Environment International 105 (2017) 20–33

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Characterising bias in regulatory risk and decision analysis: An analysis of heuristics applied in health technology appraisal, chemicals regulation, and climate change governance

Brian H. MacGillivray

Sustainable Places Research Institute, Cardiff University, United Kingdom

ARTICLE INFO

Keywords:
Risk regulation
Risk analysis
Clinical trials
Statistical inference
Model uncertainty
Methodology

ABSTRACT

In many environmental and public health domains, *heuristic* methods of risk and decision analysis must be relied upon, either because problem structures are ambiguous, reliable data is lacking, or decisions are urgent. This introduces an additional source of uncertainty beyond model and measurement error – uncertainty stemming from relying on inexact inference rules. Here we identify and analyse heuristics used to prioritise risk objects, to discriminate between signal and noise, to weight evidence, to construct models, to extrapolate beyond datasets, and to make policy. Some of these heuristics are based on causal generalisations, yet can misfire when these relationships are presumed rather than tested (*e.g.* surrogates in clinical trials). Others are conventions designed to confer stability to decision analysis, yet which may introduce serious error when applied ritualistically (*e.g.* significance testing). Some heuristics can be traced back to formal justifications, but only subject to strong assumptions that are often violated in practical applications. Heuristic decision rules (*e.g.* feasibility rules) in principle act as surrogates for utility maximisation or distributional concerns, yet in practice may neglect costs and benefits, be based on arbitrary thresholds, and be prone to gaming. We highlight the problem of rule-entrenchment, where analytical choices that are in principle contestable are arbitrarily fixed in practice, masking uncertainty and potentially introducing bias. Strategies for making risk and decision analysis more rigorous include: formalising the assumptions and scope conditions under which heuristics should be applied; testing rather than presuming their underlying empirical or theoretical justifications; using sensitivity analysis, simulations, multiple bias analysis, and deductive systems of inference (*e.g.* directed acyclic graphs) to characterise rule uncertainty and refine heuristics; adopting "recovery schemes" to correct for known biases; and basing decision rules on clearly articulated values and evidence, rather than convention.

1. Risk, decision, induction and uncertainty

Risk and decision analysis are central tools of contemporary environmental and public health governance, in contexts ranging from the appraisal of novel pharmaceuticals, to nuclear waste disposal, to climate change adaptation and mitigation planning. They explore how the future might unfold if a policy maker was to undertake a particular course of action, often using utility functions to determine which outcome is "best" (Kaplan and Garrick, 1981). These tools have their roots in theories of probability and utility maximisation, and so are often implicitly seen as deductive systems. In deductive systems, when the underlying assumptions (premises) are valid, then the conclusion is logically entailed (holds true). The validity of the inference rules themselves (*e.g.* Bayes Theorem) is secure given that they are derived from basic axioms (*e.g.* the product and sum rules of probability theory). In this view, probability theory is an extension of formal logic (Jaynes, 2003), and probability and utility theory are the intellectual core of risk and decision analysis (Kaplan and Garrick, 1981; Savage, 1972). It is well known that probability and decision theory can never solve problems of actual practice, but rather idealisations of them, and are valuable to the extent that those idealisations are good ones (Jaynes, 2003; Savage, 1972). And so unsurprisingly, methodologists, practitioners and critics of risk and decision analysis have tended to focus on whether those simplifying assumptions are reasonable (*e.g.* is model structural error well characterised, are the parameter estimates subject to large measurement errors; Smith and Stern, 2011; Spiegelhalter and Riesch, 2011), rather than on the procedures of inference themselves or their combination. This is entirely fine for problems whose structure is sufficiently developed such that the full decision-theoretic apparatus can be applied (Jaynes, 2003). However, this is often not the case in practice in environmental and public health applications. Consider the following scenarios:

E-mail address: macgillivraybh@cardiff.ac.uk.

http://dx.doi.org/10.1016/j.envint.2017.05.002
Received 10 September 2016; Received in revised form 2 May 2017; Accepted 3 May 2017
Available online 09 May 2017

Scenario 1: An expert committee is tasked with estimating the likely rise in global mean temperatures under various emission scenarios. There are multiple models available, differing across various dimensions (*e.g.* representation of physical processes, dataset calibrated on, *etc.*), and producing variable estimates of the parameter of interest. How are these differing estimates to be reconciled or combined, given that there is no clear measure of model quality?

Scenario 2: A team of analysts is tasked with evaluating the safety case for a new-build nuclear plant. Whether the plant is sufficiently robust to extreme events depends on the climate model downscaling technique adopted, yet there is no clear basis for discriminating between these. Which method should the analysts adopt, given that neither data nor theory is determinative?

Scenario 3: Observational data shows an association between a pharmaceutical and adverse outcomes in a population subgroup. Pre-marketing clinical trials found no statistically significant adverse effects. The observational study and the trial differ amongst various dimensions, with neither clearly superior. A plausible biological theory links the drug to the adverse effects, yet it is unclear why harm should be restricted to the subgroup. What causal claims can be made?

Scenario 4: An ecosystem is threatened by climate change. The mechanisms that govern its functioning are poorly understood, but analysts have identified warning signals that often presage tipping points in similar systems. One of these warning signals (a reduced rate of recovery from perturbations) has been reached for the system of interest. How should the policy maker act in the face of this surrogate data?

These diverse scenarios reflect fundamental tasks of inference and choice: hypothesis testing; weighting evidence; model selection; extrapolation; and selecting policies in the face of uncertainty. Moreover, they relate to problems where a fully Bayesian or decision-theoretic analysis is often implausible; either because the problem structure is ambiguous, there is a lack of reliable data, or because decisions are urgent. In such situations, inexact, heuristic methods of problem-solving must be relied upon (Jaynes, 2003). That is, analysts and decision makers rely on heuristics to prioritise potential threats, to define what constitutes valid data, to discriminate between signal and noise, to weight lines of evidence, to select and apply mathematical models, and to make policy recommendations. Heuristics are not problematic *per se*, but they can become so when treated as laws rather than as contingent and provisional rules (Polya, 2004), and are inescapably connected with systematic error. Whilst not drawing on the notion of heuristics *per se*, recent years have seen a growing interest in the problem of bias within regulatory science at the level of both individual studies and evidence synthesis. In the former category, concerns have been raised about data-dredging, selective reporting of results, and unacknowledged researcher degrees of freedom (Ioannidis et al., 2014; Gelman and Loken, 2014); the role of informal epidemiological conventions in shaping research designs (Greenland, 2012a); the "absurd precision" afforded to random error whilst systematic error is treated in an *ad hoc* fashion (Greenland, 2005); the role of industry funding in skewing study outcomes (Suter and Cormier, 2016); error introduced by routine misinterpretations of significance testing; a lack of attention to the compound effects of multiple uncertain inferences in epidemiological analysis (Lash et al., 2016); and bias stemming from the incomplete representation of physical processes in climate impact modelling (Brysse et al., 2013). At the level of evidence synthesis, common concerns include publication bias (Dwan et al., 2008); errors stemming from the inflexible use of hierarchies of evidence and questionable quality-scoring techniques in meta-analysis (Greenland and O'Rourke, 2001); the exclusion of novel experimental protocols from consideration in regulatory standard setting (Myers et al., 2009); and potential bias stemming from the compound effect of "conservative" assumptions (Nichols and Zeckhauser, 1986; c.f. Finkel, 1997). We draw upon, synthesize, and extend much of this work within the rubric of heuristic inference.

We define heuristics as rules of thumb for inference and choices. There are three elements to this definition: heuristics are rule-like, in the sense that there is a presumption in favour of following them (*i.e.* they structure or constrain judgment); they are rules of inference or choice (*e.g.* if-then), rather than simply assumptions; and they are simple, frugal approaches to problem solving that ignore relevant information rather than seeking optimal solutions (Gigerenzer and Gaissmaier, 2011; MacGillivray, 2014). We show that heuristics play central roles in risk and decision analysis across an array of policy domains; have sometimes been applied in ways that bias analysis outcomes and lead to sub-optimal decisions; and suggest how these heuristics might be evaluated and refined so that regulatory science becomes rigorous.

2. Conceptual framework, aims and research approach

This paper builds on the idea that many aspects of regulatory risk and decision analysis are in practice inductive. It follows that if we are concerned with robust, evidence-based public policy, not only do we need to focus on the *assumptions* behind decision analysis, but also on the chains of reasoning that transform those assumptions into conclusions. Inductive reasoning is distinct in form to deductive reasoning, and requires different approaches to evaluate its form, quality, and implications (Pearl, 2014). One key difference is that heuristic reasoning offers support for particular conclusions, but does not guarantee them in the way that deductive systems do. Valid logical arguments are "truth preserving," and in a similar vein valid probabilistic arguments are "probability preserving," whilst inductive arguments may be at best only plausible. Moreover, in inductive reasoning, it is not just the structure of the argument that shapes validity (as in, *e.g.* classical urn and ball representations), but also background knowledge and the particulars of the case being examined (Pólya, 1990). In inductive systems, model "debugging" is a different and more complex task than in deductive systems, as when outputs do not match reality, we cannot simply focus on (revising) model assumptions, but must also consider rules of inference or their combination as potential sources of error or uncertainty (Pearl, 2014).

The heuristics that we are concerned with are different from the lay rules of thumb of the heuristics and biases tradition, which are domain-general, fuzzy, and subconscious (Kahneman et al., 1982). We follow Polya (2004) in seeing heuristics as provisional methods and principles of discovery that fall short of demonstration, yet that are indispensable for domains that do not allow for formal logic or proofs. This interpretation heavily influenced research in artificial intelligence and expert systems (Simon and Newell, 1958) where heuristics are conceived and modelled as one of the foundational elements of expert knowledge. A key insight of this approach is that expertise is less about general purpose inference, and more about acquiring and representing domain-specific knowledge (Feigenbaum, 1977). This expert knowledge has widely been represented as sets of if-then rules that guide and structure inferences, decisions, and problem solving in systematic ways. These rules are often expressed in formal language rather than verbally, and work by exploiting empirical facts or causal relations that are more or less accepted or understood. Although they do not guarantee correct solutions or optimal outcomes, they may offer useful guidance through drawing on (imperfect) knowledge and making problems tractable (*e.g.* by ignoring information or options; Gigerenzer and Todd, 1999). This interpretation of heuristics—with its focus on formalism, domain specificity, and on the interaction of rules with environmental structures—is particularly relevant for characterising expert (rather than lay) risk and decision analysis.

This paper's analysis covers chemicals regulation, health technology appraisal, pharmaceutical regulation and climate change governance, necessarily in a somewhat schematic fashion. We focus on the following

regulatory agencies: the US Environmental Protection Agency (EPA) (chemicals regulation), the US Food and Drug Administration (FDA) (pharmaceuticals regulation), the UK's National Institute for Clinical Excellence (NICE) (health technology appraisal), and the Intergovernmental Panel on Climate Change (IPCC) (climate change evidence synthesis). The logic of this focus is that these organisations have pioneered the application of risk and decision analysis methods within their domains (meaning that many regulatory bodies worldwide adopt similar practices, and so the findings should have reasonable generality), and moreover their practices are relatively transparent and well documented.

The research aims are to:

1) Identify and taxonomise heuristics used within these domains of regulatory science;
2) Characterise the biases that may stem from applying these heuristics, drawing on both formal arguments (*e.g.* results of simulations or sensitivity analyses) and real-world empirical examples; and
3) Identify strategies for rigorizing heuristic approaches to risk and decision analysis, with a view to making regulatory science more robust and evidence-based.

The paper focuses on the actual *practices* of regulatory risk and decision analysis, rather than on the mathematical frameworks of probability theory and utility maximisation. Data was collected from various sources describing, evaluating, and providing context or background on the conduct and interpretation of regulatory science within these areas. These include: a) Statutes, guidelines, procedures and outputs relating to policy analysis and policy-making; b) critiques and evaluations of those practices from within the scientific and policy communities; and c) primary research papers and reviews providing detail and background on the relevant theories, methods, and assumptions adopted. Given the scope of the paper, the data collection process was necessarily schematic rather than systematic or comprehensive in nature. The analysis began by inspecting data sources that represented official or quasi-official characterisations of state of the art methodological practices (*e.g.* US EPA and FDA methodological guidelines; IPCC reports) to identify the heuristics applied within each domain. A subset of those heuristics were selected for more in depth analysis (data sources b) and c) above). This subset was selected based on the following criteria: scope or extent of application (*i.e.* significance); and the existence of data relating to the heuristic's underlying assumptions, potential biases, and alternative methods of inference. The subset of heuristics were then grouped into categories according to the functions they performed (*e.g.* weighting lines of evidence, decision rules, *etc.*), which was used to structure the analysis and discussion sections. A typology was then developed that classifies the heuristics according to their structural features (Table 1). Table 2 syntheses the overall results, cataloguing problems and biases associated with different kinds of heuristics, and prospects for rigorization.

3. Rules of thumb in regulatory risk and decision analysis: results and discussion

3.1. Screening heuristics

A basic problem in regulatory science is that not all objects of potential interest can be evaluated intensively, and so some approach has to be devised to search and prioritise the problem space. Screening heuristics are widely used for this purpose in chemicals regulation (MacGillivray, 2014), ranging from simple if-then rules to more complex categorisation trees. In some cases they draw on proxies for exposure (*e.g.* threshold values of production volume, persistence and bioaccumulation) as triggers for the degree of scrutiny for a particular chemical. In others they draw on mechanistic knowledge to guide the particular form that scrutiny should take (*e.g.* decision trees that categorise chemicals by structural properties, which are mapped to specific test batteries). Screening rules depend on the idea of *surrogates*, *e.g.* the idea that persistence and bioaccumulation (the surrogates) are indicative of expected exposure levels (the target variable), or that particular structural properties of compounds are predictive of metabolic and toxicological behaviour. They are particularly useful when collecting data on the target attribute may be impractical or costly. Screening rules are analogous to the biomarker-based approaches used in medicine to identify individuals at elevated risk of disease (*e.g.* occult faecal blood as a surrogate for identifying patients potentially having colon cancer), and to the use of surrogate endpoints in clinical trials. They can be classed as heuristic syllogisms (Polya, 2004), with the skeleton form:

Premise: If A then B
Premise: B true
Conclusion: A more credible
(Example: A: patient has colon cancer; B: patient has elevated levels of occult faecal blood.)

Generally, the conclusion will be tied to an action, which may be to better characterise the likelihood that A is true (*e.g.* performing more intensive diagnostic tests), or which may be based on the presumption that A is true (*e.g.* provide medication). There are two key issues in evaluating this class of heuristic: is the link between the surrogate and the object of interest well theorised; and how strong is the empirical link (*e.g.* in terms of sensitivity and specificity)? For example, production volume is one surrogate for exposure widely used to guide chemical test requirements. However, the empirical relationship between production volume and exposure (the IPR) has recently been shown to differ by up to five orders of magnitude (Nazaroff et al., 2012), with structural class and intended chemical uses key moderators of this relationship. This suggests that generic estimates of IPR for broad classes of chemicals could be derived from sampling, perhaps allowing it to replace production volume as a more robust screening heuristic (*ibid*).

These concerns are magnified when surrogates migrate from being the basis of screening rules to become the *objects of regulation*, as in the growing reliance on surrogate endpoints as measures of effectiveness in clinical trials (Atkinson et al., 2001; Fleming and DeMets, 1996; Ioannidis et al., 2014; Kazi and Hlatky, 2012). Here, surrogate endpoints, such as tumour shrinkage or changes in cholesterol level, substitute for clinically important endpoints such as morbidity or mortality, with the aims of maximising statistical power and reducing trial costs and duration. This approach is problematic when the surrogate lacks a well-established link to the clinically important outcomes (Ioannidis et al., 2014; Kazi and Hlatky, 2012). An example is Flecainide, a drug for reducing the risk of cardiac death from abnormal heart rhythms, which was brought to market on the basis of its performance against a surrogate endpoint (the suppression of arrhythmias). However, post-marketing trials found that it actually *increased* mortality from heart attacks in certain patient populations (Echt et al., 2001). The general point is that surrogates are not the same thing as the objects that they purport to represent. A failure to act in a way that recognises this – such as by neglecting to collect systematic evidence evaluating the presumed underlying relationship – can lead to inefficient and even harmful regulatory outcomes (Atkinson et al., 2001).

3.2. Causal inference

Approaches to representing and analysing cause-effect relations range from formal deductive frameworks (*e.g.* directed acyclic graphs (DAGs); Pearl, 2000), to purely statistical methods (*e.g.* hypothesis testing), to criteria-based approaches. Hill's (1965) criteria are a set of inductive factors used to separate causal from non-causal explanations in toxicology: strength; consistency; specificity; temporality; biological gradient (monotonicity); plausibility; coherence; experimental evi-

Table 1
A typology of heuristics applied in regulatory risk and decision analysis. Existing classifications of heuristics are typically subject-specific so we have not built on them *per se*. However, we drew on Wimsatt's (2006) class of model building heuristics, and Clancey's (1983) notion of "identification rules" has some parallels with categorisation rules. The idea of attribute substitution is central to Kahneman and Frederick's (2002) understanding of representativeness, and has parallels with our notion of surrogates within screening rules. MacGillivray (2014) classed a series of heuristics used in chemical risk assessment according to their functions, but without offering a taxonomy as such. We build on his classes of screening rules, gatekeeping rules, evidence hierarchies, and interpolation and extrapolation rules.

Heuristic class	Heuristic sub-class		Description
Categorisation rules	Screening risk objects		Decision trees used to categorise risk objects to inform both priority setting and testing requirements, often based on surrogates.
	Causal inference	Signal *vs*. noise	Variants of significance testing used to discriminate between noise from a signal, and interference from a distractor.
		Domain-specific	Domain-specific criteria used to discriminate between causal and non-causal associations.
	Data exclusion		Gatekeeping rules that exclude data generated from certain non-standard research designs from quantitative risk or benefit assessments.
	Evidence hierarchies		Hierarchies of evidence used to rank (potentially conflicting) sources of evidence according to the quality or rigor of their underlying research designs.
Model construction rules	Methodological prescriptions		Conventions prescribing *default* choices of model structure, functional form, and parameter values.
	Methodological principles		Principle-based inference rules that guide choices of model structure, functional form, and parameter values.
Adjustment rules	Debiasing		Rules that stipulate how model outputs should be adjusted to correct for perceived biases (*e.g.* safety factors).
	Extrapolation/scaling		Rules or principles that dictate how model outputs should be adjusted to extrapolate across categories (*e.g.* from human to animal), scales (*e.g.* in climate impact modelling), places or contexts.
Combination rules	Combining estimates		Rules that assign weights to different estimates of the same phenomena to allow for their combination, *e.g.* within meta analyses or multi-model ensembles.
	Combining separate outcomes		Rules that govern how different risks, impacts, or outcome types should be combined to generate aggregate measures.
Decision rules	Absolutes		Rules which deem the presence of a property (*e.g.* carcinogenic) as sufficient grounds for regulating an object.
	De minimis and de manifestis		Quantitative thresholds, based on theory, empirical data or arbitrary conventions, used to distinguish between negligible risks and those that require mitigation.
	Feasibility rules		Rules that mandate the particular technology to be used for environmental regulation, constrained in terms of the best that is available or feasible.
	Cost-effectiveness		Thresholds used to determine whether a regulatory intervention is rational in the face of budget constraints.

dence; and analogy. Save for temporality, each of these factors may be absent in genuinely causal relationships (Rothman and Greenland, 2005). Analogy and plausibility are particularly subjective, being dependent on the state of knowledge or imagination of a given investigator in a given timeframe. A precursor to these criteria are Koch's postulates, a set of principles that must be satisfied to establish a causal relationship between a microbe and a disease (including that the agent should be found in all cases of the disease). Doll (2002) recounts that these postulates were initially mapped across to chemical toxicology, and used to argue that smoking could not be a cause of lung cancer (as there were examples of cases with the disease that did not smoke). This illustrates the dangers of using rules outside of their domain of justification. Koch's postulates rest on a notion of causation – which saw it as deterministic and singular – that was useful for microbiology but quite wrong for chemical toxicology, where causes of disease are multiple and stochastic (*ibid*).

A form of causal inference central to many regulatory domains is the discrimination between signal from a target and noise from a distracter. For instance, does a spike in hospital mortality rates indicate an underperforming institution (*e.g.* sub-standard surgical practices or conditions), or is it random variation? Can a change in weather patterns (*e.g.* altered frequency or strength of North Atlantic storms) be attributed to anthropogenic causes, or is it within the bounds of natural system behaviour? Simple rules of thumb, generally variants of significance testing, are widely used to structure this class of inferences, particularly where data is generated by randomised experiments (in *theory* ruling out systematic error) and there is limited prior information. They are particularly prominent in public health:

- The FDA requires two well-designed studies where treatment effects pass the conventional threshold of significance before a pharmaceutical can be placed on the market (NRC, 2012);
- Clinical trials may be stopped early if there is substantial evidence of a clearly superior treatment, with substantial evidence widely defined in terms of p-values (*e.g.* the Haybittle-Peto boundary requires $p < 0.001$ to stop a trial early for benefit). However, these are typically interpreted as presumptive guidelines rather than strict rules (FDA, 2006); and
- Simple decision trees are used in pharmacovigilance by the European Medicines Agency (2016) to classify whether a drug-event pair is reported disproportionately (relative to an independence model). One example is: a) if the lower bound of the 95% confidence interval > 1; and b) the number of individual cases ≥ 3 for active substances contained in medicinal products included in the additional monitoring list (or ≥ 5 for the other active substances); and c) the event is classed as an "Important Medical Event", then investigate further for potential causal relationships.

Variants of the significance testing heuristic are applied in chemicals regulation, *e.g.* to determine whether chance, rather than a treatment-related effect, is a *plausible* explanation for an apparent increase in tumour formation, and to differentiate between "true" detections of a substance and those that cannot be reliably distinguished from instrument error or noise (MacGillivray, 2014). These heuristics share the basic structure of using a statistical criterion to identify a result or phenomenon that is sufficiently extreme that it qualifies as signal rather than noise (*i.e.*, sufficiently extreme to reject the null hypothesis, or more sensibly, treat it as suspect pending further investigation). A long established but sometimes overlooked principle is that a failure to reject the null hypothesis does not logically entail accepting it. A safeguard to this problem is found in chemical exposure assessment. Here, where observed values (concentration levels) are deemed to be too low to be reliably distinguished from instrument error or noise, they are not treated as zero (which would bias risk assessments downwards). Instead, proxy values are used. This parallels Pearl's (1984) notion of "recovery schemes," which are designed to guard against biases that may stem from rule-based reasoning.

A variant of the above is outlier screening, where extreme results are considered as noise rather than signal. The distinction is that outlier screening involves treating the model or theory as (provisionally) true and mistrusting the data, rather than assuming the data to be correct and mistrusting the theory (Gigerenzer and Sturm, 2007). An example is the use of significance tests to detect low outliers in flood frequency analysis in the United States (GCER, 1999). A crucial point is that

Table 2
Examples of regulatory heuristics, their associated problems, and prospects for rigorization.

Rule class	Rule description	Applications and problems	Towards rigorization
Screening risk objects	Decision trees used to categorise risk objects to inform both priority setting and testing requirements, often based on surrogates.	Screening rules based on surrogates are widely used to search the problem space in chemicals regulation, *e.g.* production volume is used to guide chemical test requirements, however its relationship to exposure has been shown to differ by up to five orders of magnitude (Nazaroff et al., 2012). Surrogates are increasingly used in the analysis of clinical trials to increase statistical power and reduce costs, but their clinical relevance is not always clear. For example, suppression of arrhythmias was used as a surrogate for mortality from heart attacks in trials of Flecainide. The drug was brought to market on the basis of those trials, yet post-marketing trials found that the drug *increased* mortality from heart attacks in certain subgroups (Echt et al., 2001).	Relationships between surrogates and endpoints should be empirically validated, rather than presumed on basis of plausibility. The identification of variables that significantly moderate the relationship between surrogate and endpoint (*e.g.* Nazaroff et al., 2012) can help to improve screening rules.
Causal inference: signal *vs.* noise	Variants of significance testing used to discriminate between signal and noise.	Significance testing is used for causal inference in numerous domains of environmental and public health regulation. Well-rehearsed problems include the arbitrary choice of thresholds, the neglect of utilities, and the neglect of domain-specific knowledge. Its application to observational data for causal inference is particularly suspect, as the no bias assumption is rarely justifiable. For example, the FDA recommended the use of oestrogen to reduce cardiovascular risk in post-menopausal women on the basis of statistically significant findings of protective effects in observational studies. Subsequent RCTs found no benefit and possible harm (Greenland, 2005).	Ritualistic use of significance testing is a problem not restricted to regulatory science. In general terms, significance tests should be used to characterise random error, not for causal inference. They should be supplemented with bias analysis – formal or informal – where applied to observational data. Standard inferential statistics is not the problem – their misinterpretation and misapplication is.
Causal inference: domain-specific	Domain-specific criteria used to evaluate whether a causal relationship has been demonstrated to hold.	Koch's postulates set out principles that must be satisfied to establish a causal relationship between a microbe and a disease. They were used to argue that smoking was not a cause of lung cancer. The error stems from the fact that Koch's postulates rest on a notion of causation that was inapplicable to chemical toxicology, where causes of disease are multiple and stochastic (Doll, 2002). Hill's (1965) criteria are inductive *factors* used to separate causal from non-causal explanations in toxicology: strength; consistency; specificity; temporality; biological gradient (monotonicity); plausibility; coherence; experimental evidence; and analogy. Save for temporality, each of these factors may be absent in genuinely causal relationships (Rothman and Greenland, 2005). Analogy and plausibility are particularly subjective, being dependent on the state of knowledge or imagination of a given investigator in a given timeframe. The system as a whole does not explicitly deal with confounding.	There are temptations and dangers in exporting heuristics across domains – generalisation is in many senses synonymous with scientific advance, but *successful* generalisation requires a coherent justification. Weed (1986) and Maclure (1985) sought to transform Hill's criteria into deductive tests of hypotheses (Rothman and Greenland, 2005). DAGs provide a deductive basis for estimating causal relationships. However, full applications to public and environmental health contexts have been limited by the strong assumptions required (Greenland, 2012b).
Evidence hierarchies	Hierarchies that categorise quality of evidence according to study design features (*e.g.* RCTs *vs.* cohort studies).	Evidence hierarchies are widely used in public health to resolve inconsistent findings. Empirical evidence shows the dangers of applying them inflexibly, *e.g.* Concato (2004) used meta-analyses to explore their underlying assumptions, finding that average results from well-designed observational studies did not overestimate exposure-outcome associations compared to those reported in RCTs. More anecdotally, two small RCTs found evidence of a protective effect of β-blockers on cardiac postoperative events (Neuman et al., 2014). Their perioperative use in patients with coronary artery disease was promoted as “best practice” on the basis of these trials. This was later overturned, largely on the basis of a large scale cohort study which associated the use of β-blockers with substantially greater mortality in patient subgroups.	Evidence hierarchies may be useful heuristic tools, but they do not absolve researchers of the obligation to evaluate the merits of the design, conduct, and analysis of *individual* studies. Risk of bias tools can structure this process (*e.g.* Higgins et al., 2011). Recovery schemes setting out justifications for down-grading or up-grading the quality ranking of specific study categories are useful, *e.g.* those found in GRADE (Guyatt et al., 2008).
Model construction rules	Conventions constraining choices of model structures, functional forms, *etc.*	Such conventions are widespread within integrated assessments in climate science, health technology appraisal, and chemical risk assessment. There is wide variation in the extent to which these defaults	Rule-bound approaches to modelling are typically difficult to justify in the social and environmental sciences, where empirical regularities are few, measurement difficult, and there is rarely a strong

(continued on next page)

Table 2 *(continued)*

Rule class	Rule description	Applications and problems	Towards rigorization
		are well-founded, *vs*. the extent to which they seek to confer standardisation for its own sake. The latter kinds of rules may introduce bias and artificially close down uncertainty, particularly when *collections* of modelling rules are standardised. Morgan (2014) describes a climate impact modelling framework within which analysts were allowed to alter discount rates, damage functions, and objective functions (typically fixed by convention). He found that the identification of the optimal climate mitigation policy was highly sensitive to alternative plausible model specifications. If stability in outcomes is contingent on standardising methodological choices that are in principle contestable, then it is unclear whether the outputs have a clear physical interpretation.	theoretical basis for constraining choices *e.g.* of distributional and functional form. Sensitivity analysis can explore the robustness of results to alternative model construction rules. Focussing on collections of rules is critical (*e.g.* Morgan, 2014). Deductive frameworks serve as a rigorous basis for constructing causal models (subject to strong assumptions) or as normative standards for evaluating specific modelling conventions. They also inform methods of multiple bias analysis applied in public health (Greenland, 2005; Lash et al., 2016).
Extrapolation/scaling rules	Adjustment rules used to extrapolate model results across scale, place, species, *etc*.	The extrapolation of model results to other locations, scales, or circumstances is often a judgment-laden, *ad hoc* process. For example, global climate impact studies tend to be based on data from wealthy nations, which is then extrapolated across to LMICs with informal adjustments to account for differences in geography, adaptive capacities, *etc*. (Smith et al., 2001). Elsewhere, explicit extrapolation heuristics are used, *e.g.* uncertainty factors to transport toxicology estimates across species, and the use of elevation as a surrogate for temperature gradients in down-scaling climate models. The latter heuristic may be biasing species distribution forecasts, as its relative coarseness leads to the neglect of refugia which play important roles in species' adaptation (Ashcroft et al., 2009).	There is lack of general theory on the assumptions required to extrapolate causal results across scale, place and context. Consequently, there is no domain-general normative standard for evaluating heuristic extrapolations. Theory-driven empirical studies (*e.g.* replications of studies across locations or species, meta-regression, and simulations) aimed at validating or refining extrapolation heuristics can be useful on a case-by-case basis.
Combining estimates	Rules (*e.g.* equal weights) for combining estimates of the same phenomena	The IPCC combines – without weighting – the outputs of multiple models into simple averages, standard deviations or ranges (Knutti, 2010). A formal justification for equal-weighting is Laplace's Principle of Insufficient Reason, which relies on the strong assumption of pure uncertainty. Even where the assumption holds, equal-weighting leads to inconsistent results in cases where the object being combined (*e.g.* parameter value) may be represented in more than one mathematical form, and those forms are non-linearly related (Frigg et al., 2015b). An example is the representation of uncertainty in the speed at which ice falls from clouds in climate modelling (*ibid.*).	Simulation studies can characterise the sensitivity of model results to alternative weighting schemes where the pure uncertainty assumption does not hold (*e.g.* Clemen and Winkler, 1999). Bayesian approaches can be used to weight climate models according to skill, although are constrained by the fact that model performance is quite context-specific.
Combining separate outcomes	Rules (*e.g.* vector addition) for estimating the joint effects of separate risks, impacts, or outcome types.	Vector addition is standard in integrated impact assessments of climate change, where impacts are typically characterised on a sector-by-sector basis, then combined to derive the aggregate economic impact (Stern, 2007). Results (Harrison et al., 2016) suggest that this presumption of additivity is biasing impact evaluations (and by extension adaptation decisions), *e.g.* through neglecting the way that changes in water availability influence the balance of irrigated and non-irrigated crops in a given area, which in turn influence food production. In clinical trials multiple endpoints are often aggregated into composite measures to reduce sample size requirements and capture the full range of impacts (Ferreira-González et al., 2007). Standard practice is to weight each component endpoint equally, however, this does not reflect patient or trialist preferences (Stolker et al., 2014) and may lead to biased estimates (Montori et al., 2005).	Harrison et al. (2016) introduce a framework for exploring how sensitive climate impact studies are to the heuristic of vector addition. Domain-general statistical approaches are available to test for the presence of interactions/violations of additivity. Experimental approaches may be useful in some contexts (*e.g.* risk characterisation of mixtures in toxicology). Expert and patient elicitation procedures can inform the weighting of component endpoints where aggregation is desired in clinical trials.
Decision rules	Absolutes De minimis and de manifestis Feasibility rules Cost-effectiveness thresholds	The general concern with heuristic decision rules is that neglecting to weight the full costs and benefits of regulatory options (and their distributional features) can lead to decisions that are inefficient, harmful, or that violate equity principles. De minimis, de manifestis, and cost-effectiveness	Decision rules should be based on clearly articulated values and a robust empirical or theoretical basis, rather than conventions. Tipping points and planetary boundaries offer promise for informing risk thresholds. Cost-effectiveness thresholds should be linked to budgetary constraints or valuations of

(continued on next page)

Table 2 *(continued)*

Rule class	Rule description	Applications and problems	Towards rigorization
		thresholds often lack clear theoretical or empirical justifications. It is unclear why the feasibility criterion should trump concerns about health and welfare. Moreover, informal arguments suggest that feasibility rules may be prone to gaming and inadvertently lead to the entrenchment of technologies for risk reduction.	statistical lives. Exploiting interpretive latitude allows decision-makers to safeguard against the perverse implications of verbal rules.

relying solely on statistical rules to *remove* outliers – as opposed to identifying suspect data worthy of further investigation – is generally viewed as bad practice. This is reflected in the (possibly apocryphal) anecdote that the automated deletion of "rogue" zero ozone concentrations over the Antarctic – presumed to stem from measurement error – prevented early detection of depletion of the ozone layer (Benedick, 1991).

These variants of significance testing are subject to the standard critiques of: using arbitrary thresholds to discriminate between signal and noise[1]; neglecting or only indirectly considering statistical aspects that are logically informative for causal inference (*e.g.* effect sizes in clinical trials); and ignoring priors and utilities (when characterised as choice rather than pure inference) (Spiegelhalter et al., 2004; Suter, 1996). Other criticisms focus on the convention that significance testing is *sufficient* for formal uncertainty analysis (leaving systematic – rather than random – sources of error to be handled informally) (Greenland, 2005). This is a powerful critique given that many regulatory agencies – such as the FDA – that previously relied almost exclusively on data generated from randomised trials are increasingly basing their decisions on observational data (*e.g.* in health technology appraisal and post-market surveillance activities; Lash et al., 2016). In such contexts, causal inferences are far from secure, implying a need for formal methods that estimate the direction, magnitude and uncertainty associated with systematic rather than purely random errors (Lash et al., 2016). As a cautionary example, the FDA previously recommended the use of oestrogen to reduce cardiovascular risk in post-menopausal women on the basis of statistically significant findings of protective effects in *observational* studies. Subsequent randomised controlled trials (RCTs) found no benefit and possible harm. Formal bias analysis may have prevented this (Greenland, 2005).

Proponents of significance testing rules tend to justify them not on the grounds of their particular form, but rather by defending them *as rules*. That is, they are promoted as ways of limiting analyst bias and personal discretion, of ensuring consistency in inference in a policy arena that may otherwise become chaotic in the absence of firm standards (*e.g.* Mayo and Cox, 2010). However, the definition of the null hypothesis can leave significant degrees of freedom, particularly in dynamic systems – such as climate – that are characterised by multiple processes and feedback mechanisms operating at different scales (Frigg et al., 2015a; Cohn and Lins, 2005). This raises questions about the reliance on significance testing in climate attribution and detection studies (Frigg et al., 2015a).

3.3. Weighting rules

How to weight and perhaps combine diverse and often conflicting kinds of data is a core dilemma in regulation. We classify three basic approaches.

3.3.1. Gatekeeping rules

Gatekeeping rules discriminate between valid and unsound data, studies, or models by assigning some a weight of zero (*e.g.* excluding models from ensemble predictions based on a threshold measure of "skill"). Although gatekeeping clearly occurs in regulatory science, it tends not to be based on explicit rules, instead relying on a mixture of factors-based judgments, expert knowledge, and peer review. Exceptions exist. For example, we find regulatory rules designed to constrain analyses undertaken by private parties for reasons of ethics or methodological assurance. For instance, the FDA advises manufacturers to exclude data derived from phase one pharmacokinetic and pharmacodynamic studies from premarketing risk assessments. The grounds are that such studies are usually short term and conducted either on healthy or fairly ill subjects with refractory or terminal conditions, and so are likely to introduce bias (FDA, 2005). Similarly, the EPA may not rely on third party research on human subjects involving pesticides that violate certain clearly defined ethical principles (*e.g.* research involving intentional exposure of pregnant women). Finally, the EPA and FDA require that industry funded (toxicology) studies adhere to "Good Laboratory Practices" (GLP) – a mixture of technical prescriptions and generic quality management practices – before they can be used as a basis for quantitative risk assessment (Alcock et al., 2011). The logic here is to ensure a basic level of methodological quality and control for any incentives that manufacturers might have to downplay their products' risks, perhaps at the cost of neglecting data produced using novel experimental protocols (Myers et al., 2009). More widespread and analytically interesting than gatekeepers are weighting rules, discussed below.

3.3.2. Weighting hierarchies

These heuristics rank sources of potentially conflicting data according to the strength of their study designs. Their role is most prominent in healthcare, where insufficient attention to the quality of evidence can have serious implications. Recall that regulatory authorities previously recommended – on the basis of observational studies – that doctors encourage postmenopausal women to use hormone replacement therapy to reduce cardiovascular risk (Guyatt et al., 2008), a recommendation later overturned on the basis of RCT results. The problem here was that the strength of their initial recommendation didn't reflect the quality of the evidence. To guard against this, explicit hierarchies of evidence have been adopted in different jurisdictions to bring logic and consistency to the provision of medical guidance (*ibid.*). These hierarchies are used to inform the development of clinical guidelines (which set out, *e.g.*, preferred treatments for different conditions) and public health interventions (broadly conceived), rather than in health technology appraisal.[2] A widely used hierarchy is (Petticrew and Roberts, 2003):

1. Systematic reviews and meta-analyses
2. Randomised controlled trials with definitive results

[1] Gigerenzer and Marewski (2015) report that the convention of using 5% (and less often 1%) as a threshold of significance appears to stem from the fact that Ronald Fisher's nemesis, Karl Pearson, refused to give him tables for any other values.

[2] Health technology appraisal typically relies on the formal analysis of clinical and economic data to determine the relative cost-effectiveness of interventions. Here, a categorical ranking of evidence isn't much use, as different sources of data need to be combined into a common numerical summary.

3. Randomised controlled trials with non-definitive results
4. Cohort studies
5. Case-control studies
6. Cross sectional surveys
7. Case reports

One critique levelled against such hierarchies is that the ordering of study designs is excessively rigid and lacking in caveats (Petticrew, 2010; Rothman, 2014). The idea is that whilst, all else being equal, randomised controlled trials are preferable to cohort studies, case-controls, and so on, things are never quite equal. An RCT with low compliance rates and substantial missing data due to high patient dropout rates is probably less likely to lead to reliable causal inference than a well conducted observational study where the relevant covariates are known and recorded (Rubin, 2008). For example, where an RCT found that those urged to cease smoking developed more lung cancer than controls, the discrepancy between these results and the many observational studies linking smoking to lung cancer was ascribed to problems with the trial (Rothman, 2014), rather than the reverse as a rigid interpretation of hierarchies would suggest. Similarly, in a meta-analysis, Concato (2004) identified specific exposure–outcome associations that had been studied with both RCTs and observational studies, and found that average results from *well-designed* observational studies did not overestimate exposure-outcome associations compared to those reported in RCTs. Put more broadly, rigid hierarchies neglect the question of how well designed and conducted the *particular* study was (rather than the properties of the ideal study in that class; Harbour and Miller, 2001), its statistical properties (*e.g.* variance), as well as questions of how well suited it was to answering the specific question of interest (*e.g.* determining effect size *vs.* identifying causal mechanisms). Anecdotal data suggests that the rigid application of hierarchies may lead to actual harm. Two small RCTs found evidence of a protective effect of β-blockers on cardiac post-operative events (Neuman et al., 2014). Their perioperative use in patients with coronary artery disease was promoted as "best practice" on the basis of these trials. This was later overturned, largely on the basis of a large scale retrospective cohort study which associated the use of β-blockers with substantially greater mortality in patient subgroups (*ibid*).

Recent years have seen several public health agencies adopt a hierarchy (GRADE) that acknowledges these caveats, for instance in specifying the criteria that may be used to justify down-grading or up-grading the quality ranking of specific study categories (Guyatt et al., 2008). This kind of "recovery scheme" is not particularly well developed in chemical risk assessment, where hierarchies are also prevalent and sometimes interpreted inconsistently or in an overly mechanical fashion (MacGillivray, 2014). For example, one weighting hierarchy, used by the US Occupational Health and Safety Administration (OSHA), is that positive test results (whether human or animal) should supersede negative epidemiological studies (Jasanoff, 1982). The logic is that epidemiological studies typically have low power and thus are prone to false negatives. The OSHA later introduced criteria allowing negative epidemiological results to trump positive animal studies: if subjects should have been exposed for a minimum of 20 years; and observed for the following 30 years; and groups were large enough to detect an increase in cancer incidence of 50% above unexposed populations (Jasanoff, 1982). However, these criteria are so restrictive that it is implausible that they could ever be satisfied (*ibid.*).

3.3.3. Combination rules

Here, we are concerned with rules for combining expert judgments (*e.g.* probability distributions, or beliefs about causation), statistical summaries (*e.g.* of multiple clinical trials), model predictions, and so forth. Approaches to combination are often classified into mathematical (*e.g.* Cooke's "Classical Model," long used for volcano management in Montserrat) and behavioural (*e.g.* Delphi workshops) camps (Clemen and Winkler, 1999). Combination can be problematic where disagreement stems from different theoretical or methodological approaches (Knutti et al., 2010), but we set aside this concern to focus on the methods applied. One common heuristic is equal-weighting:

$$W_i = 1/N$$

(where W_i is the weight assigned to the individual model, estimate, expert, *etc.*; N is total number of models *etc*. being combined).

Here, different estimates, beliefs, or predictions are aggregated with the un-weighted mean treated as the true-value. In what has been called "model democracy," the IPCC combines – without weighting – the outputs of multiple models into simple averages, standard deviations or ranges (Knutti, 2010). More formal approaches are under development, for example the most recent IPCC report discussed Bayesian techniques for model-weighting (Flato et al., 2013), yet a problem is that the quality or performance of climate models is often quite context-specific, *e.g.* with some having good representations of the Indian monsoon, and others providing better estimates of precipitation in the Pacific North West. Elsewhere, regulatory agencies such as the USEPA often resort to equal weighting when faced with expert disagreement. For instance, when members of expert advisory boards have conflicting views about matters of fact (*e.g.* beliefs about causation, or whether a fish species is endangered, *etc.*), agencies routinely simply adopt the majority view, rather than sift through the arguments advanced, or weighting votes according to expertise (Vermeule, 2008). There is a formal justification for equal-weighting in situations of pure uncertainty. This is Laplace's principle of insufficient reason or indifference, which holds that equivalent knowledges be assigned equivalent degrees of belief (*ibid.*). In short, when the combiner of information is unable to determine which of several experts (or models) is more likely to be correct – based on (lack of) knowledge of their expertise or beliefs about their relative credibility – then it makes sense to weight them equally. Yet whilst equal-weighting may appear axiomatic in these restrictive circumstances, it suffers from problems that are well known in the philosophical literature, and that may carry practical implications.

One such problem is that it leads to paradoxical outcomes in the case where the objects being combined or weighted (*e.g.* the model prediction, or parameter value, *etc.*) may be legitimately represented in more than one mathematical form, and those forms are non-linearly related (Frigg et al., 2015b). An example of this is the speed at which ice falls out of clouds, an important parameter in climate modelling. This parameter can be represented in two ways – ice fall rate or the residence time of ice in clouds – and these two values are inversely related (Frigg et al., 2015b). The true value of these parameters is unknown, and when conducting sensitivity analysis within HadCM3, the UK's Met Office uses ice fall rate and implicitly assigns equal weights to the probability that it will take a given value within a plausible range (*i.e.* the pdf of its value is flat within a middle range; Sexton et al., 2012). Full elaboration of the paradox that this leads to would be rather technically involved (see Frigg et al., 2015b), yet suffice it to say that equal-weighting in this kind of case generates different outcomes depending on the (arbitrary) choice of which representation of the parameter to select.

This is perhaps a rather esoteric case. A larger problem is how robust the equal-weights heuristic is to violations of the pure uncertainty assumption. In simulation studies where data on the relative credibility of experts is available it tends to be slightly outperformed by mathematical approaches that do include weighting, whilst typically performing better than behavioural approaches to combination (Clemen and Winkler, 1999). Crucially, the performance of equal-weighting hinges on the size and diversity of the expert sample. In inadequately diverse samples, the rule understates uncertainty and may also create a biased estimate. Moreover, institutional context shapes the performance of the rule. Bearing this out, courts and legal scholars have taken a mixed view on "nose-counting" in regulatory policy, critiquing its application in situations where the size or composition of the expert

group can be gamed, and where decision making processes encourage groupthink (*e.g.* sequential rather than simultaneous voting by panel members; Vermeule, 2008).

In health technology appraisal equal-weighting is broadly seen as naïve, and different combination rules are deployed (Petitti, 1999). Consider meta-analysis, where the outputs of several randomised trials are combined to form a common numerical summary (*e.g.* effect size). This often involves studies of vastly ranging population samples (which by extension vary widely in informativeness), a fact which any reasonable weighting scheme has to take account of. There are two major approaches to weighting here, which differ in terms of whether the underlying studies are considered homogenous (Greenland, 1994; Doi et al., 2011). In fixed effects analysis, where the studies are considered homogenous, individual studies are typically weighted according to the inverse of their variance.

$$W_i = 1/se_i^2$$

(where se_i is standard error).

There is a formal proof that this minimises the overall variance of the combined estimate, and as there is a presumed absence of bias (the homogeneity assumption), this minimises the mean squared error of the combined estimate. However, homogeneity is a strong assumption, and where the rule is used outside of this scope constraint – as is not uncommon in medical research and environmental risk assessment (Riley et al., 2011) – it can introduce bias to the overall estimate, essentially acting as a heuristic. In random effects meta-analysis, homogeneity is not assumed, meaning that the studies may either incorporate bias or are measuring true differences in effect sizes. The standard approach here is to weight according to the within *and* between study variance. However, this weighting scheme has been critiqued for lacking a clear underlying rationale (Greenland, 1994). Moreover, as it is based on statistical measures of study *outputs*, it doesn't explicitly take into account the quality of the design and conduct of the studies, nor the completeness of the data reporting. Incorporating these factors within meta-analysis weighting schemes faces serious technical and conceptual difficulties (Greenland and O'Rourke, 2001). Pearl and colleagues have recently developed a deductive framework for combining datasets collected from different populations and experimental designs (*e.g.* observational *vs.* RCTs), however, implementing this approach requires a clear understanding of the underlying data generating processes (Bareinboim and Pearl, 2016). Perhaps a less demanding approach is to use sensitivity analysis to explore the robustness of meta-analysis outputs to different study characteristics (dimensions of quality) (Juni et al., 2001).

A related task is the combination of *different* risks or impact types to generate aggregate measures. One commonly adopted combination rule is vector addition, which is premised on an absence of interactions between risks or impact categories (*i.e.* no synergistic effects). It is standard within integrated impact assessment (of climate change), where impacts are typically characterised on a sector-by-sector basis (*e.g.* impacts on water resources, *vs.* impacts on the agricultural sector), then simply added together to derive the aggregate economic impact (Stern, 2007). Recent work suggests that this presumption of additivity may be skewing the outcomes of impact evaluations (and by extension, adaptation decision-making), *e.g.* through neglecting the way that changes in water availability influence the balance of irrigated and non-irrigated crops in a given area, which in turn influence food production (Harrison et al., 2016). Similar concerns about (neglecting) synergistic effects have a long-standing history in the risk assessment of chemical mixtures. Analogously, in the analysis of clinical trial outcomes multiple endpoints are often aggregated into composite measures to reduce sample size requirements and capture the full range of impacts of interventions, especially within cardiology (Ferreira-González et al., 2007). Standard practice is to weight each component equally, however, this does not appear to reflect patient or trialist preferences (Stolker et al., 2014), and may lead to biased estimates if the number of events in the more significant components are small and the magnitude of effect varies significantly across components (Montori et al., 2005).

3.4. Interpolation and extrapolation: model choice and implementation

We turn to heuristics for interpolating datasets and for extrapolating from existing data, touching upon dose-response modelling in toxicology, the analysis of clinical trials, and climate modelling. In these domains, model selection and implementation is typically not determined by the raw data, meaning that there are multiple plausible approaches and choices that can be made. These include:

- selecting between competing model structures (*e.g.* linear quadratic *vs.* threshold models in radiation risk assessment);
- the selection of variables (*e.g.* in regression analyses of epidemiological data, which variables need to be adjusted for to avoid confounding; *e.g.* Greenland et al., 1999);
- choosing whether particular value judgments should be explicitly incorporated within a model (*e.g.* should distributional (equity) weights be applied within health technology appraisal);
- adjusting for known biases or errors in the data (*e.g.* various adjustments account for differences in size and lifespan across species in chemical dose-response modelling);
- selecting the values for model parameters that cannot be derived from the dataset (*e.g.* discount rates, the selection of priors, *etc.*); and
- choosing the technique for implementing the model (*e.g.* parameter estimation technique).

There are many forms of reasoning that can guide these inferences. For instance, there are deductive principles for confounder adjustment (subject to strong causal assumptions); model selection could be based on *ad hoc* factors that are application specific (*e.g.* the format of the output data); and deciding upon value judgments might be a deliberative process. But we find heuristics governing interpolation and extrapolation in several domains. There are two kinds of these heuristics: defaults; and inference rules. The former prescribe the *outcome* of the inference, *e.g.* the USEPA's requirement to use linear non-threshold models for carcinogen risk assessment (NRC, 2009). The latter prescribe the *processes* by which inferences are reached, *e.g.* when faced with several plausible candidate models for dose-response analysis, select that with the lowest Akaike information criterion (an approach that balances complexity and fit) (EPA, 2012). Below we give a schematic overview of the reliance on interpolation and extrapolation heuristics in chemicals regulation, before offering comparisons with other regulatory domains. Our concern is less with the individual details of the rules (for this, see Greenland et al., 1999; Greenland, 2012a; Jurek et al., 2008; and MacGillivray, 2014), than with the phenomenon of rule-bound modelling.

In chemicals regulation, the determination of a (potential) causal link between a chemical and harm is the precursor to building formal dose-response models. Here, interpolation involves structuring and regimenting the raw test data into a dose-response curve (typically covering moderate-high dose levels), with the purpose of deriving a "point of departure" (POD) from which extrapolation to the low-dose range can then be made. Key inferences include: which species and endpoint should be used? Should a biological or empirical model be used? And which parameter estimation technique should be used to apply it? What constitutes adequate fit for a statistical model? Or should the dose-response data simply be plotted by hand (*e.g.* where the "no observed adverse effect level" is taken as the POD)? Rather than leave these decisions entirely to the discretion of risk assessors, regulatory agencies have adopted a series of inference rules and default heuristics to structure the process, most famously at the USEPA (NRC, 1983, 1994, 2009). These include rules governing the preferred endpoint and species to be used (the most sensitive), and for adjusting the

raw data prior to model application (*e.g.* animal doses are scaled by ¾ power of body weight to derive toxicologically equivalent dose for humans) (MacGillivray, 2014). There are also rules for selecting amongst competing model structures for both interpolation (*e.g.* AIC is used to select amongst candidate models in benchmark dose-response modelling) and extrapolation (*e.g.* linear *vs*. non-linear extrapolation are adopted for carcinogens and non-carcinogens respectively, with the latter involving back-of-the-envelope uncertainty factors). Uncertainty stemming from the *interactions* of these kinds of conventions has received remarkably little systematic attention in both risk assessment and the epidemiological literature (Greenland, 2005). Advances in multiple bias analysis are making this task more tractable (Lash et al., 2016; Greenland, 2005), drawing in part on deductive theories of causal inference (*e.g.* Pearl, 2000) which can clarify errors stemming from informal epidemiological heuristics.

There are structural parallels between dose-response modelling in chemicals regulation, radiation protection, and health technology appraisal. And whilst we find similar heuristics for extrapolation and interpolation across these domains, there are differences in the degree to which different jurisdictions and policy domains are bound by rules and also in the particular rules that are selected. For example, there is limited published guidance on the risk-benefit methodology used to determine whether pharmaceuticals should be allowed on the market, both within the EU and the US (NRC, 2013; Hughes et al., 2007). And the actual process in the US is more reliant on the deliberative evaluation of clinical and economic data by experts, rather than formal modelling (Garrison et al., 2007). In contrast, the appraisal of health technologies – used to determine whether treatments should be publicly funded or simply to establish preference orderings – is heavily bound by default rules in many jurisdictions, including the UK. Here we find rules dictating the preferred method for valuing changes in health status (EQ-5D), the source of preference data for valuing changes in health related quality of life (the public, rather than patients), the discount rates to be used (3.5% for costs and benefits), and whether equity should be incorporated within the formal analysis (in this case, no) (NICE, 2008). Rule-bound approaches to regulatory science are a fairly recent phenomenon, replacing an earlier emphasis on sound methodological choices on the part of analysts. How might we account for such variations – across time, jurisdictions, and policy domains – in the extent to which regulatory risk and decision analysis is rule-bound?

The simplest explanation would be that interpolation and extrapolation will be heavily rule-bound in areas where there is a broad consensus on the most empirically or theoretically justified inferences. There is something to this account, as, for example, the EPA's default heuristic that linear-at-low-dose models be used for carcinogens has a long-standing theoretical basis (NRC, 2009). But other extrapolation rules are more back of the envelope, such as the uncertainty factors used for non-carcinogens. And other heuristics, if not quite arbitrary, are not rigorously supported, and indeed vary across domains and jurisdictions (*e.g.* default discount rates, and rules for deriving them; favoured methods for deriving preferences, *etc.*). And so we gain in explanatory power when we consider that rule-bound analysis reflects a desire for *standardisation*, rather than simply a consensus on the *best* methods. The logic of standardisation is that any difference between analysis outcomes should reflect true differences in likely risks and consequences, rather than variation in methodological choices (Schlander, 2009). Yet if stability in outcomes is contingent on standardising methodological choices that are in principle contestable, then it is unclear whether the outputs have a clear physical interpretation. On this reading, the substantial use of defaults in regulatory science may create consistency in analysis outcomes at the expense of robustness and generalisability (c.f. Richter et al., 2009). To develop this idea, we turn to climate science.

Climate model choices are similarly not *determined* by prior theory or observational evidence, leaving substantial discretion *e.g.* in the selection of emissions scenarios; in choices of method for downscaling GCM outputs into a form that is useful for adaptation planning; how to correct for biases introduced by idealisations (*e.g.* the exclusion of changes in dynamical ice sheet processes within models of sea level rise; Brysse et al., 2013); how to extrapolate the results of impact analyses across nations with different geographic, social and political contexts; and how to combine different kinds of climate impacts in integrated assessments. Sensitivity analysis or robustness checks can clarify these uncertainties and potentially inform these choices. For example, Coley et al. (2012) explore the robustness of adaptation strategies in the built environment to alternative rules for selecting emissions scenarios (*e.g.* worst-case *vs*. expected), whilst Hawkins et al. (2013) compare alternative methods for bias correction in crop forecasting models, finding "change factor" approaches more robust than "nudging." However, the second-order problem is evaluating chains of heuristic inferences in model-construction, particularly where a series of choices are arbitrarily constrained.

For example, Morgan (2014) and Stanton et al. (2009) have critiqued the adoption of fixed model structures and fixed functional forms across integrated assessment models in climate science. They argue that these choices are typically fixed *within* a given model, and that these choices are fixed *across* models (*e.g.* most integrated assessment models share the assumption of a quadratic form of the damage function, with little explanation or justification; Stanton et al., 2009). The result is to artificially close down uncertainty. Supporting this claim, Morgan (2014) describes the introduction of an impact modelling framework within which analysts were allowed to alter these underlying model structures (*e.g.* discount rates, damage functions, and agents' objective functions). He found that the identification of the optimal climate mitigation policy was highly sensitive to alternative plausible model specifications. Perhaps the core issue here is that rule-bound approaches to modelling have limited justification in the social and environmental sciences, compared to the physical sciences. In the former domains, the state of theoretical knowledge and the often imprecise nature of observational data typically provide limited grounds for constraining parameter values, model structures, and choices of functional form. A further implication is that rules can become *entrenched* in ways that make it difficult to displace or circumvent them in practice, even in the absence of a formal standard-setting body.

3.5. Decision rules

3.5.1. Absolutes

Absolutes are categorical rules which deem the presence of a property as sufficient grounds for regulating an object. The skeleton form is:

- IF object has property X, INFER it poses an unacceptable risk, THEN regulate.

They are distinct from de manifestis rules – discussed later – in that they regulate on the basis of categorical properties (*e.g.* carcinogenic or not) rather than quantitative thresholds. The most (in) famous example is probably the US' Delaney Clause:

- "No additive shall be deemed to be safe if it is found to induce cancer when ingested by man or animal, or if it is found, after tests which are appropriate for the evaluation of the safety of food additives, to induce cancer in man or animals".

The rule became untenable as advances in toxicological testing and analytical chemistry revealed that there were many more carcinogens present in foodstuffs than initially expected and that there were marked differences in their potencies (Majone, 2010). Absolutist rules are now largely anachronistic in environmental and public health protection, replaced by a broad acceptance that decision rules should track the *level*

of the harms and benefits associated with chemicals, drugs, health interventions, as well as their alternatives (Graham and Wiener, 1998). Although still on the books, the influence of the Delaney Clause has been diluted by statutory changes and rule-avoidance strategies on the part of the FDA.

To elaborate, the rule contains latitude in determining what an appropriate test is, what it means to induce cancer, how test results should be interpreted, and even in what constitutes a food additive (Majone, 2010). The FDA has exploited this to avoid restrictions that it perceived as inefficient or unwise (*e.g.* arguing that secondary carcinogens – those that lead to tumour formation either through reacting with other chemicals or through disrupting a bodily function – do not fall under the scope of the rule as they do not directly *induce* cancer (Kessler, 1977)). This highlights an important feature of rules that take the form of verbal rather than mathematical statements, namely that they cannot entirely determine their own application, as there is always ambiguity surrounding their proper interpretation and scope. Vague terms in such rules are analogous to adjustable parameters (Daniel Steel, pers. comm) that can be tinkered with to incorporate domain specific knowledge and respond to circumstances not foreseen at the time of rule formulation.

3.5.2. De minimis and de manifestis

De minimis decision rules, based on the notion that "the law does not concern itself with trifles," set out thresholds of risk that are negligible (Peterson, 2002). De manifestis rules are thresholds of risk deemed unacceptably high. The two are often blurred in practice,[3] and we use risk thresholds as an umbrella term. We structure the discussion according to their underlying justifications: analytical capabilities (*i.e.* detection limits), arbitrary numerical limits, and threshold models of risk.

One historic practice was to tie de manifestis thresholds to analytical capabilities, *e.g.* early 20th century exposure standards for X-ray technicians were based on the level at which fogging on a photographic plate occurred, and similar approaches were widespread in US chemicals regulation until the 1970s (Rodricks, 1994; Rodricks et al., 1987). In practice this is similar to absolute decision rules (and carries the same problems), so is largely obsolete. However, an interesting variant is FDA's sensitivity of method approach, introduced as another workaround to the Delaney Clause (NRC, 1983). Under this approach, carcinogenic compounds can be used as animal-feed additives and veterinary drugs provided that "no residue" is found in the edible portion of the animal. Crucially, the FDA are authorized to specify which analytical method is to be adopted for particular substances, and they have used this flexibility to select methods whose detection limits roughly correspond with an upper-bound lifetime incremental cancer risk of 10^{-6}.

Although the FDA did not provide a formal justification for selecting 10^{-6} as a risk threshold, it has nevertheless spread to many US policy domains, albeit with some variations in interpretation (Rodricks et al., 1987).[4] When interpreted as a de minimis clause, the 10^{-6} rule plays a reasonable enough desk-clearing function. But when applied as a de manifestis decision rule, it can lead to strikingly inefficient regulation due to the fact that it neglects the size of the population exposed and by extension the benefits of risk reduction measures. This is particularly true for contexts where exposure to risk is geographically concentrated rather than diffusely spread (*e.g.* where the EPA uses individual risk thresholds to determine whether remediation is required of hazardous waste sites). Neglecting population size has been defended on equity grounds, the idea being that members of small populations have the same rights to protection as those concentrated in cities. However, minority groups tend to be clustered near hazardous waste sites, implying that ignoring population size may prove discriminatory, which hardly advances equity (Viscusi, 2000).

Another approach is to base risk thresholds on mechanistic thresholds below which no harm occurs (often with a safety margin). This is widespread in chemicals regulation for non-cancer endpoints. The theoretical basis is that below a certain dose, clearance pathways, cellular defences, and repair processes minimize damage and make the risk of harm negligible (NRC, 2009). A variant on this approach is the toxicological threshold of concern. This involves mapping across toxicity data (*e.g.* NOAEL thresholds) from structurally similar chemicals to identify de minimis exposures for untested chemicals within the same category. This has been used in both standard setting and screening in various jurisdictions. Although there is a clear logic for basing de minimis rules on mechanistic thresholds (if the risk is practically zero then it can be considered trivial), the same is not true for *de manifestis* rules (the presence of a risk is not necessarily unacceptable). However, the two notions are often blurred in practice, *e.g.* a hazard index of unity (1) is widely used to differentiate between acceptable and unacceptable risk of chemicals within the European Union and in the US under various statutes (*e.g.* Superfund, the Clean Air Act, *etc.*). This involves the neglect of dimensions including costs, feasibility, and equity, although there is a widespread belief that these factors are often considered behind closed doors, and justified officially by creative interpretive strategies (*e.g.* where the EPA held that transient and reversible health effects stemming from exposure to ozone should not be considered *adverse* under the Clean Air Act (Coglianese and Marchant, 2004)).

A recent variant of the threshold model approach are de manifestis rules based on *transitions* in system behaviour that would entail substantial adaptation costs – "planetary boundaries" and "tipping points." The former are thresholds in control variables within ecological processes beyond which there is the risk of irreversible and abrupt environmental change (Rockström et al., 2009). Deriving these thresholds is non-trivial. For example, whilst the historic EU policy target of maintaining warming within 2 degrees centigrade was sometimes proposed to reflect a planetary boundary, it had not been clearly linked with an actual threshold in system behaviour (*e.g.* melting of West Antarctic Ice Sheet; Randalls, 2010). Tipping points, by contrast, are not based on mechanistic considerations, but rather on generic statistical cues that precede shifts in system states (*e.g.* a reduced rate of recovery from perturbations; Scheffer et al., 2012). They could potentially be translated into decision rules for governing complex systems where causal relations are not understood in mechanistic terms.

3.5.3. Feasibility rules

Feasibility rules mandate the particular technology to be used for environmental protection, constrained in terms of the best that is available or feasible. Sometimes used in concert with de minimis or de manifestis thresholds, their general structure is:

IF a technology can be implemented to reduce the level of risk, AND it is economically feasible to do so, THEN require implementation of the (best) technology by regulated industry, ELSE consider risk acceptable.

Economic feasibility is broadly understood to mean an absence of significant harm for the regulated industry, although interpretations vary (*e.g.* OSHA interprets it as either a 1% decline in revenue or a 10% profit decline, whereas the EPA interprets it as relating to job losses, plant closures, and bankruptcy (Masur and Posner, 2010)). Proponents of feasibility rules point to their frugality and speed as virtues, with some further arguing (more speculatively) that such heuristics may roughly mimic utility maximisation, on the presumption that technol-

[3] For example, where the threshold discriminates between acceptable and unacceptable risk (*i.e.* it is both de minimis and de manifestis).

[4] The 10^{-6} threshold is used to determine whether remediation is required of hazardous waste sites, is applied within the Clean Air Act, and is also adopted in industrial chemicals regulation. At times it has been interpreted as relating to the maximally exposed individual, at other times understood as an average. In some cases it has been interpreted as a de manifestis clause rather than a de minimis threshold. And in some cases regulators treat the threshold as merely one consideration amongst many.

ogy development is attuned to the point at which the cost per pound of pollution reductions begins to rise sharply (Kysar, 2010). Critics have questioned why feasibility should trump concerns about health and wellbeing (*e.g.* where OSHA discarded a welfare-enhancing exposure limit for chromium on the grounds that it threatened the survival of at least one industry (Masur and Posner, 2010)), and argued that feasibility rules may lead to the entrenchment of existing technologies (Sunstein, 1990; Majone, 1984) and may be gamed by industry in ways that minimise the regulatory burden (*e.g.* where a larger number of smaller plants are maintained to make the EPA's plant shutdown clause more likely to be invoked (Masur and Posner, 2010)). Feasibility rules may also introduce a "market distortion" that protects small firms and low-profit industries (*ibid.*). A recurring theme of the above arguments is that the outcomes of decision rules are a function of the interaction between their structural form and the features of the environment in which they are applied (c.f. Gigerenzer and Todd, 1999).

3.5.4. Cost-effectiveness rules

These rules are closest to mimicking a full decision-theoretic approach as they require the explicit calculation of costs and a quantitative measure of effectiveness (Weinstein et al., 1996), although the latter need not be monetised.[5] They are often based on the Incremental Cost Effectiveness Ratio (ICER), a method used in comparative health technology appraisal in Canada, Australia, and the UK (Gafni and Birch, 2008):

$$ICER = (C2–C1)/(E2–E1)$$

(where C2 and E2 are the cost and effectiveness of the new technology, and C1 and E1 are the equivalent for the comparator. Effectiveness is widely measured in quality adjusted life years (QALY))

The following presumptive rules are followed in the UK (NICE, 2004):

- IF ICER < £20,000/QALY; THEN recommended;
- IF ICER £20,000/QALY – £30,000/QALY; THEN require consideration of other factors (*e.g.* degree of uncertainty, innovative features of technology, wider societal costs and benefits, *etc.*);
- IF ICER > £30,000/QALY, only recommend adoption if case for supporting the technology on these other factors is "increasingly strong."

No specific justification for these thresholds was given at their introduction, beyond that they were roughly consistent with past decisions taken by the agency. Indeed, they have been criticised for lacking a reasoned basis, not being clearly linked to budgetary constraints, and being inconsistent with thresholds used in other areas of health delivery (House of Commons, 2008). Post-implementation, statistical analyses (Martin et al., 2007) revealed that the lower end of the current thresholds is higher than the marginal costs of improving health in cancer and circulatory disease, implying that inefficient technologies may be approved at the expense of more efficient ones (NICE, 2007). The thresholds have also been critiqued for neglecting the distribution of health states and (particular) conceptions of fairness (Schlander, 2010). However, the thresholds are *presumptive*, leaving space for these factors to be considered. Moreover, NICE has introduced escape valves in the form of end-of-life QALYs and relating to ultra-orphan treatments, to guard against the controversial outcomes that efficiency-based rules can lead to.

[5] Some thresholds involve the monetisation of effectiveness, *e.g.* the US Nuclear Regulatory Commission's (1995) threshold for evaluating safety improvements for existing nuclear plants ($2000 per person-rem of exposure avoided) was based on a standard valuation of a (statistical) life.

4. Conclusions

Environmental and public health policy problems are often ambiguous, lack reliable data, and require urgent decisions. In such situations, full decision-theoretic analysis may be infeasible, and inexact, heuristic methods of analysis must be relied upon (Table 1). Heuristics are not problematic *per se* – induction necessarily depends on judgments that cannot be fully justified in a formal sense (Greenland, 2012a). But problems arise when we conflate heuristics for theorems, and apply them in ways that are insensitive to their assumptions, limitations and biases (summarised in Table 2). For example, some of the heuristics we discussed have been provided with formal justifications, however, these justifications require strong assumptions. These assumptions are often violated in practical applications, leading to skewed results and inefficient decisions. Other heuristics appeal to theoretical or empirical support, sometimes presumed, sometimes robustly tested. These heuristics can be seen as empirical or causal generalisations, and distinguished from canonical laws in terms of the range and frequency with which they hold true. In effect these heuristics contain implicit *ceteris paribus* clauses, the neglect of which may lead to serious error. Other heuristics are more like conventions and lack a clear underlying logic, but are relied upon to confer a sense of stability or consistency to an analytical process which may otherwise seem chaotic. Yet this consistency – particularly when applied to chains of inferences – may come at the cost of masking uncertainty and introduce bias. Rules of choice, in contrast, serve as surrogates for values such as utility maximisation or advancing equity. But the extent to which these decision heuristics actually track those underlying values is an empirical question, as many of these rules are based on arbitrary thresholds, neglect costs and benefits, and may be prone to gaming.

Rigorization can guard against such problems (Kitcher, 1981) through filling or closing the inferential gaps in a heuristic argument (Table 2). Basic principles include formalising the assumptions and scope conditions under which heuristics should be applied; testing rather than presuming their underlying empirical or theoretical justifications (*e.g.* relations between surrogates and objects of interest in screening rules); using sensitivity analysis, simulations, multiple bias analysis, and deductive systems of inference (*e.g.* DAGs) to characterise rule uncertainty (including that stemming from rule-interactions) and refine heuristics; establishing recovery schemes in situations where the direction of bias can be predicted or to introduce flexibility to rules that may be rigidly interpreted (*e.g.* evidence hierarchies); and basing decision rules on clearly articulated values and a robust empirical basis, rather than on arbitrary conventions.

Acknowledgments

David Spiegelhalter and Crispin Cooper provided helpful guidance and critiques of previous drafts. Met Office scientists were generous with their time during informal discussions. Anonymous reviewers provided helpful comments. The usual caveats remain.

References

Alcock, R.E., MacGillivray, B.H., Busby, J.S., 2011. Understanding the mismatch between the demands of risk assessment and practice of scientists – the case of Deca-BDE. Environ. Int. 37 (1), 226–235.
Ashcroft, M.B., Chisholm, L.A., French, K.O., 2009. Climate change at the landscape scale: predicting fine-grained spatial heterogeneity in warming and potential refugia for vegetation. Glob. Chang. Biol. 15 (3), 656–667.
Atkinson, A.J., et al., 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin. Pharmacol. Ther. 69, 89–95.
Bareinboim, E., Pearl, J., 2016. Causal inference and the data-fusion problem. Proc. Natl. Acad. Sci. 113 (27), 7345–7352.
Benedick, R.E., 1991. Ozone Diplomacy. 1991. Harvard University Press, Cambridge, Mass, pp. 19.
Brysse, K., et al., 2013. Climate change prediction: erring on the side of least drama? Glob. Environ. Chang. 23 (1), 327–337.
CGER (Commission on Geosciences, Environment and Resources), 1999. Improving

American River Flood Frequency Analyses. National Academy Press, Washington DC.
Clancey, W.J., 1983. The epistemology of a rule-based expert system—a framework for explanation. Artif. Intell. 20 (3), 215–251.
Clemen, R.T., Winkler, R.L., 1999. Combining probability distributions from experts in risk analysis. Risk Anal. 19 (2), 187–203.
Coglianese, C., Marchant, G.E., 2004. The EPA's risky reasoning. Regulation 16 (2), 16–22.
Cohn, T.A., Lins, H.F., 2005. Nature's style: Naturally trendy. Geophys. Res. Lett. 32 (23).
Coley, D., Kershaw, T., Eames, M., 2012. A comparison of structural and behavioural adaptations to future proofing buildings against higher temperatures. Build. Environ. 55, 159–166.
Concato, J., 2004. Observational versus experimental studies: what's the evidence for a hierarchy? NeuroRx 1 (3), 341–347.
Doi, S.A.R., Barendregt, J.J., Mozurkewich, E.L., 2011. Meta-analysis of heterogeneous clinical trials: an empirical example. Contemp. Clin. Trials 32 (2), 288–298.
Doll, R., 2002. Proof of causality: deduction from epidemiological observation. Perspect. Biol. Med. 45 (4), 499–515.
Dwan, K., et al., 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PloS One 3 (8), e3081.
Echt, D.S., et al., 2001. Mortality and morbidity in patients receiving encainide, flecainide, or placebo - the cardiac arrhythmia suppression trial. N. Engl. J. Med. 324 (12), 781–788.
EPA, 2012. Benchmark dose technical guidance. http://www.epa.gov/raf/publications/pdfs/benchmark_dose_guidance.pdf.
European Medicines Agency, 2016. Screening for Adverse Reactions in EudraVigilance. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2016/12/WC500218606.pdf.
FDA, 2005. Guidance for Industry: Premarketing Risk Assessment. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126958.pdf.
FDA, 2006. Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf.
Feigenbaum, E.A., 1977. In: The Art of Artificial Intelligence: 1. Themes and Case Studies of Knowledge Engineering. Proceedings of the Fifth International Joint Conference on Artificial Intelligence. 1977. pp. 1014–1029 (August).
Ferreira-González, I., et al., 2007. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 334 (7597), 786.
Finkel, A.M., 1997. Disconnect brain and repeat after me: “risk assessment is too conservative”. Ann. N. Y. Acad. Sci. 837 (1), 397–417.
Flato, G., et al., 2013. Evaluation of climate models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Climate Change 2013 5. pp. 741–866.
Fleming, T.R., DeMets, D.L., 1996. Surrogate end points in clinical trials: are we being misled? Ann. Intern. Med. 125 (7), 605–613.
Frigg, R., et al., 2015a. Philosophy of climate science part I: observing climate change. Philos Compass 10 (12), 953–964.
Frigg, R., Smith, L.A., Stainforth, D.A., 2015b. An assessment of the foundational assumptions in high resolution climate projections: the case of UKCP09. Synthese 192 (12), 3979–4008.
Gafni, A., Birch, S., 2008. Incremental cost-effectiveness ratios (ICERs): the silence of the lambda. Soc. Sci. Med. 62 (9), 2091–2100.
Garrison, L.P., Towse, A., Bresnahan, B.W., 2007. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. Health Aff. 26 (3), 684–695.
Gelman, A., Loken, E., 2014. The statistical crisis in science data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don't hold up. Am. Sci. 102 (6), 460.
Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annu. Rev. Psychol. 62, 451–482.
Gigerenzer, G., Marewski, J.N., 2015. Surrogate science: the idol of a universal method for scientific inference. J. Manag. 41 (2), 421–440.
Gigerenzer, G., Sturm, N., 2007. In: Ash, Sturn (Eds.), Tools = Theories = Data? On Some Circular Dynamics in Cognitive Science, in Psychology's Territories: Historical and Contemporary Perspectives From Different Disciplines, pp. 305–342.
Gigerenzer, G., Todd, P.M., 1999. Simple Heuristics that Make Us Smart. Oxford University Press.
Graham, J.D., Wiener, J.B., 1998. Risk versus Risk: Tradeoffs in Protecting Health and the Environment. Harvard University Press.
Greenland, S., 1994. Invited commentary: a critical look at some popular meta analytic methods. Am. J. Epidemiol. 140 (3), 290–296.
Greenland, S., 2005. Multiple-bias modelling for analysis of observational data. J. R. Stat. Soc. A. Stat. Soc. 168 (2), 267–306.
Greenland, S., 2012a. Commentary: intuitions, simulations, theorems: the role and limits of methodology. Epidemiology 23 (3), 440–442.
Greenland, S., 2012b. Causal Inference as a Prediction Problem: Assumptions, Identification and Evidence Synthesis. John Wiley & Sons, Ltdpp. 43–58.
Greenland, S., O'Rourke, K., 2001. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2 (4), 463–471.
Greenland, S., Pearl, J., Robins, J.M., 1999. Causal diagrams for epidemiologic research. Epidemiology 37–48.
Guyatt, G.H., et al., 2008. Rating quality of evidence and strength of recommendations: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Br. Med. J. 336 (7650), 924–926.
Harbour, R., Miller, J., 2001. A new system for grading recommendations in evidence based guidelines. Br. Med. J. 323 (7308), 334–336.
Harrison, P.A., et al., 2016. Climate change impact modelling needs to include cross-sectoral interactions. Nat. Clim. Chang (in press).
Hawkins, E., et al., 2013. Calibration and bias correction of climate projections for crop modelling: an idealised case study over Europe. Agric. For. Meteorol. 170, 19–31.
Higgins, J.P., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., Savović, J., Schulz, K.F., Weeks, L., Sterne, J.A., 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 343, d5928.
Hill, A.B., 1965. The environment and disease: association or causation? J. R. Soc. Med. 58 (5), 295–300.
House of Commons, 2008. Health Committee report: National Institute for Health and Clinical Excellence. http://www.publications.parliament.uk/pa/cm200708/cmselect/cmhealth/27/27.pdf.
Hughes, D.A., Bayoumi, A.M., Pirmohamed, M., 2007. Current assessment of risk–benefit by regulators: is it time to introduce decision analyses? Clin. Pharmacol. Ther. 82, 123–127.
Ioannidis, J.P., et al., 2014. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383 (9912), 166–175.
Jasanoff, S., 1982. Science and the limits of administrative rule-making: lessons from the OSHA cancer policy. Osgoode Hall Law J. 20, 536–561.
Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press.
Juni, P., Altman, D.G., Egger, M., 2001. Systematic reviews in healthcare: assessing the quality of controlled clinical trials. Br. Med. J. 323 (7303), 42–46.
Jurek, A.M., Greenland, S., Maldonado, G., 2008. Brief Report How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? Int. J. Epidemiol. 37 (2), 382–385.
Kahneman, D., Frederick, S., 2002. Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich, T., Griffin, D., Kahneman, D. (Eds.), Heuristics of Intuitive Judgment: Extensions and Applications. Cambridge University Press, New York.
Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press.
Kaplan, S., Garrick, B.J., 1981. On the quantitative definition of risk. Risk Anal. 1 (1), 11–27.
Kazi, D.S., Hlatky, M.A., 2012. Repeat Revascularization is a Faulty end Point for Clinical Trials.
Kessler, D.A., 1977. Implementing the anticancer clauses of the food, drug and cosmetic act. Univ. Chicago Law Rev. 44 (4), 817–850.
Kitcher, P., 1981. Mathematical rigor—who needs it? Noûs 469–493.
Knutti, R., 2010. The end of model democracy? Clim. Chang. 102 (3–4), 395–404.
Knutti, R., et al., 2010. Challenges in combining projections from multiple climate models. J. Clim. 23, 2739–2758.
Kysar, D., 2010. Regulating From Nowhere: Environmental Law and the Search for Objectivity. Yale University Press.
Lash, T.L., et al., 2016. Quantitative bias analysis in regulatory settings. Am. J. Public Health 106 (7), 1227–1230.
Majone, G., 1984. Science and trans-science in standard setting. Sci. Technol. Hum. Values 9 (1), 15–22.
MacGillivray, B.H., 2014. Heuristics structure and pervade formal risk assessment. Risk Anal. 34 (4), 771–787.
Majone, G., 2010. Foundations of risk regulation: science, decision-making, policy learning and institutional reform. Eur. J. Risk Regul. 1, 5–19.
Martin, S., Rice, N., Smith, P.C., 2007. The Link Between Healthcare Spending and Health Outcomes: Evidence From English Programme Budgeting Data. http://www.york.ac.uk/che/pdf/rp24.pdf.
Masur, J.S., Posner, E.A., 2010. Against feasibility analysis. Univ. Chicago Law Rev. 77 (2), 657–716.
Mayo, D.G., Cox, D., 2010. Frequentist statistics as a theory of inductive inference. In: Mayo, D.G., Spanos, A. (Eds.), Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science. Cambridge University Press, pp. 262.
Montori, V.M., et al., 2005. Validity of composite end points in clinical trials. BMJ 330 (7491), 594–596.
Morgan, M.G., 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc. Natl. Acad. Sci. 111 (20), 7176–7184.
Myers, J.P., et al., 2009. Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: the case of bisphenol A. Environ. Health Perspect. 117 (3), 309–315.
Nazaroff, W., Weschler, C.J., Little, J.C., Hubai, E.A.C., 2012. Intake to production ratio: a measure of exposure intimacy for manufactured chemicals. Environ. Health Perspect. 120 (12), 1678–1683.
Neuman, M.D., Bosk, C.L., Fleisher, L.A., 2014. Learning from mistakes in clinical practice guidelines: the case of perioperative β-blockade. BMJ Qual. Saf. 23 (11), 957–964.
NICE (National Institute for Clinical Excellence), 2004. Guide to the Methods of Technology Appraisal. http://www.nice.org.uk/niceMedia/pdf/TAP_Methods.pdf.
NICE (National Institute for Clinical Excellence), 2007. Briefing Paper for the Methods Working Party on the Cost Effectiveness Threshold. http://www.nice.org.uk/media/4A6/41/CostEffectivenessThresholdFinalPaperTabledAtWPMeeting5Sep3907KT.pdf.
NICE (National Institute for Clinical Excellence), 2008. Guide to the Methods of Technology Appraisal. http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf.
Nichols, A.L., Zeckhauser, R.J., 1986. The perils of prudence: how conservative risk assessments distort regulation. Regulation 10, 13.
NRC (National Research Council), 1983. Risk Assessment in the Federal Government: Managing the Process. National Academies Press, Washington, DC.
NRC (National Research Council), 1994. Science and Judgment in Risk Assessment. National Academies Press, Washington, DC.

NRC (National Research Council), 2009. Science and Decisions: Advancing Risk Assessment. National Academies Press, Washington, DC.
NRC (National Research Council), 2012. Ethical and Scientific Issues in Studying the Safety of Approved Drugs. National Academies Press, Washington, DC.
NRC (National Research Council), 2013. Environmental Decisions in the Face of Uncertainty. National Academies Press, Washington, DC.
Pearl, J., 1984. Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley, MA.
Pearl, J., 2000. Causality: Models, Reasoning and Inference. 29 MIT Press, Cambridge.
Pearl, J., 2014. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.
Peterson, M., 2002. What is a de minimis risk? Risk Manage. 4 (2), 47–55.
Petitti, D.B., 1999. Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine. Oxford University Press.
Petticrew, M., 2010. The process of systematic review of public health evidence: quality criteria and standards. In: Evidence-based Public Health: Effectiveness and Efficiency. 327.
Petticrew, M., Roberts, H., 2003. Evidence, hierarchies, and typologies: horses for courses. J. Epidemiol. Community Health 57 (7), 527–529.
Pólya, G., 1990. Mathematics and Plausible Reasoning: Induction and Analogy in Mathematics. 1 Princeton University Press.
Polya, G., 2004. How to Solve It. Princeton University Press.
Randalls, S., 2010. History of the 2C climate target. WIREs Clim. Change 1 (4), 598–605.
Richter, S.H., Garner, J.P., Würbel, H., 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat. Methods 6, 257–261.
Riley, R.D., Higgins, J., Deeks, J.J., 2011. Interpretation of random effects meta-analyses. Br. Med. J. 342, 964–967.
Rockström, J., et al., 2009. A safe operating space for humanity. Nature 461, 472–475.
Rodricks, J.V., 1994. Risk assessment, the environment, and public health. Environ. Health Perspect. 102 (3), 258–264.
Rodricks, J.V., Brett, S.M., Wren, G.C., 1987. Significant risk decisions in federal regulatory agencies. Regul. Toxicol. Pharmacol. 7 (3), 307–320.
Rothman, K.J., 2014. Six persistent research misconceptions. J. Gen. Intern. Med. 29 (7), 1060–1064.
Rothman, K.J., Greenland, S., 2005. Causation and causal inference in epidemiology. Am. J. Public Health 95 (S1), S144–S150.
Rubin, D.B., 2008. For objective causal inference, design trumps analysis. Ann. Appl. Stat. 808–840.
Savage, L.J., 1972. The Foundations of Statistics. Courier Dover Publications.
Scheffer, M., et al., 2012. Anticipating critical transitions. Science 338 (6105), 344–348.
Schlander, M., 2009. Reference case. In: Kattan, M.W. (Ed.), Encyclopaedia of Medical Decision Making. Sage Publications, Inc.
Schlander, M., 2010. Measures of efficiency in healthcare: QALMs about QALYs? Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen 104 (3), 214–226.
Sexton, D.M., et al., 2012. Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. Clim. Dyn. 38 (11 − 12), 2513–2542.
Simon, H.A., Newell, A., 1958. Heuristic problem solving: the next advance in operations research. Oper. Res. 6 (1), 1–10.
Smith, L.A., Stern, N., 2011. Uncertainty in science and its role in climate policy. Philosophical transactions of the royal society of London a: mathematical. Phys. Eng. Sci. 369 (1956), 4818–4841.
Smith, J.B., et al., 2001. Vulnerability to climate change and reasons for concern: a synthesis. In: McCarthy, J.J. (Ed.), Climate Change 2001: Impacts, Adaptation and Vulnerability. IPCC Working Group II, Cambridge University Press, Cambridge, pp. 914–967.
Spiegelhalter, D.J., Riesch, H., 2011. Don't know, can't know: embracing deeper uncertainties when analysing risks. Phil. Trans. R. Soc. A 369 (1956), 4730–4750.
Spiegelhalter, D.J., Abrams, K.R., Myles, J.P., 2004. Bayesian Approaches to Clinical Trials and Health-care Evaluation. John Wiley & Sons.
Stanton, E.A., Ackerman, F., Kartha, S., 2009. Inside the integrated assessment models: four issues in climate economics. Climate Dev. 1 (2), 166–184.
Stern, N.H., 2007. The Economics of Climate Change: The Stern Review. Cambridge University press.
Stolker, J.M., Spertus, J.A., Cohen, D.J., Jones, P.G., Jain, K.K., Bamberger, E., Lonergan, B.B., Chan, P.S., 2014. Re-Thinking Composite Endpoints in Clinical Trials: Insights from Patients and Trialists. Circulation 130 (15), 1254–1261.
Sunstein, C.R., 1990. Paradoxes of the regulatory state. Univ. Chicago Law Rev. 57 (2), 407–441.
Suter, G.W., 1996. Abuse of hypothesis testing statistics in ecological risk assessment. Hum. Ecol. Risk. Assess. 2 (2), 331–347.
Suter, G.W., Cormier, S.M., 2016. Bias in the development of health and ecological assessments and potential solutions. Hum. Ecol. Risk Assess. Int. J. 22 (1), 99–115.
US Nuclear Regulatory Commission, 1995. Reassessment of NRC's Dollar per Person-rem Conversion Factor Policy. http://pbadupws.nrc.gov/docs/ML0634/ML063470485.pdf.
Vermeule, A., 2008. The parliament of experts. Duke Law J. 58, 2231–2276.
Viscusi, W.K., 2000. Risk equity. J. Leg. Stud. 29, 843–872.
Weinstein, M.C., et al., 1996. Recommendations of the panel on cost-effectiveness in health and medicine. J. Am. Med. Assoc. 276 (15), 1253–1258.
Wimsatt, W.C., 2006. Reductionism and its heuristics: making methodological reductionism honest. Synthese 151 (3), 445–475.