

Invited Commentary

Bounding Analysis as an Inadequately Specified Methodology

Sander Greenland*

The bounding analysis methodology described by Ha-Duong *et al.* (this issue) is logically incomplete and invites serious misuse and misinterpretation, as their own example and interpretation illustrate. A key issue is the extent to which these problems are inherent in their methodology, and resolvable by a logically complete assessment (such as Monte Carlo or Bayesian risk assessment), as opposed to being general problems in any risk-assessment methodology. I here attempt to apportion the problems between those inherent in the proposed bounding analysis and those that are more general, such as reliance on questionable expert elicitations. I conclude that the specific methodology of Ha-Duong *et al.* suffers from logical gaps in the definition and construction of inputs, and hence should not be used in the form proposed. Furthermore, the labor required to do a sound bounding analysis is great enough so that one may as well skip that analysis and carry out a more logically complete probabilistic analysis, one that will better inform the consumer of the appropriate level uncertainty. If analysts insist on carrying out a bounding analysis in place of more thorough assessments, extensive analyses of sensitivity to inputs and assumptions will be essential to display uncertainties, arguably more essential than it would be in full probabilistic analyses.

KEY WORDS: Bayesian analysis; bias analysis; bounding analysis; causal inference; epidemiologic methods; expert elicitation; Monte Carlo analysis; probabilistic assessment; risk analysis; risk assessment; sensitivity analysis; stochastic simulation

1. INTRODUCTION

From a narrow but common viewpoint there are currently only two received modes of probabilistic or statistical inference: frequentist and Bayesian. These two modes have many subtypes and variants, and the boundary between them becomes blurred when one enters the realm of hierarchical (multilevel) modeling (Good, 1983, 1987; Greenland, 2000a; Carlin & Louis, 2000). All the variants depend critically on a sampling model for the variability (both systematic and random) in observations. In the hierarchical framework both modes deploy a model for information about or relations among sampling-model parameters (known

to Bayesians as a “prior distribution,” and known to frequentists by many other names such as “mixing distribution,” “second-stage distribution,” or “penalty function”), and when they adopt the same models for all levels of variation they produce parallel mathematical deductions.

The consequence is that a savvy modeler can produce identical numerical results (e.g., interval estimates) from each model. Of course, the nominal interpretation of these results will differ. For example, a frequentist interval estimate (such as a confidence interval) is defined and evaluated in terms of a coverage rate in a very hypothetical thought experiment in which all the input models are perfectly correct and used for resampling. In contrast, a subjective-Bayesian interval estimate (a posterior interval) is defined in terms of what a hypothetical

*Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, USA; lesdomes@ucla.edu.

observer should bet, given the input models and the data.

As is often the case with competing politicians vying for our votes, these competing interpretations are less far apart than polemics often suggest. At the very least, both start with “assume the following models and distributions” and deduce from there, and both depend on full probability models (even though those models might be infinite-dimensional, as in non- and semiparametric methods). In observational epidemiology, where all the models are subjective speculation, divergence between frequentist and Bayesian inferences can just as well be attributed to differences in the input models (mainly in use of uniform versus more concentrated parameter distributions), rather than to differences in fundamental philosophy (Good, 1987).

On the principle that diversity of viewpoints is beneficial, and given that standard statistics does *not* encompass as much diversity as is often claimed (except by the virtue of rampant misinterpretations of basic statistics such as significance tests), we should welcome any attempt to produce a truly different analytic viewpoint. It seems widely accepted, however, that any such attempt must have logical clarity and consistency, and any application must be consistent with background (subject matter) information. I will argue that the bounding analysis methodology proposed by Ha-Duong *et al.* (2004) fails on both counts, and hence should be rejected in the form presented. My criticisms draw heavily on my work and that of colleagues in epidemiologic risk estimation; due to space constraints, many details are omitted but can be found in the references.

2. FLAWS IN THE BOUNDING ANALYSIS OF HA-DUONG *ET AL.*

The following sections describe three logically independent criticisms of the bounding methodology proposed by Ha-Duong *et al.* The first and most extensive criticism is important in any application, and thus must be repaired before any use of the proposed method, while the impact of the other two would depend on the context. Section 2.4 raises an ambiguity in the causal model underlying the methodology, which could have an impact in special situations.

2.1. Failure to Conceptually Define the Bounds and Deal with Bound Uncertainty

The bounding proposal begins with a conceptual gap, on which it founders thereafter: “The goal of the

analysis is to generate an upper bound on the mortality attributed to the group of poorly characterized factors . . .” (Ha-Duong *et al.*, Section 1.1). The error the authors commit is that of treating this upper bound and input bounds used to derive this bound as sharply defined quantities, without ever offering a definition of what they are. They also do not indicate how to derive input bounds from subject matter, except by expert elicitation, which I will criticize later. This definitional gap is fatal: for a mathematical analysis (like that in Section 3 of Ha-Duong *et al.*) to produce scientifically meaningful outputs, we must have operational definitions of the inputs and reliable means for measuring them.

What is one of their bounds supposed to represent? To appreciate the question, consider a central input to their analysis, the smoking attributable fraction for lung cancer (Ha-Duong *et al.*, Section 2); call it SAF for now. In what seems a nonsequitur, after discussing various estimates of this quantity, Ha-Duong *et al.* offer 0.70 as a lower bound. Later, in Section 4, Ha-Duong *et al.* condition their results on the premise “if one is confident in the bounds assigned to the well-understood risk factors”; presumably this refers to the 0.70, among others. Why not 0.67 or 0.75? Is 0.70 a point that they *are* 100% certain SAF exceeds? If so, how did they come upon that certainty? Or is it something less than certain? If less, by how much? If they are 100% certain SAF exceeds 0.70, are they as certain it exceeds 0.75? If so, why not use 0.75? If they are not as certain about 0.75, why not? The other direction deserves questioning as well: If Ha-Duong *et al.* are not 100% certain SAF exceeds 0.70, are they 100% certain it exceeds 0.67? If not, why not? If so, why not use 0.67? Finally, if the bound is not supposed to be linked to a subjective probability, then what *does* it mean, and what does “confident in the bound” mean?

These are questions about the implications of the background data and subject-matter theory. Ha Duong *et al.* write as if 0.70 is somehow implied by this scientific background, yet no estimate they present from the literature comes close to this bound. They offer no confidence interval, so apparently the bound is not based on considering random error. Nor do they offer any attempt to quantify the impact of the sources of bias, such as the uncontrolled confounding, the nonrepresentativeness of the source cohorts, the large measurement errors that must be present, or the coarsening errors from use of such crude age and exposure categories. Each of these problems is a source of uncertainty.

Instead of quantifying potential problems, Ha-Duong *et al.* either make qualitative dismissals or else fail to mention them. As an example, they say “other potential confounders . . . may not be a large problem because . . .”. What kind of confidence should we assign this claim, and how much uncertainty is omitted by ignoring this and other potential problems? Did Ha-Duong *et al.* quantify and factor in that uncertainty when setting their bounds? Apparently not. Other problems such as measurement error and excessive coarsening are not discussed, yet can have an order-of-magnitude larger impact on results than the problems Ha-Duong *et al.* do discuss (Greenland, in press; Poole & Greenland, 1997).

It appears instead that Ha-Duong *et al.* rely on expert consensus as an indication of both precision and validity of published risk estimates. This reliance might not be too misleading for smoking-lung cancer relation, given that the latter is one of the most well-studied relations in the history of epidemiology. But few relations are in that category; radon and lung cancer is certainly not. Hence, to be credible, a general risk-assessment methodology must have a means to explicitly incorporate methodologic uncertainties into the analysis. I could find no such means in the proposal or example of Ha-Duong *et al.*

So, again, what is the rationale for using 0.70? Again, why not 0.67 or 0.75? And again, what confidence are we supposed to place in this bound? I believe that the 0.70 (as well as the 0.95 upper bound they offer) represents in large measure digit preference, as arises from demanding a sharp answer to an ill-defined question. Had modern society evolved into using a base-12 number system, Ha-Duong *et al.* would have offered 0.67 or 0.75 (8/12 or 9/12, expressed as 0.80 or 0.90 in base 12). Given that subsequent analyses in Ha-Duong *et al.* depend on the remainder of $1 - 0.70 = 0.30$, the range for digit preference is not a negligible proportion of the input. Furthermore, as will be argued later, bounding analysis is more sensitive to these input bounds than probabilistic analyses. It thus appears that the input bound of 0.70 is as much a function of digit preference as subject matter; sensitivity of outputs to this preference is a clue that something very arbitrary and poorly thought out underlies the proposed analysis.

Expert elicitation does not address the above issues. If the bound is from an expert, we should ask the same pointed questions: What kind of confidence do you have in this bound? Why that bound, and not another nearby? Elicitations can be severely affected by digit preferences and cognitive biases (Kahneman

et al., 1982; Gilovich *et al.*, 2002). Worse, elicitation can incorporate personal biases of the expert, either academic (previous commitment to a value for the bounded quantity) or economic (conflict of interest). Saying that they are “expert judgments” begs these questions. Because the same problems arise in probability elicitation, I will return to these issues after discussing probabilistic assessments.

In summary, the core proposal of Ha-Duong *et al.* is nothing more than an algorithm for operating on inputs, and gives no meaning to those inputs or the output. By offering neither a precise definition of the input bounds nor their scientific derivation, the proposed bounding analysis first injects and then ignores potentially large sources of uncertainty and bias into the analysis, and leaves unclear the meaning of the outputs. No subsequent mathematical operation can compensate for these deficiencies. These criticisms require more than clarification of the proposed methodology; they require an explicit conceptual foundation for the inputs, and a means of deriving inputs from the literature.

2.2. Unjustified Assumptions

The mathematical development in Section 3 of Ha-Duong *et al.* can be criticized for the assumptions it injects. In any risk analysis, a sharp mathematical assumption represents a source of uncertainty that the analyst has chosen to ignore. A credible analysis presents scientific reasons to justify those choices, or evidence that the results are insensitive to credible assumption violations. Equation (2) of Ha-Duong *et al.* assumes independence of the exposures, an assumption with neither empirical nor theoretical support in the application; in fact, the assumption is not credible in the application, as the authors concede in a footnote. More generally, unhealthy behaviors, environments, and occupations tend to cluster (reflecting social-class and education gradients). It thus is a mistake to build a method around this independence assumption, especially when one can see from the formula that results could be highly sensitive to dependencies.

The development also assumes absence of three-way or higher interactions; the only justification given is absence of information. One should recall that absence of evidence is not evidence of absence, however; in fact, the interactions being assumed away could be large. Unfortunately, Ha-Duong *et al.* provide no clue as to whether their results are sensitive to their assumption of interaction absence. Finally,

the development in Section 3 invokes a variant of Laplace's Principle of Insufficient Reason, saying it is used in Section 4 to estimate the output bounds. I could find nothing in the article that pinpoints where and how the Principle was used to derive the bounds in Section 4. Regardless, to justify the use of the Principle in a real application one needs to demonstrate that indeed background information was accounted for and exhausted before the Principle is applied. Ha-Duong *et al.* do not do this. Their use of the Principle is not just another source of uncertainty, but a potential source of bias insofar as the Principle invokes use of a uniform distribution, which is an extreme distribution in the space of credible distributions, and is almost never epidemiologically plausible in light of background information (Greenland, 1998a).

2.3. Inferential Deficiencies

Let us for the moment forget the above criticisms and pretend (as do Ha-Duong *et al.*) that the inputs to the proposed method are logically precise, scientifically measurable, and methodologically sound, and that the assumptions made in the mathematical analysis are contextually justifiable. Even with this pretense, the inferences drawn by Ha-Duong *et al.* in Section 4 do not follow logically from their outputs. This is in part because of a new logical gap, and in part because the meaning of the output bounds is ambiguous, as with the inputs.

Of other assessments, Ha-Duong *et al.* state that "if the assessments project the lung cancer risk... from these pollutants ('other' environmental factors, X) to be in excess of 3.2% of the annual lung-cancer mortality, the assumptions of the models should be reexamined and the upper bound on the resulting estimate constrained." This inference is a complete nonsequitur. To see this, let us call the risk being bounded the X-risk. First, Ha-Duong *et al.* are assuming that their upper bound of 3.2% on this risk is a gold standard. Nothing could be further from the truth. Any disagreement between Ha-Duong *et al.*'s upper bound and another assessment could be due to faulty inputs or execution in Ha-Duong *et al.*'s analysis. Unaccounted-for biases (as well as random errors) in the data used to form the inputs would ripple through to distort Ha-Duong *et al.*'s output (Maclure & Schneeweiss, 2001). "Garbage in, garbage out" is a watch-phrase that applies to every assessment, including a bounding analysis. A competing assessment is just that: a competitor, possibly with better in-

puts and superior handling of problems such as bias. Given these possibilities, the competitor should not be judged deficient just because it conflicts with the bounding analysis.

Second, as with the input bounds, Ha-Duong *et al.* give no sense of precision or reliability for their output bound of 3.2%. It is not even clear what this bound means: Is it supposed to represent a point that we are 100% certain or confident is above the X-risk is below 3.2%? That is how they treat it in Section 4. How is such certainty justified given the inevitable uncertainty in the inputs to and assumptions in the bounding analysis? If the output bound corresponds to anything less than 100% certainty, how much less? Why use that particular less-than-100% level of certainty as a cutoff point? If we are not to invoke probabilistic concepts of certainty or confidence, how are we supposed to interpret the bound?

In a complementary fashion, the output from any competing assessment will (or should) have some uncertainty attached. At what point do we say that the competitor conflicts with the bounding analysis? Do we construct a test comparing the two outputs? If so, at what critical level do we reject the hypothesis that the two assessments are compatible, given the uncertainties involved? And how do we account for the fact that the bounding analysis and the competing assessment will be based on much the same data, and so the two will be correlated? Ha-Duong *et al.*'s comparison of their upper bound with a published point projection is naïve at best. While not likely to mislead if the point falls within the bounds, it can be very misleading otherwise: a valid comparison of two analyses cannot be based on just looking at whether bounds overlap with point estimates or other bounds; for example, two 95% confidence intervals may overlap even though the difference between the two estimates they surround is significant at the 0.05 level.

2.4. Clarifying the Underlying Causal Model

Ha-Duong *et al.* say that "there is a considerable debate over the meaning of causation in epidemiology." I think that this is a severe overstatement. Among statisticians there has been debate between a growing number that favor a potential-outcome (counterfactual) view of causation and a few opponents (Greenland, 2000, 2004; Maldonado & Greenland, 2002; Dawid, 2000). Among researchers actually developing, teaching, and applying practical causal-analysis methods, potential-outcome models

have been in use for over 80 years (Neyman, 1923), and have become textbook material (e.g., Berk, 2003; Freedman *et al.*, 1997; Newman, 2001; Gelman *et al.*, 2003; Rosenbaum, 2002; Van Der Laan & Robins, 2003). Furthermore, the general counterfactual concept on which they are based has long been established in epidemiology; e.g., see MacMahon and Pugh (1967), as well as Rothman and Greenland (1998, Chs. 2 and 4), Parascandola and Weed (2001), and Greenland (2004) for further historical as well as current citations. Models that were initially perceived as successful alternatives (e.g., causal diagrams and structural equations) have since been shown to be isomorphic to potential-outcome models, and thus not alternatives after all (Pearl, 2000; Greenland & Brumback, 2002); the remaining proposals have foundered on hidden circularities and have failed to generate acceptable data-analytic procedures (Pearl, 2000; Maldonado & Greenland, 2002).

To describe the potential-outcome models, suppose we wish to study the effect of an intervention variable X with potential values (range) x_1, \dots, x_J on a subsequent outcome variable Y defined on an observational unit or a population. One then assumes that there is a baseline vector of *potential outcomes* $\mathbf{y} = (y(x_1), \dots, y(x_J))'$, such that if $X = x_j$ then $Y = y(x_j)$; this vector is simply a mapping from the X range to the Y range for the unit. To say that intervention x_i causally affects Y relative to intervention x_j then means that $y(x_i) \neq y(x_j)$; and the effect of intervention x_i relative to x_j on Y is measured by $y(x_i) - y(x_j)$ or (if Y is strictly positive) by $y(x_i)/y(x_j)$. Under this theory, assignment of a unit to a treatment level x_i is simply a choice of which coordinate of \mathbf{y} to attempt to observe; regardless of assignment, the remaining coordinates are treated as existing pretreatment covariates on which data are missing (Rubin, 1978). Formally, if we define the vector of potential treatments $\mathbf{x} = (x_1, \dots, x_J)'$, with treatment indicators $r_i = 1$ if the unit is given treatment x_i , 0 otherwise, and $\mathbf{r} = (r_1, \dots, r_J)'$, then the actual treatment given is $x_a = \mathbf{r}'\mathbf{x}$ and the actual outcome is $y_a \equiv y(x_a) = \mathbf{r}'\mathbf{y}$; the remaining $y(x_i)$ are the counterfactuals. Viewing \mathbf{r} as the item-response vector for the items in \mathbf{y} , causal inference under potential outcomes can be seen as a special case of inference under item nonresponse in which $\sum_i r_i = 0$ or 1, i.e., at most one item in \mathbf{y} is observed per unit (Rubin, 1991).

The theory extends to stochastic outcomes by replacing the $y(x_i)$ by probability mass functions $p_i(y)$ (Greenland, 1987; Robins, 1988; Greenland *et al.*,

1999), so the mapping is from X to the space of probability measures on Y . This extension is embodied in the “set” or “do” calculus for causal actions developed for artificial-intelligence research (Pearl, 2000). The theory also extends to continuous X by allowing the potential-outcome vector to be infinite-dimensional with coordinates indexed by X , and with components $y(x)$ or $p_x(y)$.

On the matter of causation, Ha-Duong *et al.* give only one epidemiologic citation: Parascandola and Weed (2001). The latter authors note that nearly all epidemiologic proposals incorporate counterfactual elements, and hence require potential outcomes; the main point of debate is whether and how one incorporates probabilistic elements into each proposal. Traditionalists favor restricting those elements to the sampling model, separated from the underlying causal model, and argue that thinking of deterministic potential outcomes facilitates development of mechanistic hypotheses. The opposing view (sometimes labeled “black-box”) regards the actual systems under study as too complex to be modeled adequately in deterministic terms, and favors introducing stochastic potential outcomes (Greenland, 1987; Robins & Greenland, 1989; Greenland *et al.*, 1999; Parascandola & Weed, 2001). This shift to stochastic potential outcomes can affect the choice of statistical procedure, and in particular can increase final uncertainty (Robins, 1988). The black-box viewpoint is supported by the success of highly stochastic modeling in prediction problems (Breiman, 2001), although a pragmatic view would regard both types of models as useful tools with differing domains of utility.

Like traditional and modern epidemiologic views (e.g., MacMahon & Pugh, 1967; Rothman & Greenland, 1998), Ha-Duong *et al.*'s proposal appears to implicitly use potential outcomes, especially in its discussion of partitioning and interactions. Adoption of potential outcomes has statistical implications for confounding identification and control, especially in an application that uses odds ratios or rate ratios as measures of effect for a common disease (Greenland, 1987; Rothman & Greenland, 1998, Ch. 4; Greenland *et al.*, 1999). Furthermore, Ha-Duong's sharp partitioning of case numbers among risk-factor combinations implies the potential outcomes are deterministic, and so is subject to criticisms of that view (Parascandola & Weed, 2001). This is not a fatal objection (in practice to date, most potential-outcome models have been deterministic), but it is an aspect of their approach that needs acknowledgment.

3. PROBABILISTIC RISK ASSESSMENT VERSUS BOUNDING ANALYSIS

It would seem that Ha-Duong *et al.* propose their bounding analysis as an alternative to probabilistic risk assessments, or even as a standard to judge the latter. I do not see the need for such an alternative. On the contrary, I think the discipline imposed by having to specify a full probability model is an important brake on the vague hand waving that often substitutes for critical thinking in model specification and output interpretation. Furthermore, I would judge bounding analysis against probabilistic assessment methods, for the latter are far more well established and understood (e.g., Eddy *et al.*, 1992; Saltelli *et al.*, 2000), and are closely related to even more well-studied statistical methods for nonrandom exposure and missing-data mechanisms (Gelman *et al.*, 2003; Little & Rubin, 2002; Robins *et al.*, 1999).

Two types of probabilistic assessments, Bayesian assessment and Monte Carlo sensitivity analysis (MCSA), depart from traditional statistical inference by starting with a model that includes nonidentified sensitivity or bias parameters. Hence both begin with a bias modeling, a process conspicuously missing from the bounding proposal of Ha-Duong *et al.* Full Bayesian assessment requires specification of a joint prior for all unknown parameters, and in turn supplies posterior distributions of target parameters (Leamer, 1974, 1978; Eddy *et al.*, 1992; Little & Rubin, 2002; Gelman *et al.*, 2003). MCSA—a variant of familiar stochastic simulation methods (Morgan & Henrion, 1990; Vose, 2000)—requires priors for the nonidentified parameters only, and supplies simulated distributions of estimates “corrected” for biases and random errors.

The MCSA can be viewed as an approximation to a semi-Bayesian analysis and also as an extension of traditional sensitivity analysis (Greenland, 2001, 2003, in press). In traditional sensitivity analysis, bias parameters are varied in a systematic fashion, which can produce misleading range effects (Greenland, 1998b); in MCSA those parameters are instead drawn from a prior distribution. Because it is easily performed with popular software and can be based on familiar bias-sensitivity formulas, MCSA has been promoted for routine epidemiologic applications (Lash & Silliman, 2000; Powell *et al.*, 2001; Greenland, 2001, in press; Lash & Fink, 2003; Phillips, 2003; Steenland & Greenland, 2004). Nonetheless, Bayesian bias analyses have begun to appear as well (Graham, 2000; Gustafson, 2003; Greenland, 2003; Steenland & Greenland, 2004).

From a Bayesian viewpoint, the bounding analysis of Ha-Duong *et al.* entails mishandling of crucial uncertainties. Returning to my criticism of input bounds, an attempt to honestly answer the question of “why 0.70 rather than 0.67 or 0.75?” will lead one to face the fuzziness of the bound concept as well as the location of the bound. Suppose this criticism is answered by declaring the goal is to specify a greatest lower bound L that one is 100% certain falls below the parameter at issue. One will then have to face the uncertainty or dispute among experts about the value of L to specify. If one could poll the experts one could plot a cumulative distribution of L , and use that to shape the lower tail of a prior distribution. One could take the analogous approach to the least upper bound U . Although this approach would not imply a uniform density between the first 100th percentile of the L s and the first 0th percentile of the U s, uniformly filling in the gap between the highest L and lowest U would approximate averaging over the expert densities if the individual densities are relatively flat between their bounds. In the extreme case of just one expert very certain of his or her bounds, this approach would reduce to using a uniform density between the bounds. The final MCSA or Bayesian outputs would not be very sensitive to small changes in the bounds, given that the mass shifts involved would be small. As used by Ha-Duong *et al.* bounding analysis seems to correspond to placing 50% subjective prior mass at each of the uncertain bounds, which turns the bounds into leverage points, and thus grossly inflates sensitivity to bound location—a sensitivity that would at the very least need to be explored as part of a bounding analysis.

4. EXPERT ELICITATION VERSUS LITERATURE ANALYSIS

In problems with which I am familiar, I would be loathe to rely on expert elicitations, whether for bounds or full priors. In my experience, epidemiologic experts often possess highly distorted views of the literature on which they are supposedly expert, freely mixing unsupported impressions with statistical misinterpretations. They often believe a trend is monotone because a trend test (which *assumes* monotonicity) is “statistically significant” (Maclure & Greenland, 1992). When summarizing a literature, they often believe no association is present because most studies are not “statistically significant,” even when this is just an artifact of low power of individual studies, and they often believe the literature is

conflicting because some studies are “significant” and others are not, even when this is just an artifact of power differences; see Greenland *et al.* (2000a) for a meta-analysis that revealed both phenomena. When done well (with proper accounting for heterogeneity and publication bias, as needed), meta-analysis can expose such literature misinterpretations and provide a far more valid and reliable source of input to risk assessment than would an expert elicitation.

Direct inspection of relevant literature can also uncover much useful data (the best prior information) for estimating biases in published study results (Greenland, 2003; Steenland & Greenland, 2004), whereas expert judgments about methodologic biases tend to be highly prejudiced (depending on whether the bias in question would apply to the expert’s own studies, or would move results toward or away from his or her favored answer). Even when without prejudice, bias judgments are often based on incorrect or misleading heuristics, as Jurek *et al.* (2004) found regarding the oft-cited and often false rule that non-differential misclassification produces bias toward the null.

To address the problems with elicitation, it appears that Ha-Duong *et al.* would gather and synthesize relevant literature and present it to the experts for review before doing elicitation. While I have little doubt that this would improve the elicitation, I strongly doubt that it would address most of the cognitive problems described above, such as misinterpretation of lack of statistical significance. Although resource intensive, I see no way around a meta-analytic approach if one wants reliable and valid inputs to a risk analysis; the role of experts is then to conduct or review such an analysis, and interpret the findings in context.

On a more specific issue, the authors propose eliciting expert opinions regarding two-way interactions. With rare exceptions (smoking and radon perhaps being one), the epidemiologic evidence on such interactions is so imprecise that not even a direction can be reliably inferred. Because of their large number ($n(n - 1)/2$ interactions among n factors) and great instability (typically no less than twice the standard error of main effects), published interaction estimates are subject to especially severe multiple-comparison artifacts and publication bias; they are also more vulnerable to biases from data sparsity (Greenland *et al.*, 2000b). I would not expect any expert elicitation to account for these problems, even when based on a literature synthesis, which is to say I would not trust such an elicitation to be better than noise. Even in the

rare case that the evidence regarding an interaction is reliable and valid, one would again be better off deriving inputs from direct analysis of available data, bearing in mind the high likelihood of the aforementioned biases in the published literature.

5. CONCLUSION

It is an open question whether one can break from a full probabilistic framework and still produce a logically sound and scientifically sensible mode of reasoning under uncertainty. Ha-Duong *et al.*’s failure to do so might be evidence in favor of a negative answer. Still, it would be interesting to see attempts to clarify or repair their methodology. At the very least, such methodologic proposals test the hypothesis that full (not just interval) probability models are essential to rational decision and prediction. This full-probability hypothesis is received wisdom (if often implicit) in the statistics literature, which is largely devoted to proposing new probability models and deducing consequences and procedures from them.

A received hypothesis merits refutation attempts, even if it seems obviously correct; note how physicists never cease testing received theories such as general relativity. Thus, despite my strong criticisms of what Ha-Duong *et al.* have proposed and my warnings against using it, I would also strongly encourage these authors to repair or replace (not just defend) their proposal, and so provide a stronger challenge to full-probability approaches. This repair will require considerable conceptual clarification of both input and output bounds, along with far more extensive sensitivity analyses than Ha-Duong *et al.* supply. In the meantime, I again warn that bounding analysis is no substitute or criterion for probabilistic assessment.

REFERENCES

- Berk, R. A. (2003). *Regression Analysis: A Constructive Critique*. Thousand Oaks: Sage.
- Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, *16*, 199–231.
- Carlin, B., & Louis, T. A. (2000). *Bayes and Empirical-Bayes Methods for Data Analysis*, 2nd ed. New York: Chapman and Hall.
- Dawid, P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*, *95*, 407–448.
- Eddy, D. M., Hasselblad, V., & Schachter, R. (1992). *Meta-Analysis by the Confidence Profile Method*. New York: Academic Press.
- Freedman, D., Pisani, R., & Purves, R. (1997). *Statistics*, 3rd ed. New York: Norton.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. New York: Chapman and Hall/CRC.

- Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.
- Good, I. J. (1983). *Good Thinking*. Minneapolis: University of Minnesota Press.
- Good, I. J. (1987). Hierarchical Bayesian and empirical Bayesian methods (letter). *American Statistician*, *41*, 92.
- Graham, P. (2000). Bayesian inference for a generalized population attributable fraction. *Statistics in Medicine*, *19*, 937–956.
- Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analysis. *American Journal of Epidemiology*, *125*, 761–768.
- Greenland, S. (1998a). Probability logic and probabilistic induction. *Epidemiology*, *9*, 322–332.
- Greenland, S. (1998b). The sensitivity of a sensitivity analysis. In *1997 Proceedings of the Biometrics Section*. Alexandria, VA: American Statistical Association, pp. 19–21.
- Greenland, S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association*, *95*, 286–289.
- Greenland, S. (2001). Sensitivity analysis, Monte-Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Analysis*, *21*, 579–583.
- Greenland, S. (2003). The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association*, *98*, 47–54.
- Greenland, S. (2004). An overview of methods for causal inference from observational studies. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*. New York: Wiley.
- Greenland, S. (in press). Multiple-bias modeling for observational studies (with discussion). *Journal of the Royal Statistical Society, Series A*.
- Greenland, S., & Brumback, B. A. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, *31*, 1030–1037.
- Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*, 29–46.
- Greenland, S., Sheppard, A. R., Kaune, W. T., Poole, C., & Kelsh, M. A. (2000a). A pooled analysis of magnetic fields, wire codes, and childhood leukemia. *Epidemiology*, *11*, 624–663.
- Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000b). Problems from small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*, *151*, 531–539.
- Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology*. New York: Chapman and Hall.
- Ha-Duong, M., Casman, E. A., & Morgan, M. G. (2004). Bounding poorly characterized risks: A lung cancer example. *Risk Analysis*, *24*, 1071–1083.
- Jurek, A., Maldonado, G., Church, T., & Greenland, S. (2004). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *American Journal of Epidemiology*, *159*, 572.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.
- Lash, T. L., & Fink, A. K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data. *Epidemiology*, *14*, 451–458.
- Lash, T. L., & Silliman, R. A. (2000). A sensitivity analysis to separate bias due to confounding from bias due to predicting misclassification by a variable that does both. *Epidemiology*, *11*, 544–549.
- Leamer, E. E. (1974). False models and post-data model construction. *Journal of the American Statistical Association*, *69*, 122–131.
- Leamer, E. E. (1978). *Specification Searches*. New York: Wiley.
- Little, R. J. A., & Rubin, D. A. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.
- Maclure, M., & Greenland, S. (1992). Tests for trend and dose-response: Misinterpretations and alternatives. *American Journal of Epidemiology*, *135*, 96–104.
- Maclure, M., & Schneeweiss, S. (2001). Causation of bias: The episode. *Epidemiology*, *12*, 114–122.
- MacMahon, B., & Pugh, T. F. (1967). Causes and entities of disease. In D. W. Clark & B. MacMahon (Eds.), *Preventive Medicine* (pp. 11–18). Boston: Little, Brown.
- Maldonado, G., & Greenland, S. (2002). Estimating causal effects (with discussion). *International Journal of Epidemiology*, *31*, 421–438.
- Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York: Cambridge University Press.
- Newman, S. C. (2001). *Biostatistical Methods in Epidemiology*. New York: Wiley.
- Neyman, J. (1923). Sur les applications de la thar des probabilités aux expériences Agaricales: Essai des principe (English translation of excerpts Dabrowska, D., and Speed, T. (1990). *Statistical Science*, *5*, 463–472.)
- Parascandola, M., & Weed, D. (2001). Causation in epidemiology. *Journal of Epidemiology and Community Health*, *55*, 905–912.
- Pearl, J. (2000). *Causality*. New York: Cambridge.
- Phillips, C. V. (2003). Quantifying and reporting uncertainty from systematic errors. *Epidemiology*, *14*, 459–466.
- Poole, C., & Greenland, S. (1997). How a court accepted a possible explanation. *American Statistician*, *51*, 112–114.
- Powell, M., Ebel, E., & Schlossel, W. (2001). Considering uncertainty in comparing the burden of illness due to foodborne microbial pathogens. *International Journal of Food Microbiology*, *69*, 209–215.
- Robins, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, *7*, 773–785.
- Robins, J. M., & Greenland, S. (1989). The probability of causation under a stochastic model for individual risks. *Biometrics*, *46*, 1125–1138.
- Robins, J. M., Rotnitzky, A., & Scharfstein, D. O. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran & D. A. Berry (Eds.), *Statistical Models in Epidemiology* (pp. 1–92). New York: Springer-Verlag.
- Rosenbaum, P. (2002). *Observational Studies*, 2nd ed. New York: Springer-Verlag.
- Rothman, K. J., & Greenland, S. (1998). *Modern Epidemiology*, 2nd ed. Philadelphia: Lippincott.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.
- Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, *47*, 1213–1234.
- Saltelli, A., Chan, K., & Scott, M. (Eds.) (2000). *Mathematical and Statistical Methods for Sensitivity Analysis* (pp. 275–292). New York: Wiley.
- Steenland, K., & Greenland, S. (2004). Monte-Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology*, *160*, 384–392.
- Van Der Laan, M., & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
- Vose, D. (2000). *Risk Analysis*. New York: John Wiley and Sons.