An Administrative View of Model Uncertainty in Public Health

Clark D. Carrington

An Administrative View of Model Uncertainty in Public Health

Clark D. Carrington[*]

Introduction

For purposes here, a model is a mathematical statement of a theory or hypothesis, that may or may not have a verbal counterpart. However, for the most part, the discussion will also apply to theories which are formulated verbally. Thus, when talking about model uncertainty, we are largely concerned with what might also be called "theoretical" or "interpretational" uncertainty. Models are required to infer answers where facts are unavailable. Model uncertainty arises when two or more alternative interpretations give rise to different predictions.

In health risk assessment, the best known example of model uncertainty is the use of a model to extrapolate the frequency of cancer occurrence in a population from high to low doses. However, in addition to relating a dose to an effect, model uncertainty can arise at virtually every turn of an assessment. For example, in addition to relating a dose to an effect, alternative plausible interpretations may arise in judging the shape of frequency distributions used to describe a parameter in a population, the interpretation of dose-independent effects, or in the extrapolation from one species to another.

In a sense, model uncertainty may be viewed as an instance of what is known in philosophical literature as the "problem of induction" — any generalization drawn from previous experience goes beyond experience and may prove to be incorrect in the face of new experience.[1] Theories will always be constructed which predict, or posit the occurrence of virtually any future event. To use a classic example, one can propose that an object is "grue" rather than blue,

*   Dr. Carrington is employed at the Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration. He received his B.A. (Human Behavior and Institutions) from the University of Chicago and his Ph.D. (Pharmacology) from Duke University. Email: cdc@fdacf.ssw.dhhs.gov.
1   Bertrand Russell, A History of Western Philosophy 659–74 (1946).

where grue is defined to be blue until the year 2000 and green thereafter. Both the "blue" theory and the "grue" theory are consistent with experience, yet yield different predictions.

Figure 1
Model Uncertainty
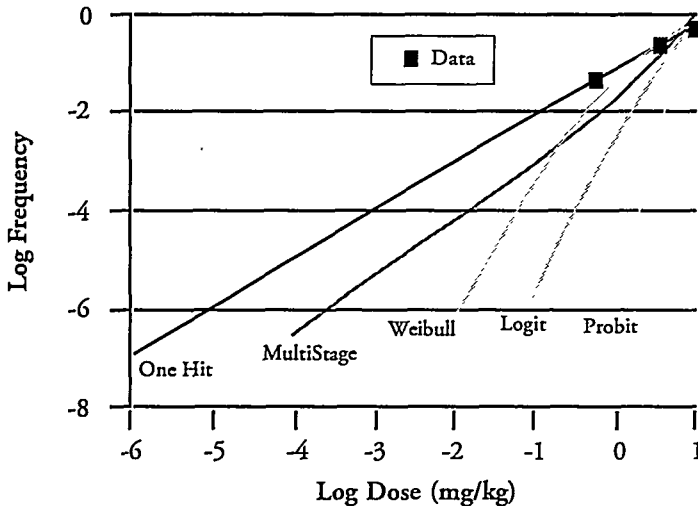
Thus, without some constraint on the choice of theories to be used, virtually anything can be predicted from anything, and all data and information is rendered useless for guiding a decision. In the process of developing information which is intended to guide decisions, constraints *are* placed. In fact, there are several methods for doing so. The remainder of this discussion will compare those methods.

Methods for Accommodating Model Uncertainty

No Model Methods

The first class of methods for dealing with model uncertainty are those that do not acknowledge the use of models at all. This is accomplished by designing a process that produces a decision rather than a prediction. With no prediction, there is no need for an inference, a model, or a question.

• *Technocracy.* The first approach to dealing with model uncertainty, which shall be referred to as technocracy, involves assigning an individual expert the entire responsibility for resolving the impact of

uncertainties arising from alternative interpretations of the decision. Because there is no objective procedure, this is not a "method" in a scientific or logical sense at all. This method is "simpler" from an administrative point of view, not because the issues that become apparent with a more explicit treatment are unimportant, but because the object of the present discussion is hidden. This may be an impetus for preferring this method precisely because model uncertainty is prevented from becoming the focus of an argument. Thus, the principal difference between subjective and explicit methods of inference lies in the appearance of model uncertainty as a topic of discussion.

Often, decisions may be made by committees of experts rather than individuals. Unless model uncertainty becomes a topic for discussion, this may still be viewed as a technocratic exercise — the process necessarily centers on discussion of the action taken, rather than any underlying belief.

• *Default factors.* The oldest and most widely used technique for dealing with model uncertainty in the context of regulatory toxicology involves the application of a fixed arbitrary factor whenever an instance of model uncertainty (or in some cases, statistical uncertainty) is encountered. For instance, factors of 10 are often applied when it is necessary to make statements about large populations based on the observations of a few individuals, about humans based on the observations of rodents, or about long-term exposures based on the observations from short-term exposures. Judging from appearances, the general rule is to divide by 10 upon each instance of model uncertainty. Because selection of the factor is not dependent on the state of the evidence, it is clear that the application of safety or uncertainty factors is not really a method for evaluating model uncertainty. Rather, it is a technique for avoiding the issue by not making a prediction.

Default factors were originally designed for setting regulatory levels.[2] However, the levels are also sometimes represented as factual statements, i.e. thresholds.[3] In this circumstance, the default factor approach may be thought of as a default model (see below), rather than

---

[2] Arnold J. Lehman, O. Garth Fitzhugh, *100-Fold Margin of Safety*, 18 Ass'n Food Drug Officials Q. Bull. 33 (1954).

[3] Michael L. Dourson, Jerry F. Stara, *Regulatory History and Experimental Support of Uncertainty(Safety) Factors*, 3 Reg. Pharm. Tox. 224 (1983).

a decision-producing policy. Similarly, default factors are sometimes represented as instruments of a technocratic exercise (see above) – in which case the factors are merely window dressing for a decision that has already been made. They cannot, however, be considered to be both procedures and products of technocratic judgment as these concepts are antithetical. However, it is possible to mix in some of each to create a hybrid.[4] For instance, a factor of 3 may be used instead of 10 because, on the basis of expert judgment, the degree of model uncertainty is smaller. The process is still semi-procedural to the extent that the factors must follow half-orders of magnitude.

Single Model Methods

Single model methods acknowledge the use of models in making predictions. However, because only a single model is employed, model uncertainty does not appear in the result.

• *The default model.* Another common technique for evaluating data in the context of public health is to employ an explicitly specified model. This technique has arisen out of the acknowledgment that the model employed may be a determinant in making the prediction. However, in spite of the fact that there may be multiple plausible interpretations of the data, a default model treatment of model uncertainty resolves the issue, by dictating one of the many models available as the basis for the prediction, irrespective of the state of the evidence. The default model approach deals with model uncertainty by suppressing it. The best known example of default model methodology in regulatory toxicology is the employment of the linear multistage model in cancer risk assessment.

In theory, one can depart from the default model if the evidence is sufficient. However, the conditions under which a departure might be justified are never specified and are rarely discussed, and in practice, almost never occur. Because the default model is not chosen on the basis of the available data, it is singled out from competing in the face of the critique to which the alternative models are subjected. It may therefore be selected even if it is contrary to the available data.

---

[4] *See, e.g.,* Robert L. Maynard et al., *Setting Air Quality Standards for Carcinogens: An Alternative to Mathematical Quantitative Risk Assessment,* 14 Hum. & Exper. Toxicol. 175 (1995).

• *The best model.* A single model could be selected from available alternatives by explicitly formulated criteria that define the "best" model. Somewhat surprisingly, this method has not been employed in regulatory toxicology, although it appears to be the primary technique recently advocated by a National Research Council committee.[5] Perhaps the main reason that this method has not found favor is that there may often be very little to choose from between the best model and the alternatives. Thus, relatively small changes in the evidence may result in rather large consequences to assessment outcome.

Model Weighting

Model weighting methods are intended to convey to decision makers the extent of uncertainty arising from model selection. This requires model validity or correctness to be treated as a matter of degree by assigning relative probabilities to each alternative. This purely epistemic use of the term "probability" differs from the interpretation employed in the discipline of statistics.[6] Instead of describing a measurable frequency, it is intended to provide a measure of the degree of belief or degree of credibility associated with a theoretical statement.

• *Weighting by experts.* By interviewing experts in the field studied, alternative models or theories may be identified and assigned relative probabilities[7]. The relationship between data and theory may be viewed as a psychological exercise in which the scientist or expert is the subject. This technique has been employed in health risk assessment in the context of an evaluation of chloroform.[8] The major focus in this evaluation was to identify which data was important, which theories were plausible and the culmination in assignment of the probability that each theory was correct. Each of these tasks was accomplished subjectively, i.e. by consulting the opinion of a group of experts.

[5] Committee on Risk Assessment of Hazardous Air Pollutants, National Academy of Sciences, Science and Judgment in Risk Assessment (1995).

[6] Ian Hacking, The Emergence of Probability (1975). The philosophical term for model weighting is probability logic or logical probability. Hacking presents Pascal's wager on the existence of God as the first historical employment of probability logic.

[7] Granger M. Morgan & Max Henrion, *Performing Probability Assessment, in* Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis 141 (1991).

[8] John S. Evans et al, *A Distributional Approach to Characterizing Low-Dose Cancer Risk*, 14 Risk Anal. 25 (1994).

• *Weighting by algorithm.* Instead of selecting the best model, another possible use of explicit criteria for model preference is as a basis for assigning relative model probabilities.[9] The essential notion which defines this approach is that the weights assigned to the models for any given evidence is preordained by an explicit algorithm. Although models which are judged to be best are given more weight in generating a distribution of possible outcomes, any model judged *a priori* to have some likelihood of being correct will have an influence on the outcome of the assessment. As with weighting by expert, the generation of a weighting algorithm is a normative endeavor or process of explication. The most important difference is that an algorithm is expected to transcend particular issues.

To attain comparable results, it will also be necessary to preordain a list of plausible theories which will be used for a particular problem. Still, the constraint is loosened: Instead of one default model, the methodology is defined by a set of default models, the size of which is limited only by practical considerations. While it should be possible to revise the list of plausible theories, doing so constitutes a revision of the methodology, which may make the present evaluation incomparable to previous assessments. In such a circumstance, the value of improving the methodology by adding a new theory must be weighed against the cost of recalculating prior assessments to yield comparability.

A Functional Evaluation

Having grouped techniques for dealing with model uncertainty into six categories, let us examine them in terms of administrative criteria by which evaluation procedures are commonly judged. The ability to meet these criteria is in many cases interrelated, sometimes negatively so.

Separation of Risk Assessment and Risk Management

In public debates it is often desirable to separate the process of developing information (i.e. risk assessment) from acting upon that information (i.e. risk management). While virtually any process could be criticized for embedding some critical judgment within the technical evaluation, methodologies for dealing with model uncertainty may be

---

[9]   Clark D. Carrington, *Logical Probability and Risk Assessment,* 2 Hum. Ecol. Risk Assess. 62 (1996).

distinguished by the extent to which they endeavor to separate the discussion of what is done from what *ought* to be done. Attaining this goal is largely dependent on the recognition that risk assessment is a process which is directed by a practical question rather than facts, and in consequence, the relationship between risk assessment and risk management is a dialogue.

With regard to this criteria, technocratic decision making and the default factor approach may be quickly placed at one end of the spectrum. These methods do not attempt to separate risk assessment and risk management; at the end of the process the decision has been made.[10] The two single model approaches are better in this regard, but they both dictate to some degree the impact model uncertainty will have (or not have) on the outcome of the decision.

The application of "conservative" criteria in selecting a level or model also necessarily compromises the goal of delivering information because it is clearly intended to bias the result of the decision. There is also a potential problem with selection of best or central tendency criteria for selecting a single model — the possibility of other interpretations is not allowed to play a role in the outcome of the decision process and therefore, in some way biases the decision. If the assessment process is to be truly separate, then it is the decision-maker or policy-process which must determine the impact of the uncertainty, including model uncertainty, on the outcome of the decision, not the technical person or process. Model uncertainty cannot under any circumstances be portrayed with the results of a single model.

Transparency

Transparency is a matter of communication and justification. Transparency is to some extent a moving target and audience-dependent: Making the resolution of one issue transparent will often lead to another "subjective" issue. Simply stating a procedure does not guarantee that it will be accepted as reasonable. In fact, giving someone an opportunity to object may be the principle motivation for transparency. This discussion will focus on transparency only with respect to model uncertainty.

---

10 For instance, it may dictate some level of a chemical above which some action should be taken, or toss a compound into the jaws of the Delaney clause. Of course, this "decision" can be, and often is, second-guessed and preempted.

To make model uncertainty transparent, the development of a suitable vocabulary, whether it be words[11] or numbers, for describing model uncertainty is essential. Conversely, to the extent that any of the evaluation is subjective or requires direct participation of expert judgment, it is not transparent. Therefore, virtually the entire weight-of-the-evidence method and in part the methods employing defaults (assuming there is ever to be some deviation from the default) and decision analysis cannot be said to be transparent. Although the application of uncertainty or safety factors appears to be transparent as a policy, it is not transparent with regard to model uncertainty.

Transparency and the ability to separate risk assessment and risk management correlate strongly. Yet, as difficult as it may be to separate risk assessment from management, it may be even harder to combine them. In examining the rationale behind a putative action, discussion of intent or desire vs. belief or practicality with apparent necessity become separate issues. However, transparency may yield different decision-making paradigms. For instance, the default model approach, by seeking to resolve the issue as a matter of policy makes model uncertainty an issue of risk management, while the model weighting technique makes model uncertainty a matter of risk assessment.[12]

Cost of Execution

The cost of implementing a method will increase with the degree of difficulty in implementing the method, which in turn will limit the extent to which it can be used. The cost of making a health assessment methodology can be segregated into the cost of execution and the cost of administration. The first refers to the cost of actually carrying out the calculation. The second refers to the cost of defending the result. It may be necessary to weigh the difficulty in defending an included procedure against the difficulty in defending its omission. The selection of a particular approach may be thought of as a cost-minimization decision involving the sum of these two costs. Administrative cost is not

---

[11] Paul Krause et al., *An Argumentation-Based Approach toRisk Assessment,* 5 IMA J. Math Appl. Bus. Ind. 249 (1994). It is argued that it is inappropriate to use numbers to describe uncertainty because the reasoning is non-quantitative.

[12] For instance, in the recently proposed Environmental Protection Agency Cancer Risk Assessment Guidelines (16 Fed. Reg. 7998), the agency declined to represent uncertainty as a series of competing models because "This would obviate the function of the policy default."

considered in this discussion, as it will vary greatly with both the question and audience. With regard to cost of execution, the methodologies may be grouped into three categories:

• The cost of execution is kept to a bare minimum through the use of default factors.[13] This method can often be used without even the benefit of a calculator and generally bypasses any possible concern with relevant information, other risks, benefits or costs of avoiding exposure.

• A close second may be the single model approach using default methodology. Although the use of a predesignated model to predict an outcome will generally require a computer rather than a calculator, the calculation is simple enough to perform within a spreadsheet.

• Somewhat more costly are the best-model and weighting by algorithm approaches. Because the criteria are explicit, it is possible to use a computer to carry out the necessary evaluation of each and every model. Performing the task, however, is likely to require some programming which may be time-consuming. The amount of work required for evaluating which model is best as opposed to weighing the alternative models will generally not be very different. Employing the distribution of models resulting from a model-weighting scheme as a predictive tool will require an iterative (Monte-Carlo) simulation.

The evaluations conducted subjectively may be very simple or very difficult, depending on the level of effort expended by the individual(s) involved. A dictum by a single technocrat may be quite efficient. However, if the issue is to be decided by committee, the task may become very difficult. Because the need for expert opinion about model uncertainty arises from the absence of knowledge, the likelihood that an issue will be resolved may decrease as the expenditure for expert opinion increases — it may be very difficult to get experts to agree about what they don't know.

Correspondence to Scientific Opinion

If a risk assessment is expected to result in a statement of belief (as opposed to a regulatory standard or a decision point) then it must also embody some standard of truth. In science, the standard of truth is scientific opinion. For instance, the ultimate test for a statement in a

---

13 The use of experts to apply default factors (as opposed to someone else) may be thought of as an administrative cost, rather than a cost of execution.

scientific journal is peer review — a statement is considered to be true if it is accepted by disciplinary peers. Similarly, one of the criteria by which a risk assessment technique may be judged is by its ability to reflect scientific opinion. In this sense, a risk assessment may be seen as the normative element of science stripped of its empirical elements.

As long as there is no deception involved, methods which rely entirely on expert judgment will presumably reflect the opinion of the experts queried. Attempting to make an analysis transparent will tend to reduce correspondence to expert opinion, because it is fraught with the difficulty of communicating the state of knowledge or degree of uncertainty to a less knowledgeable (at least with respect to the issue) person. Also, because transparency requires general measures, there is always the danger that important considerations will be excluded from the analysis. Therefore, to the extent that the goal of the assessment is to meet a standard of scientific objectivity, expert opinion is the gold standard by which other methods must be judged.

However, if the expert is entrusted with the entire decision, then there is more than belief involved, the expert also charts an action. If there are multiple experts involved (i.e. a committee), there will be no way of knowing whether any disagreement that the experts may have over an action involves belief or valuations of the beliefs, unless they are able to convey (at least among themselves) their state of belief. Therefore, one cannot determine whether or not correspondence to scientific opinion has been attained without some transparent segregation of risk assessment and risk management.

At the other end of the spectrum, employment of a default factor or model virtually guarantees that the result will not accurately reflect current scientific opinion. First, default methodologies may not meet the decision-maker's criteria. The same model or factor cannot always be the best model for all sets of evidence. Similarly, if the decision-maker wants to be "conservative", the same model cannot be equally conservative all the time. The motivations of reproducibility and simplicity which favor default-based methodology necessarily make the evaluations relatively data-insensitive. Even if departures from the defaults are permitted, rather substantial evidence must be assimilated before this result is obtained and there will inevitably be a bright line which in some circumstances can greatly accentuate small differences in

the bodies of evidence for two different chemicals. Second, default methodologies do not, by design, give the decision-maker any representation of the extent of the uncertainty arising from interpretation. Therefore, the decision-maker is given an appearance of certainty which matches no one else's opinion.

Between judgment given free reign and thoughtless application of rules lie those methods which employ criteria for evaluating models. In the final analysis, objective model preference criteria must also be thoughtless rules, but because they are applied more abstractly, good rules may be expected to generate a result which is better adapted to the evidence under scrutiny, and will (by design) more closely match the judgment of an expert. Consistent variance of rules with scientific opinion may be taken as evidence that the rules are in need of revision.

The desire for an accurate portrayal of the state of current knowledge must often be balanced against the need for a method to be simple and/or transparent. Meeting the demand for objectivity tends to introduce superficiality.[14] This can be overcome by making the process more elaborate, but the cost of executing the process is then sacrificed. If the issues are at all complex, then responding to those issues in a reasonable manner will necessarily become more difficult, and communicating or explaining the results of an analysis will require more effort.

Impersonality

The pursuit of truth can be endless. In contentious circumstances where decisions must be made, the truth may not be worth having and there may be no alternative to preempting further discussion with a dictatorial statement. In a democracy, the responsibility for defending such a statement is unlikely to stand for long on the opinion of an individual or even a small group of experts. It is thus often desirable for a technique to take on a life of its own by becoming mechanical and impersonal, a quality of an institution rather than a person.

To put this issue in another light, it may take on the heading of risk comparison. Even when there is only one interest at stake, there is often

---

[14]   Theodore M. Porter, *Trust in Numbers, The Pursuit of Objectivity in Science and Public Life* (1995). Exploring the tradeoff between relying on rules and relying on expert judgment.

a need to compare risks. If uncertainty is an issue to be decided by each expert, one has no common standard unless the same experts perform all the evaluations. There is also the matter of justice. Regardless of whether or not the method is intended to have separate steps for information development and decision-making, there is often a need to be able to demonstrate that an even-handed process is applied.

The clear example of such law-like preemption of discussions of model uncertainty is use of the default model. Default factors, best model and weighting by algorithm also have potential for the attainment of impersonality — so long as the involvement of experts is a thing of the past. Since expert judgment and weighting involve the direct expert participation in the decision or information development, impersonality is not achieved. However, even if it doesn't explicitly say where weights come from, the weighting by expert method may be given some credit for stating which theories are employed.

Transparency is prerequisite for impersonality; thus, impersonal methods will also be transparent. This correlation is not entirely reflected in summary Figure 2.

Figure 2
Summary Evaluation of Six Methods for Dealing with Model Uncertainty

	Transparency	RA vs. RM	Cost of Execution	Scientific Opinion	Impersonality
Technocracy	P	P	I/P	I	P
Default Factor	P	P	I	P	G
Default Model	F	P	G	P	I
Best Model	F	F	G	F	I
Weighting by Expert	G	G	P	I	F
Weighting by Algorithm	I	I	F	G	I

Fair, Good, Ideal or Poor?

The correlation is not reflected because the methods have been rated for impersonality based on the entire process, whereas the transparency rating refers to model uncertainty only. Yet, transparency might refer to expert judgment, clearly not impersonal. Impersonality may also affect execution cost if a process can be performed by computer.

Beyond that, Figure 2 may be used in two ways: Given motivation or criteria, a method may be selected. Or, given a method, a motivation may be gleaned. The split rating for Technocracy in terms of Cost of Execution reflects the efficiency of the method when a single expert is used versus its inefficiency when a a committee is employed.

Conclusions

All of the techniques for dealing with model uncertainty involve trade-offs with regard to the criteria that have been offered for judging them. This is in no small way attributable to the fact that some of the criteria are mutually exclusive. Which technique is selected for use may therefore be expected to vary with the needs of the administrator:

• Although a technocratic dictum is not at all transparent, this is not a limitation as long as no one questions the decision. However, as soon as an argument arises, the motivation and/or the expertise underlying the judgment may both be questioned and transparency becomes a critical issue.

• Although methods employing defaults are undesirable for other reasons, their administrative and quantitative simplicity may still make them attractive in some circumstances. In particular, they may continue to be useful for screening out non-problems, which are issues that do not merit the attention and effort another method would require. A technique employing a best model concept might provide a better match with scientific judgment without greatly encumbering the analysis with an account of alternative models. However, if for some reason the default fails to triumph over controversy, it is unlikely that the argument can be resolved without resorting to a more transparent analysis.

• Although weighting by experts is difficult to implement and generates results which are difficult to compare, the cost may be

justified for particularly important and novel issues. If the goal is to obtain the best available statement about a single issue, then having a result which is comparable to others may not be very important. In addition, an exercise in eliciting expert judgment may present an occasion for initially formulating or testing model preference criteria.

To the extent that comparability among risks is desired, weighting by algorithm would appear to have the greatest merit. It is transparent, reproducible and achieves impersonality while maximizing, by design, concurrence with opinion (scientific or otherwise). The principal negative aspect of this approach compared to some of the others is the additional effort required to complete the necessary calculations. In fact, a computer is necessary to make the project conceivable. However, because much of the effort required would necessarily (to meet a criterion of reproducibility) be transferable across issues, the expenditure required might be expected to diminish with use.