of this type hold that: (1) there must be a discontinuity in exposure–response at a point where the incremental effects (e.g., numbers of mutations) are no longer as large as the spontaneous or "background" rate of these effects (Clewell et al., 2019); and (2) there must be a dose-dependent qualitative change in the mechanism

[3] This claim, in its obverse formulation, is also frequently made as a straw-man complaint against the linearity assumption; see, for example, Moghissi, Love, and Straja (2012), who write that "the LNT [linear, no-threshold] hypothesis is based on the single event process implying that either one photon or one molecule is needed to produce the effect."

[4] Although the precise magnitude of this large exponent does not change the argument much, consider actual occupational exposures to a common substance like benzene, when expressed on a molecules per lifetime scale. The OSHA Permissible Exposure Limit for benzene is 1 ppm, or 3.2 mg/m^3 in air. In a working lifetime of 13,500 workdays, a worker would inhale about 135,000 m^3 of air (at the standard 10 m^3/workday), or about 432 g of benzene at 1 ppm (and we note that the average benzene concentration of about 10,000 personal samples OSHA has taken in U.S. workplaces over the past 30 years is about 2 ppm). Since a mole of benzene weighs 78 grams, this represents about 5.5 moles of benzene, or about 3.3×10^{24} molecules.

of toxicity as dose increases, and only when the limiting mechanism is overwhelmed with increasing exposure is a threshold crossed (Slikker et al., 2004). The most common observational arguments for "thresholds somewhere" are either that observed data "upstream" from whole-animal bioassay results (e.g., mutation rates) can be better fit to a threshold function than to a linear function (Clewell et al., 2019), or, more sweepingly, that because human cancer rates have not decreased despite significant decreases in exposures over the past several decades, the assumption of linear dose–response must be faulty (see, e.g., Golden, Bus, & Calabrese, 2019, p. 4: "despite commitment of enormous societal resources to comply with LNT-based risk assessments, LNT-based cancer regulatory practices have failed to fulfill the promise of making meaningful differences in overall cancer incidence and mortality.")

In addition to the precariousness of these general claims ("must" there be a threshold? "must" this statement hold true for all subpopulations?), few of these studies attempt to ground their claims in terms of current exposures or pending policy decisions, and thus do not acknowledge that the only policy options on the table may in fact seek to reduce "high" doses of a substance to "slightly less high" ones (or "low" ones to "slightly lower"). The adjectives "high" and "low" are subjective, and what we need is a reliable means to infer specific "lower" responses from "higher" ones, not blanket statements that the two domains are unrelated or unrelatable.

Taken together, these two general lines of argument for thresholds may add up to less than the sum of their parts. First, the logic asserts that vanishingly small exposures "must" be innocuous (almost a truism, except that "vanishingly small" is often not defined or given context). Next, damages from exposures somewhat below frank effect levels are hard to "prove" (also a truism, but much more a statement about our power to observe what may be there than a statement about what is there). But the union of these two statements in no way demonstrates that a threshold exists in a policy-relevant window of exposure, only that one *may* exist somewhere between zero exposure and *some* level we might seek that might be below a level capable of causing harm. So to the extent that even a claim of a policy-relevant threshold for a specific substance (see below) merely results from reaching a lack of power to detect adversity in the underlying toxicology or epidemiology, we think it will be important to estimate the *lower statistical bound* on where the substance's "threshold" may fall. This calculation could be based on the power of the data to rule out a linear or sublinear dose–response, so we could evaluate whether this more precautionary "threshold" in fact still remains relevant for any completed or pending decision about the substance. Clearly, as in many other arenas, using the appropriate lower (or upper) bound on an uncertain quantity provides useful incentives for some interested parties to conduct more research, increase sample size, and so on, in order to narrow the range of uncertainty and thereby change the estimator in one direction or the other.

And just as first-principles sorts of arguments can and should be mustered to support the existence of thresholds, other such arguments can cast some general doubt on how policy-relevant true thresholds might be. In order for thresholds to commonly exist just at or near the exposure levels where our epidemiologic or toxicologic studies begin to lack power to discern statistically significant effects, it would have to be the case that these thresholds *just so happen* to fall where our studies (with their arbitrarily defined sample sizes based in large part on financial constraints) would have revealed them clearly, had the studies been much more powerful. That strikes some as more a coincidence than a finding. Similarly, it would have to be the case that humans have generally evolved to be at significant risk of cancer from levels of contaminants that are readily produced by common industrial processes when unregulated, and yet have evolved to be completely able to ward off any risk when those levels are reduced by (say) an order of magnitude or two. This is not far-fetched, but perhaps suggests that similar logical arguments that "all we need" are small reductions to move from risky to completely innocuous should be examined more carefully than they are at present.

2.1.2. *Practice Regarding Thresholds*

Many of the studies that assert a threshold for a *particular* substance (as opposed to the generic investigations above) conclude simply that "practical thresholds of exposure must exist below which there is no risk for cancer" (Slikker et al., 2004, p. 267), but do not even purport to pin down its location. Others (see, e.g., some of the 13 case studies in Slikker et al., 2004) provide a location for the threshold but do so via arguments that are not necessarily coherent. For example, their study of vinyl acetate concludes that exposures to this substance that yield concentrations of the metabolite (acetaldehyde) that are equal to or smaller than endogenous levels are "below biological thresholds," a conclusion that ignores the possibility that small percentages of the human population may in fact develop cancer due to endogenous exposures (and that also presumes this mode of action is known and unique). Still other investigations offer more thoughtful and persuasive evidence for the location of a specific threshold, but rarely to our knowledge do any of these articles *ground* the discussion of the threshold by comparing it either to prevailing exposure levels or (more importantly) to the health benefits or lack thereof associated with any pending attempts to reduce exposures by a given amount through regulation.

We are not at all suggesting that investigations into thresholds must explore regulatory policy within their publication, only that they should acknowledge somewhere that the threshold they identify (assuming they quantify at all) *may* not be relevant to any conceivable decision. This task is trivial and particularly necessary in extreme cases. We note that in at least one case with contemporary policy implications, advocates have invoked the threshold concept *even when assessing the benefits of the applicable risk reduction only requires* ***interpolation*** *within known frank-effect levels*. In 2015, for example, a consulting firm (Gradient Corp. 2015) commented to EPA that a threshold model for the dose–response of *n*-propyl bromide (nPB, also known as 1-bromopropane) was more appropriate than a monotonic function (linear or nonlinear). The comments stated further that because "the exposure concentrations used by [the U.S. National Toxicology Program] NTP (62.5–500 ppm) are several orders of magnitude greater than those modeled for ambient air for the general population … [the NTP results may not be] reliable for quantitative extrapolation from animals to humans." But one of us had pointed out in previous comments to EPA that the *current exposures of U.S. workers to nPB averaged 60 ppm*, and that more than one-third of all the samples OSHA has taken in U.S. workplaces exceeded 62.5 ppm, the exposure at which rodents showed an 800% increase in tumor incidence over controls. So, even if a threshold may exist outside of the range of observable data, it certainly will not be relevant in a situation where prevailing exposures exceed frank effect levels!

In future work, we hope to carefully evaluate all the peer-reviewed articles claiming that a particular toxicant has a threshold, by partitioning the set of articles into those which do—and those which do not—attempt to quantify where the threshold occurs, and then subdivide the first set further into those which do or do not make any reference to prevailing exposure levels and to levels contemplated by any agency for future reductions.

For those cases where analysts present a claim that a relevant threshold exists and a pending decision hinges on recognizing that fact, another question then arises: how should risk managers adjudicate the controversy and decide whether to regulate differently because one or more of the possible options may involve "going too far" and seeking meaningless and expensive exposure reductions? Clearly, this is just another situation where an existing evidence-based "default" assumption (see Section 2.3 below) could reasonably be supplanted by an evidence-based alternative. We offer no judgment here as to how receptive we think regulatory agencies should be to claims of a *decision-relevant* threshold; we simply observe that the errors of assuming a threshold when there is none, versus assuming there is none where there is, are different in kind. Therefore, as we discuss below, we urge agencies to consider the regrets of being too willing to accept speculative claims about relevant thresholds, versus the regrets of being too unwilling to do so. We also support, when reasonable, the addition of notations or other caveats in the EPA Integrated Risk Information System (IRIS) and other compendia of risk values, to signal to users that a given risk value might yield a substantial overestimate at some exposure levels because of the possibility of a relevant threshold.

2.1.3. Policy Regarding Thresholds

Mistaking a "threshold somewhere" for "a threshold that matters" would be merely a conceptual problem, if decisionmakers did not *act* on this distinction in questionable ways. We are concerned that the very notion of a "threshold carcinogen" is enticing decisionmakers to assess some carcinogenic risks in a very different way *without* considering whether prevailing exposure levels and/or desired postregulatory levels are in fact above the threshold. Papers such as that by Bevan and Harrison (2017) encourage regulatory agencies to treat the No Observed Adverse Effect Level (NOAEL) as a "perfect threshold" whenever they decide that the dominant mode of action is a nongenotoxic one. Indeed, EPA has already embraced this risk assessment policy change, although it has had few opportunities to date to implement it (in part because of doubts about how compelling the evidence for nongenotoxicity has been in many cases, but in larger part because the Agency has not regulated many carcinogens in the past 15 years). In its 2005 *Guidelines for Carcinogen Risk Assessment*, EPA (p. 3–23) established a policy that "in cases with sufficient data to ascertain the mode of action and conclude that it is not linear at low doses… an oral reference dose or an inhalation reference concentration, or both, should be developed." To date, we believe EPA has only used a Reference Dose (RfD) approach for a carcinogen in its 2001 appraisal of chloroform. However, stakeholders continue to proffer RfDs and RfCs (Reference Concentrations) to EPA for presumed "threshold carcinogens" in accord with this guidance. For example, Pecquet et al. (2018) recently developed an ingestion level of 0.26 mg/kg/day for tetrabromobisphenol A, which they state is a "no-significant-risk level," by estimating the lower bound of exposure causing a 10% tumor increase (from the animal tumor data) and dividing by a factor of 100 to adjust for inter- and intraspecies sensitivity differences.

The problem with this general approach, of course, is that it urges no concern about exposures below the RfD or RfC, which in turn *requires that the point of departure (NOAEL) truly* ***is*** *a threshold for the test animals, and hence that POD/100 truly is a threshold for humans with above-average sensitivity*. If instead, all the mode of action analysis is truly telling us is that there is a "threshold somewhere," and we mistakenly assume the threshold occurs just where our assays or studies lose sufficient power, then the RfD/RfC will *not* be a safe exposure, or even a "no-significant-risk" exposure. Assessing a "threshold carcinogen" via the RfD/RfC approach, in an exposure region above the *true* threshold, is a potentially serious error of underprotection.

The potential magnitude of this error is easy to estimate in the general case. The BMD_{10}, by definition, poses a risk to test animals of 10^{-1}; the NOAEL is well-known to pose a risk of approximately 5×10^{-2} (Leisenring & Ryan, 1992), because of the limited power of chronic bioassays to detect risks smaller than this. Therefore, *even if* humans are no more sensitive than the test species on average (with doses converted across species by a power of body weight or via a pharmacokinetic model), *and* if no human is more sensitive than the average human, the risk to humans at the NOAEL/100 will be approximately 5×10^{-4} if the true exposure-response relationship is linear in this region (the true threshold exists, but is not coincident with the NOAEL). And if the adjustment factors are doing their assigned job, (that is, for substances where humans *are* 10× more sensitive than the test species and where some humans *are* 10× more sensitive than the average human), the risk at the NOAEL/100 could be as high as 5%, the NOAEL risk "adjusted" properly. Needless to say, both 5% and 5×10^{-4} are risk levels (far) higher than Congress has instructed EPA to strive for in regulation of carcinogens.

In summary, the very concept of "the threshold carcinogen" encourages abandoning the exposure–response concept that is absolutely fundamental to our field. "Monotonic plus threshold," with the location of the inflection point estimated scientifically rather than by decree, and with uncertainty in both the slope (in the monotonic region) and the location of the threshold appropriately quantified, is in our view the way that the traditional "linear all the way" function should be improved upon. Only then can we thoughtfully undertake the central task of human health risk assessment—estimating the life-prolonging benefits of specified exposure reductions.

2.2. There is Too Great a Focus on Human Health Reference Values from "Authoritative Bodies" that Work too Slowly and Sometimes Work on the Wrong Things

Chemical risk management requires information to guide many decisions, including chemical substitutions, protective measures, or remediation efforts. For many chemicals, no authoritative body (e.g., EPA or the International Programme on Chemical Safety) has yet developed health reference values to inform these decisions. For example, EPA has no health reference value for tellurium, although it is on the 4th Contaminant Candidate List (CCL) (US EPA, 2016), required by the Safe Drinking Water Act, because of potential exposure in public water systems. In many other cases, even though a regulatory body ultimately produced a reference value, it was only after inordinate delays during which such a value did not exist, and the substance was therefore treated as if its risk was zero. For example, at this writing EPA is finalizing a risk assessment for 1-bromopropane that will eventually set cancer and noncancer reference values (US EPA, 2020). But quantitative toxicologic information sufficient to estimate this solvent's adverse reproductive risk was available in 2003; human studies showing a Lowest Observed Adverse Effect

Level (LOAEL) for neurological damage were available in 2004; and the NTP reported the final results of a strongly positive cancer bioassay in 2009.

When potency (and hence risk, and hence risk-based control) values will be assumed to be nonexistent (zero) until sufficient evidence accrues for an authoritative body to carry out an assessment, the incentives for all those advocating in favor of these substances' use flow in the direction of making it harder and harder to agree on those values. This complicates many decisions—especially those that involve chemical substitutions and the like—and simply sets up a risk treadmill as we move from one problem to the next. This yields a system wherein chemicals are "innocent until proven guilty"—so we urge that EPA's IRIS and the other "potency exercises" switch gears from the "gold or nothing" standard to a "provisional first; gold second" process in which the "10-year risk assessments" are done to *improve* provisional potency values, rather than being a precondition for having *any* official potency estimate. Under the status quo, important risks may be missed while the very slow gears of official assessment grind, and risk-increasing substitutions become a plausible outcome. When manufacturers wanted to remove bisphenol A (a chemical with authoritative values and hence in the spotlight) from their products they turned to unassessed chemicals with similar properties. An EPA evaluation demonstrated that for use in thermal printing paper the substitutions may have increased, rather than decreased, risk, based on assessing the risk of the substitutes using data currently in hand (US EPA, 2014).

Our current process for developing official human health reference values (HHRVs), such as those from IRIS, can take years or even decades to complete, and potentially endanger public health in the meantime. The slow pace of review leaves many potentially dangerous chemicals without risk values needed for good public health decisions. Even when they are published, they are invariably challenged by stakeholders, National Academies of Science (NAS) committees, and many others. We need something John Evans has advocated for decades, namely, faster ways to use existing information to generate risk values, even for chemicals with little chronic toxicologic data. And when this is done, we need to reflect the uncertainty in these values to help with decisions and guide future research (Gray & Cohen, 2012). As John Evans would say, uncertainty does not mean ignorance, and we can use the information available to help avoid the "missing risk" problem of unassessed compounds and provide better information for chemical management decisions.

Approaches to developing provisional HHRVs based on *in vitro* tests, structure–activity modeling, and empirical relationships exist now. Many of these have been around for a long time and are designed to help provide numbers useful for regulatory decisions (e.g., Layton, Mallon, Rosenblatt, & Small, 1987). Other approaches seek to use short-term data to predict points of departure for chronic risk assessment (e.g., Kratchman, Wang, Fox, & Gray, 2017; Pennington et al., 2002). These are usually independent of the specific toxic effect, which we know does not predict well across species anyway (Wang & Gray, 2015). Therefore, these approaches are subject to many of the same quantitative concerns that plague other risk management values. Especially of concern here is the claim that there may be a tautological relationship between risk values in different species due to constraints of experimental design (Bernstein, Gold, Ames, Pike, & Hoel, 1985; Brand, Catalano, Hammitt, Rhomberg, & Evans, 2001; Freedman, Gold, & Slone, 1993; Krewski, Gaylor, Soms, & Szyszkowicz, 1993). On the other hand, we know the status quo approach is maximally arbitrary, in that it guarantees false-negative conclusions by treating absence of (strong) evidence as evidence of absence. Provisional HHRVs may introduce some false positives and errors of overestimation of risk, but at least these errors would be overt and not hidden by the "missing risk" convention. It is very true that these estimates will be uncertain, and a real challenge is developing methods to characterize that uncertainty.

Perhaps the greatest challenge is getting people comfortable with using these provisional HHRVs. Decisionmakers will have to contend with uncertainty in an explicit way (Finkel & Gray, 2018). Other stakeholders will likely object too. For many, a lack of authoritative risk values is a feature, not a "bug," and the current slow and contentious approach avoids public and consumer scrutiny of their products. Using alternative methods to develop HHRVs means the default position will be that all chemicals for which some acute or chronic toxicity tests have been conducted pose some risk which may need to be managed—even those we can only apply read-across (Kovarich, Ceriani, Gatnik, Bassan, & Pavan, 2019) or quantitative structure–activity (Wignall et al., 2018) models to analyze. Toxicologists will tend to complain that decisions are being made without full testing of chemicals, while public-health advocates will focus on possible but untested sensitive populations or specific endpoints that might be of concern.

Despite these concerns, implicit treatment of no authoritative HHRV as meaning zero risk makes for bad decision analysis. We urge the development, perhaps by the current authoritative bodies but perhaps by new groups, of provisional potency values, with their attendant uncertainty, to ensure that public health decisions are well informed. These need to be evidence-based and theoretically-sound values with utility for risk management decisions despite their provisional status (and need be clearly labeled as such). In many cases, choices will be simple, with the provisional value sometimes clearly indicating a significant risk that is easily addressed, and at other times indicating a situation with no need for further action. In a decision context with greater stakes, tools like value of information (VOI) analysis (below) can then be used to characterize whether, and which, new data might need to be gathered to revise a provisional HHRV.

2.3. Two Fundamental Desiderata in Risk Analysis and Management May be Incompatible: The Desire to Fully Characterize Uncertainty, Versus the Desire to Apply Reasonable "Default" Assumptions and Models to Avoid Paralysis

Sources of uncertainty and variability abound in the quest to characterize, and explore the decision ramifications of, the risks of typical contaminants in the environment. Many of us (students and colleagues of John Evans, or not) have written articles and books categorizing the sources of uncertainty in risk analysis, and improving ways of quantifying and communicating them individually and collectively (e.g., Morgan & Henrion, 1992; Finkel, 1990; Cullen & Frey, 1999; Finkel & Gray, 2018). Most of these advances have emphasized the relatively uncontroversial treatment of *parameter uncertainty;* for example, using first principles, simulation, or other methods to account for uncertainty in the slope of the exposure–response function, the conversion of exposures from rodents to humans (Watanabe, Bois, & Zeise, 1992), the concentration of the contaminant at any location-time coordinate, and so on.

But uncertainties of equivalent or larger extent involve *model uncertainty*: for example, what about the substantial additional uncertainty contributed by the possibility that the proposed exposure reduction crosses a biological threshold? What about the possibility that effects seen (or not seen) in test animals are wholly irrelevant to humans, or that elevated incidence rates in human studies are confounded and not caused by the exposure? What about the possibility that the health effects associated with exposure are treatable and therefore not grave? Model uncertainty is far from straightforward to quantify, and more significantly, there is no firm consensus that it *should* be quantified, or how and if it should affect decision making.

But can (how can?) one be "in favor of uncertainty analysis" and yet be willing to put *any* significant uncertainty to the side? So we pose this fraught question: d*oes it violate a basic principle of uncertainty assessment to analyze risk and uncertainty* ***conditional on*** *a set of assumptions about causality, evidence, and relevance, or must we acknowledge (all) possible alternative assumptions and widen our uncertainty bounds because we can't be sure our assumptions are correct?* We have learned how to think about this conundrum (though not how to resolve it!) from John Evans.

There are two quite reasonable ways to confront a situation where the uncertainty contributed by not knowing which of two or more theories is correct about a risk dominates the uncertainty that would remain if we knew the correct theory. One way, which could be summarized as "full weight of evidence analysis," requires us to articulate and array all of the plausible models and derive risk estimates (with their parameter uncertainty and variability) for each. The other way, which happens for whatever reason(s) to have arisen earlier in the history of risk analysis (IRLG, 1979), involves making a judgment about which one model or theory will predominate, deliberately relegating other possible theories to "footnotes" in decision making.

This "determine the appropriate model" approach resembles how juries work in the court system: they weigh evidence, but not with a goal of reaching a verdict about "how guilty" or "how culpable" a defendant is, but whether he/she is simply "guilty" (more specifically, "guilty with enough confidence to be treated as such"). So this approach by definition requires consensus on two matters: (1) what assumption(s) will be used "by default," in the absence of sufficient information to the contrary; and (2) how much contrary information will be enough to discard the default assumption and substitute a different theory or model. Continuing the analogy to our jury system, we have long since become used to innocence (in criminal cases) or a verdict for the defendant (in civil cases) being the default, with the burden on the state or on the plaintiff to overturn the default presumption. The two legal realms differ, though, in how much of a hurdle the burdened party faces: in criminal matters, our system generally requires "proof beyond a reasonable doubt," whereas in civil cases the plaintiff prevails if she demonstrates her case via a "preponderance of the evidence."

To oversimplify, our system for human health risk assessment contains many inference assumptions in which a more precautionary stance prevails by default (e.g., that adverse responses in test animals are relevant to humans absent specific reason to doubt this general presumption), but also some significant assumptions that amount instead to a "presumption of innocence" (e.g., that humans all have the same "typical" extent of susceptibility to carcinogenesis (Finkel, 2014b), or that elevated relative risks in exposed subjects in toxicologic or epidemiologic studies are not considered unless they meet a strict statistical test of significance such as $p < 0.05$). As for the "how much contrary information?" standard, EPA and the other agencies have steadfastly refused to articulate one (despite repeated insistence by NAS/National Research Council [NRC] committees that they do so—see NRC 1994, 2009). We think it's fair to summarize that for most of the period 1980–2010, EPA and the other agencies were looking for "compelling evidence" to overturn a default, whereas more recently they have emphasized a more permissive approach, wherein the assumption chosen is the one that has "the most evidence" behind it, without regard to whether it would have been considered a precautionary default in times past (although there have not yet been many opportunities to implement this variant approach).

Each of the two very different ways to handle model uncertainty has distinct advantages. The multiple-models approach is far more faithful to the honest appraisal of uncertainty and the avoidance of overconfidence; the defaults/departures approach is generally far more efficient in avoiding strategic delay (by using default assumptions in the absence of contrary information, there is no need to ask "is anyone there?" ready to proffer some alternative assumption(s) and wait for someone to respond). A system based on defaults is also arguably (NRC, 1994, Appendix N-2) more practical, in that a generally somewhat precautionary risk assessment will emerge unless an interested party with the resources to conduct research that might lead to a lower risk estimate has the financial or other incentives to do so. Where alternatives are either far-fetched or not worth the trouble because "conservative" decisions are both obvious and acceptable to the regulated, defaults bring the finish line much closer to the present moment.

Both approaches, of course, also pose difficulties. In order for the "let all models bloom" approach to truly differ from a system built on defaults and reasoned departures therefrom, it must provide *weights* (subjective estimates of the degree-of-belief to be assigned to each model). Without weights, we would be left with multiple and incompatible risk estimates with no way to *either* combine them into a single uncertain estimate (you need weights for that) or no way to assess the expected consequences of acting as if the wrong model was correct (you need weights for that too). Otherwise, all one can say is "we might be very wrong, with unknown likelihood." So, without subjective weights, the first approach *becomes* the second approach—acting as if one single model is correct, without an estimate of how likely that model error is. Of course, the approach of assigning default assumptions, and of only switching from reliance on a default to reliance on a specific alternative in the face of persuasive evidence for it, can also be characterized as relying on subjective weights—here the weight given to the preferred assumption is always 1.0.

And the assignment of weights to more than one model at one time, though it has been accomplished and refined over many iterations (Evans et al., 1994, Oppenheimer, Little, & Cooke, 2016), is controversial and frequently criticized for being arbitrary and easily manipulable (NRC, 2007). In the limiting case, where one assumption predicts substantial risk and an alternative predicts zero risk, the composite uncertainty distribution is completely determined by the values given to p and (1-p), the weights assigned to each of the two incompatible states of nature—which means that the views of one expert can have more influence on a downstream risk management decision than a data set or a test result might have. And there is no way to avoid subjectivity in the assignment of model weights: starting from the premise that assumptions that are controversial should be given equal weights unless we can justify more precise parsing is *itself* a very value-laden and restrictive form of weighting (Finkel, 2018, p. S23).

The two of us do not necessarily fully agree about which of the two ways is better, but we agree that while model uncertainty is of great importance, it is possible to pay too much attention to it (or the wrong kind of attention) as well as too little.[5] We also fully agree that *if* multiple models are to be combined for a given step in the risk assessment equation, there is a right way and a wrong way to do so. We will explain our views in this regard by considering a special case of model uncertainty; namely, one in which there are only two possible (and incompatible models), one of which would predict a significant risk (of magnitude *X*) and the other of which would predict de minimis risk. To declare that "the risk is either of size *X* with probability *p* or very small (let's call that risk zero) with probability (1-*p*)," because one of two incompatible theories is right and the other is wrong, is a reasonable place to start. But one thing the quoted utterance does *not* mean is that "the risk is *p* times *X*." The average value of the uncertain risk may be *pX* (in the same sense that "the average human has approximately one ovary"), but we believe decisionmakers and affected persons need to understand that in such cases, the risk is actually either zero or *X*, and never *pX*.

It's not that averaging *per se* is wrong-headed at all—it can be worse *not* to average—but that analysts, decisionmakers, and consumers of risk information need to think carefully about what to average and *why*. When interindividual variability is the reason that a distribution exists rather than a single number, most people understand that averaging the data makes a profound statement: to say that the "best estimate" of the height of an adult human is 5 feet 4 inches should obviously not govern building codes setting doorframe sizes. When parameter uncertainty creates the distribution, using the average imposes a very specific value judgment upon the resulting action: that we regard errors of overestimation as precisely of equal consequence as errors of underestimation of equivalent size. Choosing a point estimate corresponding to any percentile of such a distribution (as opposed to the mean) merely imposes a *different* value judgment: perhaps we should be more concerned about needless expenditures than about needless (monetized) "lives lost," in which case we should tend to use a lower-bound estimate of the uncertainty distribution for risk. And when the distribution is really a bi- or multimodal distribution composed of two or more mutually exclusive assumptions, acting as if the risk is a weighted combination of the possibilities says that we are indifferent between erring by incorrectly giving credence to one assumption or the other.

Instead, we suggest that the right way to handle a situation wherein "the risk is either huge or tiny, depending" is to use that information to compare the performance of *two or more decision options*, not to average away the uncertainty in the risk. In these situations of fundamental uncertainty, we need to contrast two eventualities, and consider *what we might gain or lose if*: (1) the risk is huge but we make the decision that makes sense for a tiny risk; versus (2) the risk is tiny but we act as if we know it is huge (Brand & Finkel, 2019). This advice amounts to considering both competing models so as to *minimize the regret of choosing the wrong control strategy*. We can, of course, also consider the performance of a third decision—acting as if the risk is in fact *pX* and we control the risk as such—but whenever the harms (economic and/or physical) are nonlinear as the risk increases, or the available solutions are not continuous, but "lumpy," the optimal act for a risk known to be *pX* in size may be (very) different from the optimal choice in the real situation where the average value *pX* never manifests.

It may, of course, be difficult to predict how any given control option will perform *a priori*: both its efficacy and its cost will likely be to some extent a gamble. But we face this problem *whenever* we seek to choose the option with the greatest net benefit, so the problem is not with the uncertainty, but with how we (mis)handle it. To the extent that any kind of averaging (computing the expectation of something) is helpful, we stress the gaping contrast between "decision averaging" and "risk averaging." See NRC (1994, p. 173), where the NRC Committee provides an example showing the difference between estimating the average net benefit of each of two sensible actions to evacuate a city where a hurricane might or might not be headed, versus estimating the "weighted average location" between the two cities and evacuating the unpopulated area at that coordinate. John taught us both that expected utility and "the utility of the expectation" are different, back in the days when we would have had to fill a stadium full of mainframe computers to match the processing power of a modern iPhone, and yet it is still not followed today as the truism it is.

[5] It is also possible to set up a system wherein analysts construct both a risk estimate contingent on a single inference option (either a robust default or a compelling alternative), *and* an estimate that eschews this choice but instead considers all plausible models. Decisionmakers could then consider the "truncated" approach to model uncertainty alongside a baroque approach to it.

It may also be somewhat facile for us to posit that competing models should be judged based on their effects on the relative performance of decision options, given that traditionally, risk assessment precedes risk management and the options are often not arrayed until later (for an exhortation that we reverse this ordering, and array the possible control options *before* we begin to estimate the risk under any option, see Finkel, 2011). One intermediate strategy that does not require explicit control options, but that improves upon one-size-fits-all risk averaging, was described in detail in a doctoral dissertation John Evans supervised (Brand 1999), in which the author advocates for the combining of disparate models via an explicit and generally unequal weighting of the decision regret associated with incorrectly choosing one model over another.

And we offer one other cautionary note about "a full treatment of model uncertainty." The main attraction of incorporating model uncertainty rather than relying on defaults is that the former approach may allow decisionmakers and the affected publics to "see the full light of uncertainty." But if incorporating model uncertainty carries the baggage of subjective weighting, delay, and possible "decisions guaranteed to be wrong," and *does not even fully depict uncertainty*, then it may be a marginal improvement with substantial downsides. We suggest here that there is *much* more to model uncertainty than simply supplementing, overturning, or "watering down" precautionary and reasonable defaults with other reasonable interpretations of mode of action, interspecies scaling, and the like.

What would an exhaustive treatment of model uncertainty in risk, benefit, and cost look like, one that could not be criticized for leaving anything out? It would include various model uncertainties that few if any risk and cost-benefit analyses ever consider. Certainly, there are alternative exposure and fate-and-transport models that are rarely considered alongside the traditional ones. Ditto with the way we currently erase most of the uncertainty in the "value of a statistical life"—as the central tendency of many stated-preference experiments *or* revealed-preference studies, but rarely incorporating the interindividual variability in each subject's responses or the model uncertainty that makes it difficult to choose one type of study over the other (Finkel & Johnson, 2018). We believe that the *economic cost* aspects of cost-benefit analysis are also especially handicapped by the tacit and pervasive use of unacknowledged default assumptions without appreciation of model uncertainty. For example, regulatory economics routinely assumes, without explicit mention, that partial-equilibrium cost estimates are good surrogates for over general-equilibrium ones, that technological learning and/or economies of scale are unimportant, that price rises will reduce demand rather than spur demand for substitute goods. (Hazilla & Kopp, 1990; Mannix, 2014).

And even within the realm of dose–response modeling, there are ways in which we censor important uncertainties. For example, suppose that an epidemiology study shows a relative risk (RR) with confidence limits going from 0.8 to 4.0. That is a classic "negative" result because the lower confidence limit is below 1, and so we would never give *any* weight to the alternative possibility that the exposure *does* cause disease, because we can't rule out with confidence that the exposure is inconsequential. Why, however, shouldn't we include the (let's say) 80% chance that the RR is > 1, in our risk/uncertainty estimation? The reason invariably given is that when the p value is larger than 0.05, we cannot rule out the null hypothesis with "anything approaching certainty." Indeed, recent recommendations have been offered (see, e.g., Benjamin et al., 2018) to make the dividing line for "statistical significance" even more stringent, to $p < 0.005$. We are not advocating for either the 0.05 status quo or for a stricter (or a less strict) criterion. We merely point out that either 0.05 or 0.005 amounts to an "anti-conservative default assumption"; the risk analysis system chooses to strongly guard against false positives at this important step in the evaluation of epidemiologic or toxicologic data (Greenland et al., 2016). In at least one important court case (Flue-Cured Tobacco, 1998), an EPA risk assessment for second-hand smoke inhalation was invalidated in large part because the agency *relaxed* the criterion to (in effect) $p = 0.1$ without adequately explaining this departure from convention. Our concerns expressed above about including "fringe" assumptions and giving them expert-derived weight is really no different from the perennial objections to relaxing the p-value threshold and using more of the entire confidence interval on "negative" bioassay and epidemiology results. The only difference, actually, is that substantial momentum is behind the view that "minority" theories of causation or mode of action must be given some weight in analyses and decisions, whereas we are unaware of any serious effort to suggest that a hazard we can be "only" 90% confident is associated with adverse effects might be deemed worthy of attention.

2.4. Decisionmakers Who Refuse to Require VOI Calculations to Guide the Choice between Action and Analysis (and to Guide the Contours of any Further Analysis) Ought to Admit They are Walking Around with Sunglasses on When the Skies are Dark

John Evans has been a strong advocate of the use of VOI analysis for many years. Sometimes he has urged its use in its formal sense, as the expected gain that would come from reducing the uncertainty in estimates of the consequences of alternative choices in a decision. Often, though, John would advocate merely stopping to think more qualitatively about how valuable new information would really be in making choices. He would encourage the enormously useful thought-question "how much would the information have to change my risk estimates in order to make me change my mind?" Often, it is difficult to imagine any experiment or survey or sampling effort that would make a big enough difference to change a choice – it is not that the information has no value, but that it cannot be *used* in a valuable way. This conclusion could result from a situation where the uncertainties in exposure or risk are sufficiently small that further reductions are of little decisional value, or where they are sufficiently large that valuable uncertainty reductions are hard to imagine occurring given constraints on expense, difficulties in measurement, and so on. But, we hasten to add, information can also be of little value to refining choices when it is the choices themselves that are deficient. The tendency of some regulatory analysts to present to their managers a carefully orchestrated set of decision options with one "winner" (sometimes referred to as "Stupid Option A, Stupid Option C, and Brilliant Option B") yields a situation where further information will have no decisional value, but this indicts the choices, not the uselessness of actual knowledge.

John is not the only one who has encouraged the use of VOI information in environmental, health and safety decisions. For example, Committees of the NRC have urged the EPA (National Research Council, 2009) and the Food and Drug Administration (National Research Council/Institute of Medicine, 2011) to expand the use of VOI in making research and information gathering investments. There are a wide range of technical papers and reports that identify ways in which VOI could be applied by agencies (Bates et al., 2015; Dakins, 1999; Keisler, Collier, Chu, Sinatra, & Linkov, 2014; Laxminarayan and Macauley 2012; Mitchell et al., 2013; Yokota, Gray, Hammitt, & Thompson, 2004).

However, VOI, in any formal sense, has not caught on at all in the regulatory world (Gray, 2019). There are both implementation issues and technical challenges that seem to have stymied its use. Perhaps the biggest implementation issue is that in order for VOI to be applied, risk estimates need quantitative estimates of their attendant uncertainty, something regulatory agencies have very rarely developed, especially on the "cost side" (Finkel, 2014a). In addition, any formal decision analytic tool needs to specify *a priori* the options being considered. It is clear that many in decision-making positions are uncomfortable stating the options under consideration in advance. There also seems to be a general belief that uncertainty analysis and VOI are too difficult for decision makers to understand and use appropriately—a view we believe reflects badly on the decision makers, not on the analysts (Finkel & Gray, 2018)!

There are also technical challenges to using VOI in EHS risk decisions. One of the greatest is actually knowing how much information a given data collection event will deliver and by how much it will reduce uncertainty. Uncertainty analysis will always be subject to "unknown unknowns" and truncation of possible models, which means that information might reduce uncertainty more than predicted. Many of the early studies of VOI focused on the expected value of perfect information, but we know that an animal toxicology experiment, exposure assessment, or cost of implementation survey will provide only partial information for an analysis. Estimating how much uncertainty will be reduced with different sources of information is a continuing challenge. Dealing with model uncertainty provides another challenge, since data to effectively rule out, or even greatly change the probability of alternatives may be difficult to acquire. It may also be difficult to know how much it will cost to deliver a specific piece of information. Some information may be generalizable across decisions and would therefore be even more valuable than assumed for a single choice. These and other technical issues, while they need to be considered, are not obstacles to the use of VOI today.

Perhaps a place we can start is with John's question of bounding the magnitude of uncertainty reduction necessary to change a decision. For example, imagine an abandoned hazardous waste site with chemicals identified as carcinogens in the soil but no contamination of groundwater. There are three decision options: Do nothing, put a cap of clean soil above the contaminated dirt, or dig up and remove the contaminated soil. A risk assessment, using default procedures such as a linear dose-response relationship between exposure and cancer risk, and standard dust exposure assumptions, finds the lifetime cancer risk to be 1×10^{-3}. At this risk level, the relatively best remediation choice is to remove the soil. If the risk were 1×10^{-5} the choice would be capping, and if under 1×10^{-6} the choice would be do nothing. Because the remediation options are "lumpy," John's question would have us ask if and how new information could ever move us from one option to another. In this case, if the assumption of linear dose-response is a major source of uncertainty, we would have to think that we could do an experiment that would leave no more than a 1% chance that the true dose–response is linear, in order to change our choice from soil removal to capping. It is highly unlikely that any information gathering would yield this level of precision, so this form of VOI thinking has helped solidify a decision. In other cases, it may be that readily obtainable information, on cost of alternatives or exposure profiles, could indeed matter and more formal analysis would be called for.

Those who make research and data gathering decisions, intended to help guide and improve decisions, but do not use VOI approaches, are likely to squander resources and miss opportunities to shape better choices.

3. CONCLUSIONS

These four themes may seem disparate, but they are interrelated, and all hearken back to the answer John provided in the *Air Pollution Control Association* volume in 1986: "analysis IS useful." Earlier we emphasized that analysis is more useful, perhaps only is useful, when it is done quantitatively, with careful attention to uncertainty and variability, and in ways that allow feasible choices to be compared along multiple dimensions. But more importantly, we suggest that his 1986 title was pointing the field toward a probing examination of what "useful" means. Of course, good analysis has value and utility—but we do not merely want the analysis because it is a tool; we (should) want the *results of our actions* to be "useful."

So perhaps the real question behind the 1986 question was and is "are unanalyzed actions useful?"—and John and we, his students, would say "no." For an easy target, consider the first 24 launches of the Space Shuttle between 1981 and 1986. Arguably, NASA risk managers did not heed the results of the risk assessments done there, and while the launch decisions made before the *Challenger* disaster were useful, they were not optimal (Dalal, Fowlkes, & Hoadley, 1989; NRC, 1988). On the other hand, John set both of us on careers that included substantial time in regulatory agencies, where we can look back on decisions we made affirmatively but also on equally weighty decisions we made by failing to decide, by waiting to express our views until after we could no longer influence policy, or by changing the subject in order to "make" a decision about some other problem (thereby making a decision about the problem we were shunting or punting). So we have learned, from working with John and from life after we "left without the pebble," that whether one sees oneself as an analyst or a decisionmaker (a somewhat arbitrary and unfortunate bifurcation), one needs to be relentless in asking as often as possible "what did I decide *today?*" If the answer is "I decided we weren't ready to decide," *John's work challenges us to then ask, with great humility but also with great urgency, "what are we waiting for, and why?"*

So, all four of the topics we discussed here tell the same story, with variations. Analyzing dose–response data (toxicologic, epidemiologic, or both) to categorize a stressor as "threshold or not" can be valuable, *but only if* knowing which is which is expected to improve an outcome. Using provisional HHRVs can seem deflating, *but only if* refining them is expected to improve an outcome. "Reducing reliance on defaults" can increase real or perceived sophistication, *but only if* doing so is expected to improve an outcome. And if any of these refinements matter, which they certainly often will, some form of VOI analysis is waiting in the wings to provide the answers to questions like these three.

ACKNOWLEDGMENTS

We would like to thank the two anonymous reviewers for very insightful and useful feedback, Lisa Robinson and Jim Hammitt of the Harvard Center for Risk Analysis for convening a workshop in John Evans' honor, and Richard Belzer for use of Fig. 1. Finally, we want to acknowledge the wisdom that John Evans has shared with both of us—see https://sites.google.com/view/tributetojohnevans for a tribute.

REFERENCES

Bates, E. M., Keisler, J. M., Zussblatt, N. P., Plourde, K. J., Wender, B. A., & Linkov, I. (2015). Balancing research and funding using value of information and portfolio tools for nanomaterial risk classification. *Nature Nanotechnology*, *11*, 198–205.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981.

Belzer, R. B. (2012). Risk assessment, safety assessment, and the estimation of regulatory benefits. Mercatus Research, Mercatus Center, George Mason University, Arlington, VA. Oct. 10. 32. Retrieved from https://www.mercatus.org/system/files/RiskAssessment_Belzer_v1-0_2.pdf

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E- J., Berk, R., … Johnson, V. E. .(2018) Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Bevan, R. J., & Harrison, P. T. C. (2017). Threshold and non-threshold carcinogens: A survey of the present regulatory landscape. *Regulatory Toxicology and Pharmacology*, *88*, 291–202.

Bernstein, L., Gold, L. S., Ames, B. N., Pike, M. C., & Hoel, D. G. (1985). Some tautologous aspects of the comparison of carcinogenic potency in rats and mice. *Toxicological Sciences*, *5*(1), 79–86.

Bogen, K. T. (2019). Inflammation as a cancer co-initiator: New mechanistic model predicts low/negligible risk at noninflammatory carcinogen doses. *Dose-Response: An International Journal*, *17*(2), 155932581984783.

Brand, K. P. (1999). *Interpreting bioassays for policy: Analysis of extrapolation uncertainties*. (Sc.D, dissertation, Harvard School of Public Health, Boston, MA, May 1999. On file with authors).

Brand, K. P., Catalano, P. J., Hammitt, J. K., Rhomberg, L., & Evans, J. S. (2001). Limitations to empirical extrapolation studies: The case of BMD ratios. *Risk Analysis*, *21*(4), 625–640.

Brand, K. P., & Finkel, A. M. (2019). A Decision-analytic approach to addressing the evidence about football and CTE. *Seminars in Neurology*, *40*(4), 450–460. Retrieved from https://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0039-1688484

Calabrese, E. J. (2004). Hormesis: From marginalization to mainstream: A case for hormesis as the default dose-response model in risk assessment. *Toxicology and Applied Pharmacology*, *197*, 125–136.

Clewell, R. A., Thompson, C. M., & Clewell, H. J. III (2019). Dose-dependence of chemical carcinogenicity: Biological mechanisms for thresholds and implications for risk assessment. *Chemico-Biological Interactions*, *301*, 112–127.

Cox, L. A. (2008). What's wrong with risk matrices? *Risk Analysis*, *28*(2), 497–512.

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: A handbook for dealing with variability and uncertainty in models and Inputs*. New York: Plenum Press.

Dakins, M. (1999). The value of the value of information. *Human and Ecological Risk Assessment: An International Journal*, *5*(2), 281–289.

Dalal, S. R., Fowlkes, E. B., & Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-*Challenger* prediction of failure. *Journal of the American Statistical Association*, *84*, 945–957.

Evans, J. S. (1986). *Environmental risk management: Is analysis useful?* Pittsburgh, PA: Air Pollution Control Assoc.

Evans, J. S., Gray, G. M., Sielken, R. L., Smith, A. E., Valdezflores, C., & Graham, J. D. (1994). Use of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. *Regulatory Toxicology and Pharmacology*, *20*(1), 15–36.

Finkel, Adam M. (1990). *Confronting Uncertainty in Risk Management: A Guide for Decision-Makers*, Washington, DC: Resources for the Future. https://www.researchgate.net/publication/245769033_Confronting_Uncertainty_in_Risk_Management_A_Guide_for_Decision_Makers.

Finkel, A. M. (2008). Protecting people in spite of—or thanks to—the 'veil of ignorance'. Chapter 17 In R. Sharp, G E. Marchant, & J. Grodsky (Eds.), *Genomics and environmental regulation: Science, ethics, and law* (pp. 290–342). Baltimore, MD: Johns Hopkins University Press.

Finkel, A. M. (2011). Solution-focused risk assessment: A Proposal for the fusion of environmental analysis and action. *Human and Ecological Risk Assessment*, *17*(4), 754–787 (and 5 concurrent responses/commentaries, pp. 788–812).

Finkel, A. M. (2014a). The cost of nothing trumps the value of everything: The failure of regulatory economics to keep pace with improvements in quantitative risk analysis." *Michigan Journal of Environmental and Administrative Law*, *4*(1), 91–156.

Finkel, A. M. (2014b). "EPA underestimates, oversimplifies, miscommunicates, and mismanages cancer risks by ignoring human susceptibility." *Risk Analysis*, *34*(10), 1785–1794.

Finkel, A. M. (2018). Demystifying evidence-based policy analysis by revealing hidden value-laden constraints." In Governance of Emerging Technologies: Aligning Policy Analysis with the Public's Values, Gregory E. Kaebnick and Michael K. Gusmano, eds., *Hastings Center Report*, *48*(S1), S21–S49.

Finkel, A. M., & Gray, G. (2018). Taking the reins: How regulatory decision-makers can stop being hijacked by uncertainty. *Environment Systems and Decisions*, *38*, 230–238.

Finkel, A. M., & Evans, J. S. (1987). Evaluating the benefits of uncertainty reduction in environmental health risk management. *Journal of the Air Pollution Control Association*, *37*(10), 1164–1171.

Finkel, A. M., & Johnson, B. B. (2018). The limits of self-interest: Results from a novel stated-preference survey to estimate the social benefits of life-prolonging regulations. *Environmental Law (Lewis & Clark Law School)*, *48*(3), 453–476.

Flue-Cured Tobacco Co-op. v. US EPA, 4 F. Supp. 2d 435 (M.D.N.C. (1998). U.S. District Court for the Middle District of North Carolina, decided July 17, 1998. Retrieved from https://law.justia.com/cases/federal/district-courts/FSupp2/4/435/2349856/

Freedman, D. A., Gold, L. S., & Slone, T. H. (1993). How tautological are interspecies correlations of carcinogenic potencies? *Risk Analysis*, *13*(3), 265–272.

Golden, R., Bus, J., & Calabrese, E. (2019). An examination of the linear no-threshold hypothesis of cancer risk assessment: Introduction to a series of reviews documenting the lack of biological plausibility of LNT. *Chemico-Biological Interactions*, *301*, 2–5.

Gradient Corp. (2015). *Comments on the Petition to Add n-Propyl Bromide to the List of Hazardous Air Pollutants Regulated under §112 of the Clean Air Act*. May 7, 43 , Retrieved from https://oehha.ca.gov/media/downloads/proposition-65/crnr/comments/0904151bromopropanecomments2.pdf

Gray, G. M. (2019). Workshop organizer: Use of value of information analysis in federal agencies. George Washington University Milken Institute School of Public Health, Washington DC. July 25.

Gray, G. M., & Cohen, J. T. (2012). Rethink chemical risk assessment. *Nature*, *489*, 27–28.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337.350.

Hazilla, M., & Kopp, R. J. (1990). Social cost of environmental quality regulations: A general equilibrium analysis. *Journal of Political Economy*, *98*, 853–873.

IRLG (Interagency Regulatory Liaison Group) (1979). Scientific bases for identification of potential carcinogens and estimation of risks. *Journal of the National Cancer Institute*, *63*, 241–268.

Keisler, J. M., Collier, Z. A., Chu, E., Sinatra, N., & Linkov, I. (2014). Value of information analysis: The state of application. *Environmental Systems and Decisions*, *34*, 3–23

Kovarich, S., Ceriani, L., Gatnik, M. J., Bassan, A., & Pavan, M. (2019). Molecular Informatics. 38. Retrieved from https://zoom.us/j/97646813282

Kratchman, J., Wang, B., Fox, J., & Gray, G. (2017). Correlation of non-cancer benchmark doses in short and long-term rodent Bioassays. *Risk Analysis*, *38*, 1052–1069

Krewski, D., Gaylor, D. W., Soms, A. P., & Szyszkowicz, M. (1993). An overview of the report: Correlation between carcinogenic potency and the maximum tolerated dose: Implications for risk assessment. *Risk Analysis*, *13*(4), 383–398.

Laxminarayan, R., & Macauley, M.K. eds., (2012). *The value of information: Methodological frontiers and new applications in environment and health*. Dordrecht, The Netherlands: Springer Science & Business Media.

Layton, D. W., Mallon, B. J., Rosenblatt, D. H., & Small, M. J. (1987). Deriving allowable daily intakes for systemic toxicants lacking chronic toxicity data. *Regulatory Toxicology and Pharmacology*, *7*, 96–112.

Leisenring, W., & Ryan, L. (1992). Statistical properties of the NOAEL. *Regulatory Toxicology and Pharmacology*, *15*(2), 161–171.

Mannix, B. (2014). Employment and human welfare: Why does benefit-cost analysis seem blind to job impacts?" Chapter in Coglianese, In C. A. M. Finkel & C. Carrigan (Eds.), *Does regulation kill jobs?* (pp. 312). Philadelphia, PA: University of Pennsylvania Press.

Mitchell, J., Pabon, N., Collier, Z. A., Egeghy, P. P., Cohen-Hubal, E., Linkov, I., & Vallero, D. A. (2013). A decision analytic approach to exposure-based chemical prioritization. *PLoS ONE*, *8*(8), e70911.

Moghissi, A. A., Love, B. R., & Straja, S. R. (2012). Linear non-threshold: Separating facts from fiction. *Dose-Response*, *10*(2), 297–305.

Montague, P., & Finkel, A. M. (2007). Two friends debate risk assessment and precaution. Rachel's Democracy and Health News, #920, August 16. Retrieved from http://www.rachel.org/?q=en/newsletters/rachels_news/print/920#Two-Friends-Debate-Risk-Assessment-and-Precaution

Morgan, M. G., & Henrion, M. (1992). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.

National Research Council (1988). *Post-Challenger evaluation of space shuttle risk assessment and management*. Washington DC: National Academy Press.

National Research Council (1994). *Science and judgment in risk assessment*. Washington DC: National Academy Press.

National Research Council (2007). *Scientific review of the proposed risk assessment bulletin from the office of management and budget*. Washington DC: National Academy Press.

National Research Council (2009). *Science and decisions: Advancing risk assessment*. Washington DC: National Academy Press.

National Research Council/Institute of Medicine (2011). *Risk characterization framework for decision-making at the food and drug administration, Chapter 2*. Washington DC: National Academies Press.

Oppenheimer, M., Little, C. M., & Cooke, R. M. (2016). Expert judgement and uncertainty quantification for climate change. *Nature Climate Change*, *6*, 445–451.

Pecquet, A. M., Martinez, J. M., Vincent, M., Erraguntla, N., & Dourson, M. (2018). Derivation of a no-significant-risk-level for tetrobromobisphenol A based on a threshold non-mutagenic cancer mode of action. *Journal of Applied Toxicology*, *38*(6), 862–878.

Pennington, D., Crettaz, P., Tauxe, A., Rhomberg, L., Brand, K., & Jolliet, O. (2002). Assessing human health response in life cycle assessment using ED10s and DALYs: Part 2—noncancer effects. *Risk Analysis*, *22*, 947–963

Schnell, F. (2016). Chemicals, cancer and common sense. Webpage. American Council on Science and Health. Retrieved from https://www.acsh.org/news/2016/04/15/chemicals-cancer-and-common-sense-12094

Schrenk, D. (2018). What is the meaning of 'a compound is carcinogenic'? *Toxicology Reports*, *5*, 504–511.

Slikker, W. Jr., Andersen, M. E., Bogdanffy, M. S., Bus, J. S., Cohen, S. D., Conolly, R. B., … Wallace, K. (2004). Dose-dependent transitions in mechanisms of toxicity: Case studies. *Toxicology and Applied Pharmacology*, *201*, 226–294.

Stelljes, M., Young, R., & Weinberg, J. (2019). 28-day somatic gene mutation study of 1-bromopropane in female Big Blue B6C3F1 mice via whole-body inhalation: Support for a carcinogenic threshold. *Regulatory Toxicology and Pharmacology*, *104*, 1–7.

Tennekes, H. A. (2016). A critical appraisal of the threshold of toxicity model for non-carcinogens. *Journal of Environmental & Analytical Toxicology*, *6*(5), 408

US EPA (2014). *Bisphenol A alternatives in thermal paper*. Retrieved from https://www.epa.gov/sites/production/files/2014-05/documents/bpa_final.pdf)

US EPA (2016). *Contaminant candidate list–4* Retrieved from https://www.epa.gov/ccl/contaminant-candidate-list-4-ccl-4-0)

US EPA (2020). *Risk evaluation for 1-bromopropane (n-Propyl bromide). Office of chemical safety and pollution prevention*. August 2020, EPA #740-R1-8013. Retrieved from https://www.epa.gov/sites/production/files/2020-08/documents/risk_evaluation_for_1-bromopropane_n-propyl_bromide.pdf

Watanabe, K., Bois, F. Y., & Zeise, L. (1992). Interspecies extrapolation: A reexamination of acute toxicity data. *Risk Analysis*, *12*(2), 301–310.

Wagner, W. E. (1999). The triumph of technology-based standards. *Illinois Law Review*, *2000*, 83–113.

Wang, B., & Gray, G. (2015). Concordance of non-carcinogenic endpoints in rodent chemical bioassays. *Risk Analysis*, *35*, 1154–1166.

Wiener, J. B. (2001). Precaution in a multi-risk world. In D. D. Paustenbach (Ed.), *The risk assessment of environmental and human health hazards* (2d ed). Retrieved from https://ssrn.com/abstract=293859 or https://doi.org/10.2139/ssrn.293859

Wignall, J. A., Muratov, E., Sedykh, A., Guyton, K. Z., Tropsha, A., Rusyn, I., & Chiu, W. A. (2018). Conditional toxicity value (CTV) predictor: An *in silico* approach for generating quantitative risk estimates for chemicals. *Environmental Health Perspectives*, *126*(5), 057008. https://doi.org/10.1289/EHP2998

Yokota, F., Gray, G., Hammitt, J. K., & Thompson, K. M. (2004). Tiered chemical testing: A value of information approach. *Risk Analysis*, *24*(6), 1625–1639.