EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

31 August 2010
EMA/549682/2010 - Revision 1
Human Medicines Development and Evaluation

Benefit-risk methodology project

Work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment

7 Westferry Circus • Canary Wharf • London E14 4HB • United Kingdom
Telephone +44 (0)20 7418 8400 **Facsimile** +44 (0)20 7418 8613
E-mail info@ema.europa.eu **Website** www.ema.europa.eu

An agency of the European Union

Table of contents

Executive summary

This paper reviews approaches for balancing benefits and risks in decision making about medicinal products. Each approach is defined, illustrated with a case study, and set in a theoretical or empirical context. Our view of its usefulness to regulators, for decisions at both pre- and post-approval stages, completes each review.

The paper begins by establishing working definitions of 'benefit' and 'risk', then details criteria of logical soundness, comprehensiveness, acceptability of results, practicality and generativeness for evaluating the approaches.

A generic qualitative approach of eight steps for decision making, PrOACT, is presented as it might apply to decision-making by regulators, followed by descriptions of three approaches currently under development: PhRMA BRAT, CMR CASS study and FDA BRF.

The paper continues with descriptions of 18 quantitative approaches under four major headings: simulation, models, statistics and measurement methods.

The review concludes with four suggestions to guide work in WP3:

1. Any quantitative method or approach requires a qualitative framework within which the model can be effectively developed. Indeed, the qualitative approach may be sufficient by itself for simpler benefit/risk decisions.
2. Only three quantitative approaches are sufficiently comprehensive to enable the benefit-risk balance to be represented numerically (as a difference or a ratio) by incorporating the value or utilities of favourable and unfavourable effects, along with probabilities representing the uncertainties of those effects: Bayesian statistics, decision trees and influence/relevance diagrams, and multi-criteria decision analysis (MCDA).
3. Five other approaches, while more restricted in scope, may well prove useful for particular cases: probabilistic simulation when the focus is on uncertainty of effects; Markov processes and Kaplan-Meier estimators for changes in health states over time; QALYS for modelling multiple health outcomes; and conjoint analysis to explicate trade-offs among effects, especially for eliciting patient preferences.
4. Combinations of approaches will prove useful in situations characterised by more than one of the following issues: the magnitude of favourable effects, the seriousness of unfavourable effects, uncertainty about the effects, transitions in health states and the time spent in each state, and trade-offs between effects.

Disclaimer

[*] Revision 1: This revision refers to the update of the disclaimer in order to state the opportunity for public consultation

1. Introduction

The purpose of this paper is to provide a brief description of all approaches that have appeared in the literature for balancing benefits and risks in drug regulatory decision making, and to present our view of each approach. We began this task by searching the literature and listing every suggestion, but we soon discovered that the approaches could be classified first into two distinctly different categories: qualitative and quantitative approaches. The quantitative approaches are further articulated into the four sub-categories: simulation, models, statistical approaches, and measurement methods. Within some of these sub-categories, we distinguish between static and dynamic approaches. A static approach represents some aspect of reality at a slice in time, while dynamic approaches capture changes in the reality over time. We provide our view of the potential usefulness of each approach to regulatory decision making about medicinal products both before and after approval.

Several assumptions underlie our work. First, we recognise that for decision making, human judgement plays an essential role; tools, models or processes by themselves are insufficient. Roles for judgement include framing the decision problem, identifying the relevant features of the problem, agreeing what evidence is relevant to the decision, forming preferences about the relative desirability of favourable effects and the undesirability of unfavourable effects, and assessing uncertainty about the effects. Second, we take for granted that the final decision to recommend approval of a drug, or not, is an act of human judgement, so any of the approaches reviewed here must in some way provide an aid to those making the decision. The approach does not make the decision; an individual or group does. Decision aids can do no more than assist human judgment; they do not make the judgments. Third, we assume that balancing benefits with risks, and the process of deciding to recommend approval or not, demands that multiple sets of information be processed and combined to make the final judgment that, overall, benefits exceed risks sufficiently to approve.

Balancing benefits and risks is no small task: a regulatory authority might receive a dossier of 10GB or more, parts of which are farmed out to relevant experts, then reassembled and discussed, without the help of any models, to decide 'yes' or 'no'. However, a substantial literature exists showing that people are limited in the amount of information they can combine intuitively, and the problem is particularly acute for integrating uncertainties [2-4], so this is where the most egregious errors occur. Fortunately, the research literature shows that model-based, structured approaches to problem-solving can not only avoid or correct the biases and errors of intuitive aggregation, they can also yield solutions that were not evident initially to any of the participants in the process [5].

These three assumptions led us to search for approaches that can aid, supplement and enhance human judgement on the one hand, and also provide support for combining the relevant information on the other. Let experts take a problem apart into its pieces and exercise judgement to turn data into useful information, but then let a computer or other decision aid put the pieces back together. This should make the benefit-risk judgement more explicit, more communicable, and possibly smarter and quicker than if no aids had been employed.

Finally, we note that much of the drug approval process is carried out in groups, so we considered the extent to which an approach could provide support to both individuals and groups in their work.

1.1. The meaning of 'benefit' and 'risk'

In reviewing this literature we frequently stumbled when trying to understand an author's meaning in explaining issues of benefit-risk balance. For example, consider a drug that can lower the incidence of heart attacks in otherwise healthy people. "There is a risk this drug won't lower your risk and there are risks from taking the drug." Three possible meanings here: the possibility the drug won't work for an individual, the chance of a heart attack, and side effects. Indeed, our interviews in 2009 with 55 people at six European drug approval agencies confirmed the observation reported 11 years earlier in CIOMS IV [6] that "There is no standard, widely acknowledged definitions of the terms *benefit* and *risk* as applied to medicine and particularly to medicinal products…".

We have, therefore, avoided these terms in this review. Instead, we have adopted the terms from Work Package 1 to the CHMP's *Assessment Report Guidance* [7], as summarised in Figure 1.

Figure 1. The EMA's four-fold model of 'benefits' and 'risks'

Favourable effects	**Uncertainty of favourable effects**
Unfavourable effects	**Uncertainty of unfavourable effects**

Definitions
Favourable effects are any beneficial effects for the target population (often referred to as "benefits" or "clinical benefits") that are associated with the product.

Unfavourable effects are any detrimental effects (often referred to as risks, harms, hazards both known and unknown) that can be attributed to the product or that are otherwise of concern for their undesirable effect on patients' health, public health, or the environment.

Uncertainties about both types of effects arise from variation, important sources of bias, methodological flaws or deficiencies (including GCP, compliance, etc.), unsettled issues, and limitations of the data set, e.g., due to sample size, study design, or duration of follow-up.

It will be important in reading this paper to keep in mind that a favourable effect could mean an expected efficacy outcome as well as a clinically-meaningful consequence, such as the elimination of an existing disease state, or the prevention of a negative consequence in a healthy person. Curing tuberculosis and preventing heart attacks are both favourable effects. Unfavourable effects are usually side effects, which can include the elimination of normal, healthy effects.

Within this framework, balancing benefits against risks is a matter of comparing the favourable and unfavourable effects, with an account of how that comparison is affected by consideration of the uncertainties. The remainder of this paper covers approaches to the assessment or measurement of favourable and unfavourable effects, and their uncertainties, as well as methods for combining the features of the four-fold model to effect a comparison of benefits with risks.

1.2. Evaluation criteria

The approaches are evaluated by considering the following criteria, which are based on past experience in evaluating decision models [8] and experience gained from Work Package 1:

Logical soundness

- The overall benefit-risk evaluation is decomposed into separate elements that are demonstrated theoretically and/or empirically to be meaningful.
- The elements are recombined according to a theoretically sound rule.
- The approach is coherent, that is, it ensures that related decisions based on the approach do not contradict each other or the objectives that are to be met.
- The approach aids rational thinking about benefits and risks.
- The approach gives results that do not change relative evaluations when alternatives are added or removed.

Comprehensiveness

- The approach can handle any form of data, continuous or discrete, qualitative or quantitative data, objective or subjective.
- The approach can accommodate uncertainty and value judgements, time preferences and risk attitudes.
- The approach makes multiple objectives and trade-offs explicit.

Acceptability of results

- The approach provides consistency checks that identify inconsistencies in the data and in people's judgements.
- The outputs of the approach should be understandable and interpretable in the user's terms, readily understandable and in quantitative form to facilitate comparison between options.
- The approach should be 'scrutable' in that it should make sense to anyone using it and be seen as a realistic way to evaluate benefits and risks.

Practicality

- Implementation of an approach should be economical in the use of participants' time.
- The approach should be easy to teach and easy to use.
- Additions or deletions to the approach should be possible without having to re-do existing inputs.
- Extending a model based on the approach should grow linearly with its inputs.
- Computer support should be available for any approach, enabling the user to make changes quickly and provide immediate feedback. The functionality of the software should include clear and effective graphical displays, and support for sensitivity analyses.

Generativeness

- The output of the approach should link clearly to action.
- The approach should provide a clear audit trail so that all aspects of the benefit-risk evaluation can be traced.
- The approach should develop insight and promote learning about benefit-risk evaluation.
- The approach should transform a fragmented, covert benefit-risk evaluation into an overt structure and set of rational processes.
- The results should be readily communicable and easily understood.

No approach satisfies all these criteria. Indeed, many of the criteria are not even relevant to some of the approaches. Thus, the following evaluations will make use of only those criteria that are relevant to the approaches.

2. Qualitative approaches

Several organisations are developing qualitative approaches to benefit-risk decision making. This is, of course, the first step in applying quantitative modelling: structuring the problem [9]. An eight-step generic framework, PrOACT, developed by Hammond, Keeney and Raiffa [10] and applied to decision making in health care by Hunink et al [11], provides a generic problem structure, which is here adapted to benefit-risk decision-making by regulators.

1. PrOBLEM. Determine the nature of the problem and its context: what is the medicinal product (e.g., new or marketed chemical or biological entity, device, generic); what sort of decision or recommendation is required (e.g., approve/disapprove, restrict); who are the stakeholders and key players; what factors should be considered in solving the problem (e.g., the therapeutic area, the unmet medical need, severity of condition, affected population, an individual's social context, time frame for outcomes). Then frame the problem (e.g., as mainly a problem of uncertainty, or of multiple conflicting objectives, or as some combination of the two).
2. OBJECTIVES. Identify objectives that indicate the overall purposes to be achieved (e.g., maximise favourable effects, minimise unfavourable effects), and develop criteria against which the alternatives can be evaluated (i.e., what are the favourable and unfavourable effects?).
3. ALTERNATIVES. Identify the options (actions about a medicinal product or the products themselves) to be evaluated against the criteria (e.g., pre-approval: new treatment, placebo, active comparator; post-approval: do nothing, limit duration, restrict indication, suspend).
4. CONSEQUENCES. Based on available data, describe how the alternative would perform on the criteria (e.g., describe the magnitude of possible favourable and unfavourable effects). It may be helpful to consider intermediate outcomes, such as safety and efficacy effects. Consequences describe clinically relevant effects. Create a 'consequence table' with alternatives in rows and criteria in columns. Write descriptions of the consequences in each cell, qualitative and quantitative. (See the reference in 'Our view' at section 3.2.8, below.) It may at this stage be helpful to record the basis for uncertainties about the consequences in preparation for step 6, if relevant.
5. TRADE-OFFS. Assess the balance between favourable and unfavourable effects.

These five steps are common to all decisions in which the consequences are known with certainty. In approving drugs, regulators typically must face uncertainty and risk, in which case three additional steps are relevant:

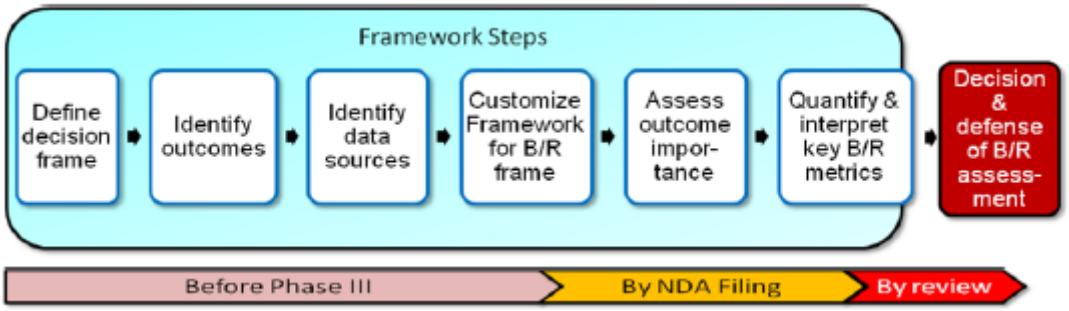
6. UNCERTAINTY. Consider how the balance between favourable and unfavourable effects would change by taking account of the uncertainty associated with the consequences.
7. RISK. Judge the relative importance of the Agency's risk attitude for this medicinal product (by considering, e.g., the therapeutic area, the unmet medical need and patients' concerns) and adjust the uncertainty-adjusted balance between favourable and unfavourable effects accordingly. Consider, too, how risks would be perceived by stakeholders (according to their views of risk).
8. LINKED DECISIONS. Consider the consistency of this decision with similar past decisions, and assess whether taking this decision could impact future decisions either favourably or unfavourably (e.g., would it set a precedent or make similar decisions in the future easier or more difficult).

This eight-step framework will be used throughout the remainder of the paper, as it provides a definition of what is meant by a comprehensive approach.

2.1. PhRMA BRAT

Since 2006, the PhRMA Benefit-Risk Action Team (BRAT), a collection of academics, regulators and pharmaceutical staff interested in improving decision making in pharmaceutical industries and regulatory agencies, has developed a broad framework that could inform the steps in the benefit-risk evaluation process. As can be seen in Figure 2, the six-step process can guide the development of a new medicine, help to interpret available clinical evidence and serve to improve communications between a pharmaceutical company and the regulator, with particular emphasis on the benefit-risk balance. It is currently being tested.

Figure 2. Steps in the BRAT



Source: Unpublished report by PhRMA-BRAT.

2.2. CMR CASS study

The CMR International Institute for Regulatory Science has been exploring approaches for benefit-risk assessment in various workshops since 2002. A six-step framework reported by Walker et al [12] requires identifying options and benefit-risk criteria, organising the latter in a value tree, scoring options on the criteria, weighting the criteria, calculating weighted scores at each level in the value tree and overall, and conducting sensitivity analyses. A subsequent paper by Liberti [13] et al gives a similar five-step qualitative framework that omits the assessment of numerical scores and weights. This latter approach is being tested by a task force from Health **C**anada, **A**ustralia's Therapeutic Goods Administration, **S**wissmedic and the **S**ingapore Health Science Authority, the CASS Group.

2.3. FDA BRF: Mullen and Korvick

While work on an integrating approach is still in progress, the currently-proposed framework is shown in Table 1 [14]. The FDA's stated intention is to provide "a high-level snapshot – the 'big picture' – of the issues relevant to the regulatory decision." It is intended to provide a standardized structure, which can be updated as new information is received, and to help focus discussions on the evidence and improve consistency in assessments. As it was developed from the mental models of regulators, the framework makes explicit current regulator perspectives of important topics that are considered in any benefit-risk assessment. It can be used throughout the regulatory process, and could facilitate communicating decisions outside the FDA.

Table 1. The FDA's developing framework for identifying key issues in the benefit-risk deliberations of regulators

Consideration	Favorable benefit-risk	Non-contributory	Unfavorable benefit-risk
Severity of condition			
Unmet medical need			
Clinical benefit			
Risk			
Risk management			

Source: Unpublished presentation by John Jenkins, 23 April 2010.

Our view

It would be premature to comment on any of these approaches as they are still in development or testing. However, it is worth noting that none makes explicit mention of the last three steps, U-R-L, in the PrOACT model. Uncertainty in particular is of serious concern to regulators [15]. Ignoring it is possibly due to the influence of multi-criteria decision analysis (MCDA); certainly both the BRAT and CMR approaches acknowledge MCDA, and they appear to have been guided by the steps shown in Figure 6.1 in *Multi-Criteria Analysis: A Manual*[8], none of which mention uncertainty. As the discussion later in Chapter 6 of the Manual makes clear, uncertainty and risk can be accommodated in MCDA, when they are relevant, but that seems to have been omitted in the transfer to frameworks for drug regulation.

3. Quantitative approaches

Quantitative approaches tend to focus on formal models and numbers, and often leave out some of the steps included in the qualitative frameworks, particularly in framing the problem at the start. It is for this reason, that none of the following approaches is in itself comprehensive; an act of judgement is needed, for a start, just in choosing the way to model a problematical situation.

At this writing, three reviews of quantitative approaches have appeared in print. Four methods were presented and discussed in October 2007 at a workshop sponsored by the Office of Health Economics in London [16]. Mussen, Salek and Walker [17] reviewed three approaches, and Guo described 12 quantitative benefit-risk methods. All of these are included in the 18 methods described and evaluated here.

3.1. Simulation

This approach attempts to mimic the behaviour of a system. For example, a flight simulator mimics the behaviour of an aircraft and can be used to train pilots. The cockpit of the simulator, which is mounted on hydraulic jacks, looks and feels like the interior of a plane. Special curved-mirror displays outside the cockpit provide realistic distance vision scenes to both pilot and co-pilot. Input is required from a human operator while the simulation is operating. This simulation attempts to mimic the whole system, plane, environment and pilot. However, other simulations are more restricted in scope, as is evident in the probabilistic simulation approach below.

3.1.1. Discrete-event simulation

Dynamic simulation models use differential equations and continuous variables, working at different levels of detail, from the interactions of physiological processes to the interaction of patients with the health care system. The complexity of these models provides challenges to transparency and validation of the model [18], though Brandeau [19] claims that they support decision making, not that they predicts events with certainty. The most comprehensive discrete-event simulation model for health care, Archimedes, is described by Eddy and Schlessinger [20] as "broad, spanning from biological details to the care processes, logistics, resources, and costs of health care systems." At this writing, 12 different diseases or conditions have been modelled with Archimedes (see: http://archimedesmodel.com/index.html).

Krishna [21] proposes its use for benefit-risk assessment. He explains that Archimedes is configured to deal well with predicting cardiometabolic risk, and "is one of the most advanced commercially available tools for predicting risk in diabetes." However, he points out several shortcomings, such as the "lack of confirmatory data related to adverse and side effects," and he also comments on the lack of validation based on completed outcome studies. He suggests that discrete-event simulation is worth developing as an aid to pharmaceutical companies in assessing the risk of new molecular entities. As for its use by drug regulators, Eddy himself admits that Archimedes will never replace clinical trials for evaluating the favourable and unfavourable effects of new medicines [22].

Our view

In general, simulation models are logically sound and comprehensive, although many are stronger on the URL stages of PrOACT than the multi-criteria aspects. That said, their outputs can be very useful in describing possible future outcomes and consequences. However, the models are typically large, complex and non-transparent when the system being simulated is complex, as is the case with weather simulation models. In addition, if the reality they are simulating is not static, then new data require modifications and extensions of the model to improve its validity, so keeping the model updated is costly.

Archimedes appears to be logically sound in modelling events, but may be less than comprehensive in it ability to model alternatives, their consequences and trade-offs. Thus, it is premature to suggest its use for regulatory decisions about approving drugs. Eddy and Schlessinger themselves point out that a major reason for clinical trials is that they can throw up surprises which are beyond current knowledge, which no simulation model can anticipate. Even so, the very high correlations, 0.97 to 0.99, between outcomes observed in trials and those predicted by Archimedes, coupled with the potential for the model to include practical issues like physician and patient behaviours (like non-compliance), suggest that at some point the model may help regulators to examine real-world deviations from clinical trial results. This would be especially true for decisions about drugs when adverse signals have been received after approval. The complexity of these models makes it difficult to understand why any particular results were obtained, so their outputs may not be acceptable, although the models can help users to generate new insights. A watching brief may be advisable for large-system simulations like Archimedes.

3.1.2. Probabilistic simulation (Credence decomposition)

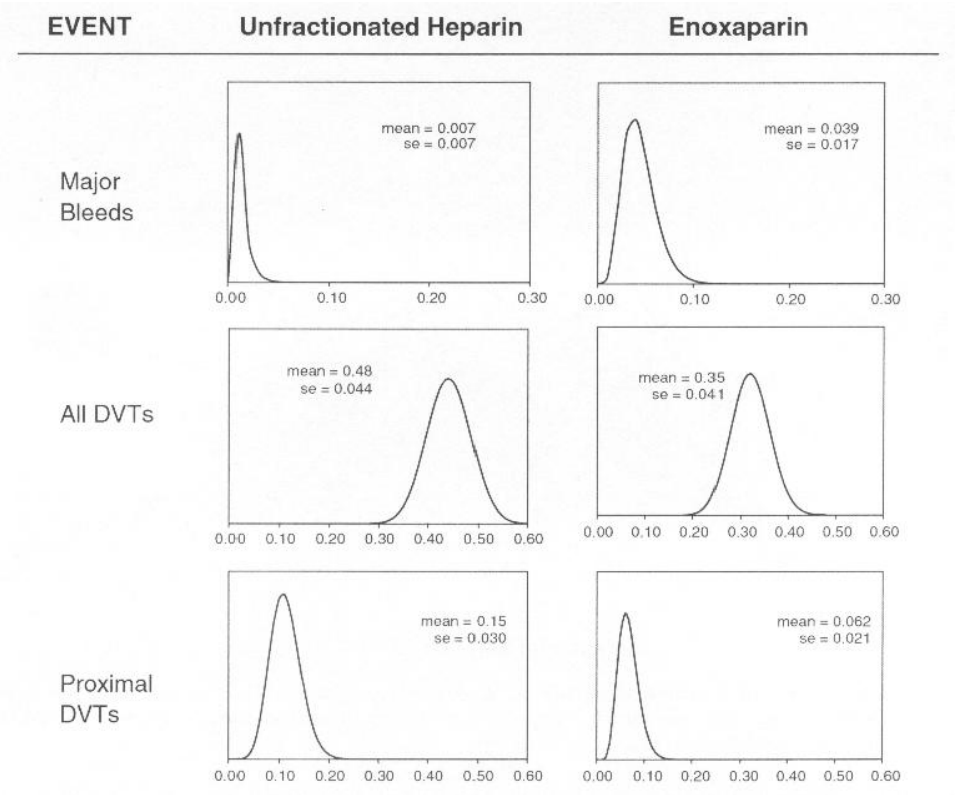
Uncertainty about the overall benefit-risk balance is decomposed into uncertainty associated with the separate benefit and risk criteria, probability distributions are assessed for each criterion, and the

overall result is obtained by Monte Carlo simulation. The approach is also known as probabilistic simulation [23], or risk analysis [24].

For example, Lynd and O'Brien [23] compared the reduced chance of deep vein thrombosis (DVT) associated with enoxaparin compared to unfractionated heparin, as against the increased chance of a major bleed. Because these chances are not fixed values for all patients, probability distributions derived from the clinical data were used to describe the variability across patients in the clinical trials, as shown in Figure 3. Random sampling from those probability distributions defined 3,000 different combinations of chances, resulting in as many pairs of favourable effects (benefit: reduced chances of DVT), and unfavourable effects (risk: increased chances of major bleed). Those 3,000 data points were plotted on a graph of incremental benefits versus incremental risks and overlaid with various thresholds indicating different judged trade-offs between numbers of DVTs compared to major bleeds. Any one of those trade-off judgements equates a unit of benefit to a unit of risk so that benefits can be compared meaningfully to risks.

The simulation showed "there is a 10% probability that the number of major bleeds induced by enoxaparin is greater than the number of proximal DVTs averted, which exceeds the conventional frequentist threshold of p = 0.05." Thus, taking account of the uncertainty in both favourable and unfavourable effects gave a different result from traditional significance tests, which showed no significant difference in the chance of a major bleed and a significant difference (p = 0.012) in the favourable effects of enoxaparin. The authors concluded that "This analytic approach to risk-benefit evaluation shows that, depending on the threshold for risk, even when two therapies are not statistically significantly different in terms of risk or benefit, most often there remains a nonzero probability that there is a difference between therapies."

Figure 3. Probability density functions of the population proportion of patients experiencing a major bleed, a proximal DVT, and any DVT occurring in patients treated with unfractionated heparin or enoxaparin



Source: Lynd & O'Brien, Figure 3, page 799.

The Lynd-O'Brien study dealt with just two drugs, and single risk and benefit criteria. Additional risk and benefit criteria can also be accommodated, as Lynd et al [25] have shown in a study of alosetron compared to placebo for irritable bowel syndrome. This simulation was based on point estimates rather than probability distribution; the simulation is of patients going through a 1-year treatment (or placebo) regime, so it might better be classified as a dynamic simulation. It is included here simply to illustrate one way to deal with issues of trade-offs between benefits, between risks and between risks and benefits. In this case, multiple thresholds were invoked to ensure the comparability of all units of benefits and risks. This moves simulation into the arena of multi-criteria decision analysis, which is certainly doable, but has not yet been reported in the literature. An opportunity exists for combining these two approaches to assist regulators in simultaneously looking at the effects of both uncertainty and multiple objectives with no loss of data.

Our view

Probabilistic simulation could prove to be useful for regulatory agencies, both pre- and post-approval. The approach is based soundly on probability theory, is comprehensive in the scope of inputs, provides readily interpretable results, and can be implemented using existing software, such as @Risk or Crystal Ball sitting in Excel, or Analytica. Its outputs are clear, graphical and easy to understand, though some will not be familiar to regulators. This approach can display two-dimensional probability distributions for the differences between a new drug and a placebo or a comparator for either measures of favourable or unfavourable effects, or the two combined.

Currently, regulatory decisions are informed by looking at statistical summaries such as means, risk ratios and confidence intervals, which ignore the richer information found in entire probability distributions. Particularly when probability distributions are skewed, this extra information might lead to a decision that is different from one that relies on single estimates, as Savage forcefully points out [26]. Recent developments in near-instantaneous simulation technology now make it possible for these differences of probability distributions to be calculated as quickly as the click of a mouse [27].

3.1.3. System dynamics

System dynamics is an approach to understanding the behaviour of complex systems in which feedback and time delays create non-linearity that is difficult to understand in simple cause-and-effect terms [28]. The approach is usually implemented using computer simulation. As far as we know, it has not been used, or even suggested for use, in modelling regulatory decisions about medicinal products.

Our view

There may be potential for applying systems dynamics, particularly for post-approval decisions, as data would be available about the time sequence of health states, compliance by patients, and other factors that could create a non-linear system.

3.2. Models

The approaches in this section cover methods for decomposing a problematical situation into its constituent pieces, for making the necessary assessments about the pieces and then recomposing the pieces into a whole that facilitates, or actually expresses, the benefit-risk tradeoff.

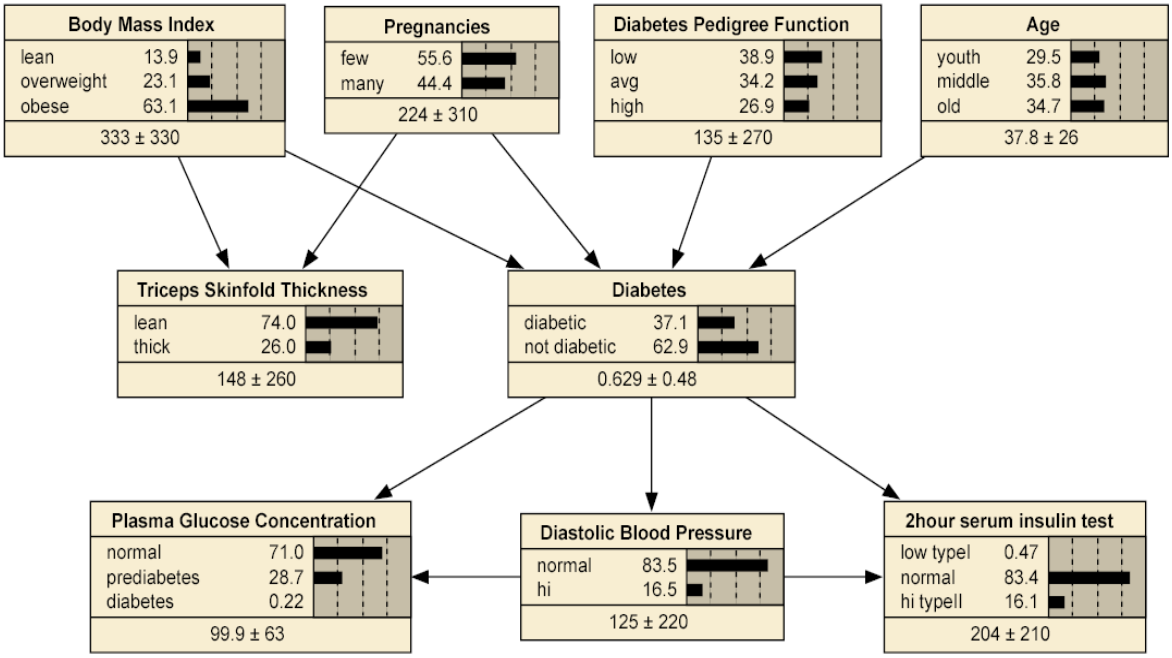
3.2.1. Bayesian belief networks

Used to model uncertainty, BBNs depict uncertainty about events and quantities represented at nodes in directed graphs, with arrows between the nodes denoting the conditioning of one node on another [29]. This conditioning can represent either causal or diagnostic relationships. For example, one node could represent the overall benefits of a drug, and another node the risks, with a target node providing, say, an overall assessment that a drug is both safe and efficacious. Arrows from the benefit and risk nodes would connect to the observable benefits, and to the observable risks, and any of those nodes might be connected if they are not conditionally independent of each other.

The network of connected nodes is known as a directed graph, which expresses the conditional relationships, or lack of them, amongst the nodes. Conditional probabilities, derived from data or judgements, or a combination of both, are entered into the model to express the relationships. In use, many items of observed data for a particular case are entered into the model, and the model then propagates the relevant probabilities throughout the network to give revised probability distributions at all nodes.

These models are particularly useful for situations in which the structure of the model, the interconnections and the conditional probabilities remain the same from one problem to the next, but where the actual pattern of data is different. If these conditions are met, then the BBN can learn as new problems are presented to it, updating its conditional probabilities so that the accuracy of the model improves over time. The example in Figure 4 is for diagnosing diabetes.

Figure 4. Directed graph for diagnosing diabetes



Conditions in the top row are predispositions to diabetes, with measurable indicators in the bottom row nodes. Clicking on a single symptom in any one or all of the three bottom-row indicators revises the probabilities shown in the Diabetes node, as will choosing the relevant predispositions in the top four nodes and the one at middle left. It is instructive to see how a single selection causes probability changes in the other nodes; some will change substantially, others not at all. As more data are collected the probability of diabetes becomes more definitive.

Source: "Diabetes Learned" in Netica Library.

Our view

This approach is basically a network of conditional probabilities, but for unique drugs the required data will be limited or absent. There might be some scope for eliciting the required probabilities using expert judgment, provided that the drug is of a class in which relevant relative frequency data are available. BBN's are usually limited to modelling uncertainty, with perhaps a single outcome measure, but they can provide inputs to more comprehensive decision models. We see more potential for applying BBNs to the approval of me-too drugs and in post-marketing decisions.

3.2.2. Bayesian statistics

At the heart of this approach to statistics is the definition of probability as a degree of belief, which leads to a crucial role for Bayes's Theorem in providing the means for revising the degrees of belief as new information from confirmatory trials is obtained. The initial probabilities, which summarise all relevant evidence, such as from exploratory trials, are known as 'prior probabilities,' while the resulting revised probabilities, which take account of the confirmatory trial data, are known as 'posterior probabilities.' Those posterior probabilities form a complete expression of the uncertainty attending the scientific inference about a hypothesis or a treatment effect that was the reason for the confirmatory trial.

In contrast, the Neyman-Pearson approach to statistical inference, which currently dominates statistical practice in regulatory decision making, provides a way of making statistical inferences without regard to prior probabilities, but still requires the exercise of judgement in other matters. Professor Pearson, quoted in Savage [30], indicated what he and Neyman were thinking in the mid-1920s:

"We were certainly aware that inferences must make use of prior information and that decisions must also take account of utilities, but after some considerable thought and discussion round these points we came to the conclusion, rightly or wrongly, that it was so rarely possible to give sure numerical values to these entities that our line of approach must proceed otherwise."

Today, statisticians of all persuasions recognise judgement is an essential ingredient of inference (e.g., a study cannot be powered without some knowledge of the size of the effect), and the old debates have largely disappeared from the literature as Bayesians have answered their traditionally-minded critics. Methods for assessing probabilities and utilities are now well developed [31-32] and numerous applications in many areas, including medicine [33], show the benefits of the Bayesian approach.

Significance levels, which are commonly used for reporting inferences in clinical trials, are probability statements about data rather than about the hypotheses and uncertain quantities that are directly relevant to decisions. Consequently significance levels cannot be formally integrated into decisions about benefits and risks. However, provided that the sufficient statistics of clinical trials are reported (i.e., summary statistics that contain all the information necessary to make a proper inference) it is possible to develop decision-relevant posterior probabilities using 'non-informative' prior probabilities. Ashby and Smith [33] argue that evidence-based decision making in medicine requires an integration of evidence, uncertainty and the utility of outcomes, which is provided by the Bayesian approach.

Our view

Bayesian statistics focuses on valid inferences from evidence, providing probabilities as part of the wider whole that is Bayesian decision theory. Although traditional significance testing still dominates in regulatory decision making, the data used to calculate the significance levels can be used to determine relevant posterior probabilities for decision models. As they

are based on probability theory, posterior probabilities are logically sound, and readily understandable to regulators, but Bayesian statistical models do not generally deal with multiple criteria. However, integrated into decision models that do include multiple criteria, probabilities are an essential ingredient for sound regulatory decision making both pre- and post-approval.

3.2.3. Decision trees & influence/relevance diagrams

These are models that incorporate, in diagrammatic displays, decisions (options), subsequent uncertain events, consequences and multiple criteria describing the consequences [34]. Decision trees show these as branching structures, like trees tipped on their sides, with roots (decisions) at the left, and branches to the right showing possible outcomes of the uncertain events, followed by more decisions, etc., until finally the tree is chopped of at the right, representing some time in the future when consequences will be apparent. This form of representation typified the early textbooks on decision analysis [32], and are still widely used today [35], largely because they can represent almost any decision situation whatever the topic. They are often seen in the journal *Medical Decision Making*.

One problem with decision trees is that they can expand exponentially as more and more nodes are included, thereby becoming very complex. Influence/relevance diagrams, which are graphical networks of decision, event, consequence, criteria and 'no-forgetting' (time sequence) nodes, are more compact representations than decision trees. (Bayesian belief networks are special cases of influence diagrams that mostly depict only event nodes.) Arrows connect nodes that 'influence' each other in a causal sense, though diagnostic relationships are also possible (knowing the time shown on my wristwatch reduces my uncertainty about the time on yours, though there is no direct causal relationship between our watches, only a diagnostic one). To accommodate both causal and diagnostic relationships, these structures are also known as 'relevance' diagrams (the time on my watch is relevant to my knowing the time on yours, but there is no direct causal connection between them). Stonebraker [36] reports how Bayer used an influence diagram to guide data collection for information relevant to a new blood-clot-busting drug. The elements of a decision tree and its associated influence/relevance diagram, for assessing a regulatory decision about approval of vaccines for the 2009 swine flu pandemic are shown in Figures 5 and 6, respectively.

Both decision trees and influence diagrams are models derived from decision theory, which is a theory grounded in Frank Ramsey's concept of coherent preference [37]. As elaborated by Savage [30], the theory starts by establishing self-evident axioms of preferences that exhibit simple consistency principles, like transitivity: if A is preferred to B, and B is preferred to C, then A should be preferred to C. From these axioms, three theorems are established: (1) probabilities exist, (2) utilities exist, and (3) the alternative associated with the highest expected (weighted average) utility should be most preferred. Essentially, this theory shows that coherent preference logically implies that just two quantities are needed for decisions: numbers that express the relative values of possible consequences, and numbers showing how likely these consequences are to occur. Multiplying utilities by their associated probabilities and summing those products over all consequences for a given alternative provides an expected utility figure that is a guide to action. Note that it is not just utilities that are compared, nor probabilities, but probability-weighted utilities, i.e., expected utilities.

The theory is thought by some to be flawed because the preferences of real people are not always coherent, but that argument fails to recognise that while these axioms logically imply the theorems, so the theorems also logically imply the axioms. Thus, decision analysts start by decomposing a complex problem into its elements, then assessing probabilities and utilities about the relevant pieces, and finally reassembling the pieces using the expected utility calculation. That result is examined to help people form their preferences. In other words, decision analysis is used to help people construct

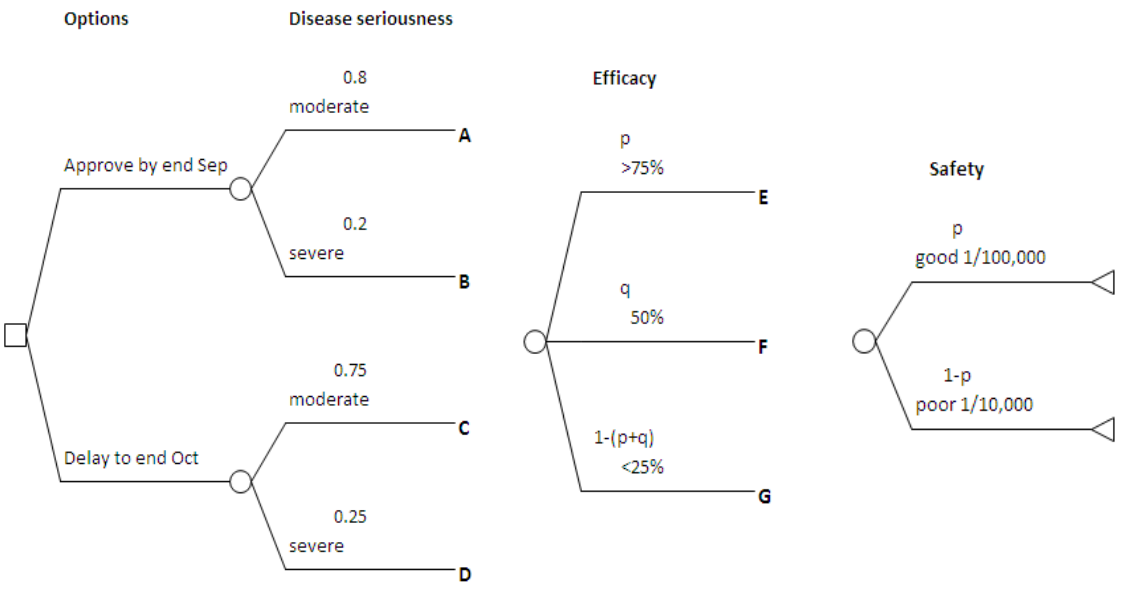
coherent preferences; it doesn't start by assuming preferences are coherent. Thus, any of the models derived from decision theory are classed in this paper as logically sound. Indeed, any model that does not exhibit in-built potential for violating consistency is classed here as logically sound.

Our view

From a theoretical point of view, decision trees and influence/relevance diagrams are certainly logically sound. They are particularly applicable to unique situations, and they can integrate data from many sources (in the form of probabilities of outcomes at each event node) and value or utility judgements (of the consequences). It is also possible for a single event node to be modelled as a Bayesian belief network, and for multi-criteria decision analysis to model consequences characterised by multiple objectives. Decision trees are often found in medical decision making papers to extend data-based statistical inferences to include costs and payoffs of taking a decision.

However, building decision trees and influence/relevance diagrams is an art, so care and experience is needed to ensure a realistic representation of the problem facing regulators, and some problems are so complex that the time and effort to build a decision tree are not easily justified. On the other hand, if the problem is very complex, unaided human judgement can also be questioned as an acceptable alternative.

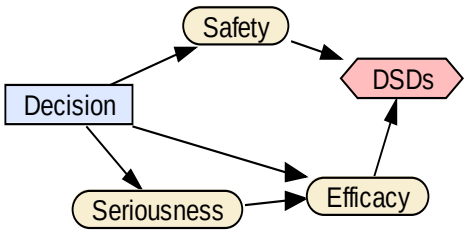
Figure 5. Decision tree for approving swine-flu vaccines



These are the elements of a decision tree on 1 September 2009 for approving swine flu vaccines by the end of September or delaying until the end of October. The left branches show the decisions followed by disease seriousness events and their probabilities. Events about efficacy and safety of the vaccines follow. The efficacy node, with its three outcome branches, attach at points A through D; the safety node and its two outcome branches attach at points E through G. This creates 24 scenarios of decisions and event outcomes. The 24 triangles at the end of the branches indicate the end of the decision tree, where they receive the number of deaths and serious debilitating adverse events appropriate for the outcomes of the uncertain events in that path through the tree.

Source: Phillips et al[1].

Figure 6. Influence diagram of the swine-flu decision tree



Directed arrows indicate the direction of relevance. The probabilities of Seriousness, Safety and Efficacy depend on the decision taken; the probabilities of Efficacy depend also on Seriousness, but the probabilities of Safety and Efficacy are conditionally independent of each other.

Source: Phillips et al[1].

3.2.4. Evidence-based benefit and risk model

This model, suggested by Jürgen Beckman [38] of the Federal Institute for Drugs and Medical Devices in Germany, depicts benefits as a 'box' whose sides are efficacy, responder rate and evidence for benefit. Risks are shown as several boxes, one for each ADR, with the dimensions "seriousness of the ADR, if it occurs," "frequency of the ADR in the exposed patient population" and "evidence for the respective risk." A two-armed balance depicts the relative weight of the benefit box under one arm against the sum of the weights of the risk boxes under the other arm.

Our view

This is, of course, a simple multi-criteria model, with all risks and benefits evaluated on three criteria each. Two of the benefit criteria, efficacy and responder rate, are favourable effects, while the third dimension, evidence for benefit, captures uncertainty about the favourable effects. Similarly, risks show two unfavourable effects, and an uncertainty. Beckman admits it is a simplified model, but suggests that "many properties of drugs can often be seen as irrelevant if not dissimilar so that only a few 'box sides' remain to be compared." His approach might be useful in displays of multi-criteria analyses, especially if the boxes are coloured to represent a fourth dimension. It might then be possible visually to balance benefits with risks.

However, the approach is silent on how the dimensions can all be expressed in equivalent units of value, which is essential for assessing and comparing the total volumes of the boxes. Thus, it is an incomplete technology, but worth thinking about for graphical displays of any multi-criteria analysis in summarising data analyses for both pre- and post-approval decisions. Dr James Felli has developed graphical displays showing box cars of different sizes and colours in a railway train that have helped managers at Lilly visualise their development portfolio and make decisions about it (personal communication), so the simplicity of this approach should not be summarily dismissed.

3.2.5. Incremental net health benefit

The net improvement in a patient's health state from taking a drug over a placebo, comparator or current treatment can be represented by a single number, incremental net health benefit (INHB), in this simple equation:

$$INHB = (FE_d - FE_c) - (UFE_d - UFE_c)$$

The difference in the unfavourable effects, UFE_d - IFE_c between the drug, d, and the comparison, c, is subtracted from the difference in favourable effects, FE_d - FE_c, when all effects are measured in the same units, which allows multiple effects to be considered. The previously-mentioned study by Lynd, Najafzadeh et al [25] applied this approach using the common metric of RVALYs (relative value-adjusted life-years), but QALYS (see section 3.2.9) or utilities or any other health outcome metric could be used.

Our view

INHB is one version of a multi-criteria model, which is discussed below. In the case of the Lynd, Najafzadeh et al [25] study, they extended the model to include multiple favourable and unfavourable effects, and used simulation to capture the uncertainty associated with the effects. Overall, a good demonstration of the power of applying multiple approaches to modelling. On its own, however, INHB deals only with multiple criteria, so is restricted in comprehensiveness.

3.2.6. Markov processes

Markov models capture the dynamic element of processes that develop or change over time, such as the progression of a disease in a patient[39]. Each stage in the developing process is represented by a node, a clinical state, and progression to the next one of several possible nodes is given by a probability distribution across the several nodes. Thus, the time-dependencies among event probabilities and state utilities can be represented. By running the model many times, it is possible to see at any given time in the future what proportion of trials were found in the relevant states, such as full health and death, thereby providing an insight into the probability that an individual patient will arrive at a particular future health state.

Theoretically speaking, a decision tree analysis should arrive at the same result as a Markov model that includes alternative decisions. The advantage of the Markov representation is that it more clearly shows the transition of patients from one health state to another over one time period to the next. An excellent example was reported by Thompson et al [40], who used a Markov model, displayed as a decision tree, to examine the relative long-term effectiveness of natalizumab over interferon β-1a in the treatment of MS. The model showed the superiority of natalizumab in terms of the utilities of health gains as compared to the small disutility of health losses from developing PML. In conclusion, Thompson et al conclude that "Understanding the long-term risks and benefits of treatment has never been more important given the serious limits to the old paradigm of short-term clinical trials, FDA approval, and weak postmarketing oversight."

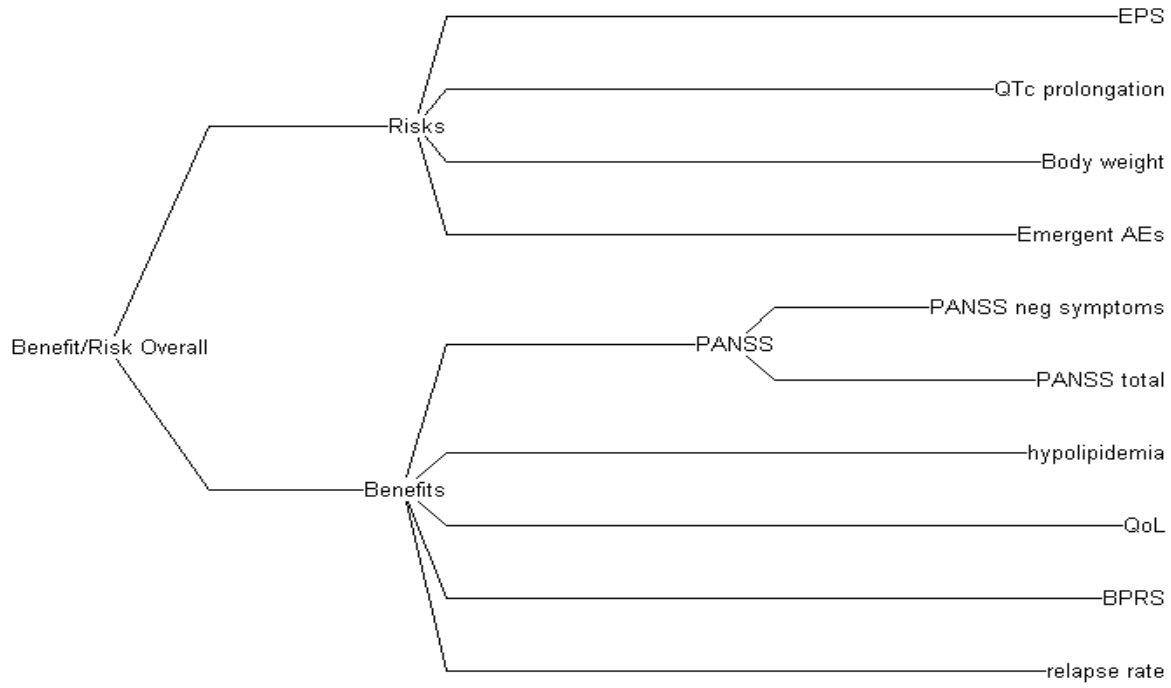
Our view

Markov models require clearly-defined health states and conditional probabilities of moving between the health states (as well as utilities or preference values of the health states). Clearly, assessments of these probabilities will be better if data are available over many years. Therefore, short-term clinical trials may yield insufficient confidence in the observed proportions of rare events like PML, limiting the use of Markov models for pre-approval decision making. The potential for Markov models may be greater for post-approval decisions. In any event, as we have already commented favourably on decision trees, we see Markov models as a useful extension enabling the dynamic modelling of disease progression.

3.2.7. Multi-criteria analysis

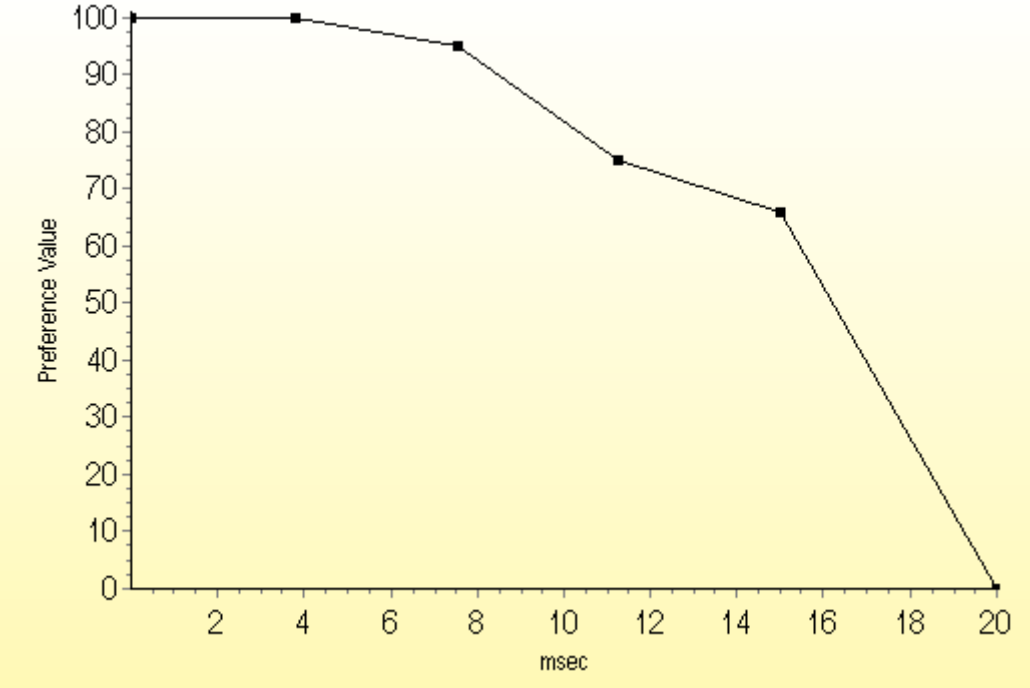
This general class of models accommodates decision making with multiple objectives. A review appears in Belton and Stewart [41], including the approach based on decision theory, multi-criteria decision analysis (MCDA), which was first introduced in 1976 by Keeney and Raiffa [42]. MCDA, like decision trees and influence/relevance diagrams, is based on Decision Theory (See 3.2.3). An accessible introduction can be found in Chapter 6 of the publicly-available report *Multi-Criteria Analysis: A Manual* [8]. The main purpose of MCDA is to bring together evaluations of options on different criteria into one overall evaluation. It does this through two separate processes: scoring and weighting. Scoring is the process of measuring the value of options, one criterion at a time, using scaling techniques borrowed from psychology [43], in particular psychometrics (psychological measurement of a person's abilities, attitudes, personality, etc.) and psychophysics (the measurement of sensation caused by an objectively-measurable stimulus). Weighting ensures that the units of value on all the criteria are comparable, which is necessary for combining the scales into one overall scale. MCDA solves the problem of comparing benefits and risks by providing a common unit of value so that the added value of favourable effects can be compared to the loss of value from the unfavourable effects. In practice, the two-step process of scoring and weighting means that all numerical assessments can be carried out by applying simple paired comparison techniques in which just two entities are compared at a time. Hierarchical value trees are often used to structure the benefit and risk criteria, as in Figure 7, while value functions, as in Figure 8, assessed by the relevant decision makers, show how measurable outcomes can be translated into relative values.

Figure 7. Value tree of the risk and benefit criteria for comparing an antipsychotic drug with a comparator and a placebo



Source: Walker S, Phillips L, Cone M. Benefit-Risk Assessment Model for Medicines: Developing a Structured Approach to Decision Making. Epsom, Surrey: CMR International Institute for Regulatory Science; 2006.

Figure 8. A value function showing relative preference values as a function of QTc prolongation



Source: As for Figure 7.

Well before 2009, pharmaceutical companies had recognised the value of MCDA, particularly for constructing portfolios of projects under development [44]. A recent example showing how MCDA enables benefit-risk comparisons to be made is provided by Felli [45]. Other researchers have provided alternative multi-criteria analysis approaches, for example the model proposed by Chuang-Stein [46] have essentially re-invented MCDA without, apparently, knowing of its existence. They use constructed scales to quantify the intensity of the favourable effects and the unfavourable effects, and then they propose using a 'proportionality constant' which effectively equates the units of favourable with unfavourable effects. They invoke statistical significance to handle uncertainty.

A 1993 paper showed the relevance of decision theory, including MCDA, to pharmaceutical medicine [47], and noted that "one of the most vexing problems facing all who are concerned with pharmaceutical medicine is the balancing of risks against benefits." The paper reported the results of two workshops at the London School of Economics, one in 1989 sponsored by Ciba-Geigy as part of their RAD-AR initiative, and the other in 1990 by the Centre for Medicines Research, which showed how MCDA can resolve decisions at intervention points for marketed pharmaceuticals. Nothing more was heard until two more seminars sponsored by CMR in 2004 and 2005 resulted in CMR publications reporting how MCDA could be applied to regulatory decision making [48-49]. Filip Mussen provided an introduction to MCDA for assessing benefit-risk [50], and this was followed by a more detailed exposition in 2009 with the publication of the book by Mussen, Walker and Salek [17].

Our view

Based on an extension of the axioms of decision theory, MCDA is a logical, coherent model for decisions with multiple objectives. As mentioned in the section on qualitative approaches, MCDA is described in Figure 6.1 of the Multi-Criteria Manual as an eight-step procedure on which the other qualitative frameworks have drawn. It is comprehensive in its ability to accommodate all forms of data and time preferences, and provides a way of transforming input data into values (or utilities). As one of several decision-theoretic

models, it can handle uncertainty, though the recent expositions in the medical literature have ignored this aspect of MCDA, so are limited in application. In 1971 Edwards introduced an easy-to-implement version of MCDA called 'SMART' [51], which employs a minimal number of criteria, simple scoring and swing weighting. Its current version is described in Chapter 3 of Goodwin and Wright [52].

The four seminars referred to above all took place with a mixture of pharmaceutical and regulatory specialists, who agreed that the modelling was understandable, practical and communicable. Five current or former regulators have urged further exploration of MCDA to aid benefit-risk decision making [53]. MCDA is a strong candidate for testing in a live regulatory setting.

3.2.8. Principle of threes

Either of both benefits and risks are described by three criteria using just three states for each, high = 3, medium = 2 and low =1 (and for some cases, no effect = 0). For risks, generic criteria of seriousness, duration and incidence are suggested, while for benefits, seriousness, chronicity (e.g., acute, chronic, or duration of disease) and extent of control or cure. Simple totals within benefits and within risks are calculated without weighting the criteria. This approach was first mention by Edwards, et al [54], and was then picked up in CIOMS IV [6].

Our view

This is a particular realisation of a 'performance matrix' or 'consequence table' (see section 4.3.2 of *Multi-Criteria Analysis: A Manual* [8]). It is a two-dimensional summary of the performance of options (rows) against criteria (columns), with each cell entry describing in words, symbols or numbers the performance of an option on a criterion. This approach is universal in magazines that rate consumer products. For regulatory decision making, it can provide useful displays, but the Principle of Threes is too simplistic for even moderately complex cases, and unweighted totals fail to recognise that some benefit or risk criteria are relatively more important than others.

3.2.9. QALYs/DALYs

Favoured by health economists and health technology assessment (HTA) organisations, these approaches focus on outcome dimensions of health [55]. The QALY considers quality adjusted life years, while the DALY is about disability adjusted life years. Each is a kind of multi-criteria model insofar as health states are defined for each dimension and rating scores given, higher scores for more desirable states. Also, weights are assigned to reflect the relative importance of the health dimensions, enabling a weighted average to be taken for any particular combination of health states across the dimensions.

An excellent short introduction to QALYs is provided by Weinstein, Torrance and McGuire [56], who explain that QALYs measure preference or desirability, and were developed to inform cost-effectiveness decisions about health care, define health outcomes in terms of multiple, weighted criteria about health states, and are applied to population-level health, value-weighted over time. The entire supplement to *Value in Health*, volume 12, 2009, is devoted to discussing the value of QALYs [57], so is a good reference to the extensive debates about these measures. Many alternative measures of health outcomes have been developed and compared to QALYs [58-61], and it is now apparent that different measures suit different purposes.

Our view

There is no doubt that QALYs and other health-outcome measures have improved decision making about resource allocation in many countries, though there is also disagreement about potential biases and inadequacies in the various measures that are being used. It remains unclear how useful these measures would be for regulatory decision making, which must be concerned with outcomes of safety and efficacy as well as health consequences. Also, QALY models, with some exceptions [62], are not differentiated by therapeutic area; regulators need to include considerations that are unique to disease states, so regulators would find QALY models insufficiently comprehensive.

Still, as all QALY models are versions of multi-criteria models, it is in principle possible to use any mix of criteria in these models, from measures of efficacy on endpoints through to health consequences over time. For the most part, these models are weak in acknowledging uncertainty about the outcomes and consequences, as they rely heavily on point estimates. Still, there are good reasons to believe that further developments in these multi-criteria models will occur to ensure that the models provide genuine assistance to the decision makers in specific domains. Indeed, it may well be possible for regulators to include QALY-type modelling in their models, which would make it possible to deliver the regulator's model to an HTA so they would then only have to add the cost criterion to the model to determine cost-effectiveness.

3.2.10. TURBO – transparent uniform risk-benefit overview

A drug is scored on one primary and one ancillary benefit criterion, and on the most serious adverse effect and an additional risk, using 5-point scales. Relative weights are assigned to each pair of scores, and the weighted average of the two benefit scores and of the two risk scores are computed. These two scores determine a single point in a two-dimensional benefit versus risk plot, in which acceptable, unacceptable and indeterminate combinations are indicated. This approach is explained in Appendix F of CIOMS IV [6]. This is, of course, a simple multi-criteria analysis model, but lacking the theoretical foundations of MCDA.

Our view

In complex multi-criteria cases involving many criteria, a typical finding is that the same results could be obtained with fewer criteria. Unfortunately, it isn't possible to determine at the start of modelling which criteria are the ones that are mainly responsible for the results. For regulators, focusing on only two benefits and two criteria from the start will not, in the majority of cases, be sufficient, and certainly would not satisfy regulatory standards. Yes, TURBO is transparent, but it is also too simple for use in regulatory decision making.

3.3. Statistics

This section covers various suggestions for measures, based on statistical analysis of data, to be taken into account in making benefit-risk comparisons. None of these is as comprehensive as the models discussed above.

3.3.1. Kaplan-Meier estimator

The intention here is to represent a survival function over time: for a particular condition or context, the proportion surviving over time from some initial starting point. These curves enable comparisons to be made between conditions. For example, the number of people surviving after contracting some

disease at discrete times in the future, based on statistical data over the time period. The curve for a treated sample of patients can be compared to the curve for those untreated to gain an indication of the difference. A ratio of the two data points at any given time can be used to express the magnitude of the difference in survival. This approach can also be used for any single measurable quantity of either unfavourable or favourable effects to see the change over time of the effect.

An extension of the Kaplan-Meier curves that accommodates the trade-off between the durations of different health states is provided by the Q-TWiST method: Quality-adjusted Time Without Symptoms of disease and Toxicity of treatment [63-64]. Basically, this is a trade-off between the duration of suffering from the toxicity of treatment against the prolongation of life. The authors indicated that "it is not intended to provide a unique result combining quality and quantity of life". However, it can display utilities and sensitivity plots of utilities of the two duration dimensions that reveal how the trade-off affects survival time.

Our view

Kaplan-Meier curves are one way of displaying the results of a Markov model or decision tree with repeating event nodes at each time period. Any of these might be incorporated in a more comprehensive model. The Q-TWiST extension demonstrates one way of incorporating utilities into the analysis, though it does not extend to a full multi-criteria analysis of all favourable and unfavourable effects, nor does it formally incorporate uncertainty. Thus, the K-M approach is one way that could be useful, when data are available, for showing time sequences of health states. Our view about Markov models applies here as well.

3.3.2. NNT/NNH

NNT, the 'number needed to treat', is the average number of patients that would have to be treated in order for just one of them to receive the expected favourable effect. NNH, the 'number needed to harm', is the average number of patients that would have to be treated in order for just one person to experience a particular unfavourable effect. Both NNT and NNH are calculated as the inverse of the difference in proportions of the effects between the treatment and control groups [65]:

$$NNT(NNH) = \frac{1}{p_t - p_c}$$

The denominator is often referred to as the absolute risk reduction. For a given disease, a smaller value of the NNT (i.e., a big improvement in the probability of a favourable effect—which might mean a reduction in the chance of a negative outcome) is better as it indicates a drug that is effective for more people, while a larger value of the NNH (a small increase in the chance of an undesirable effect) is preferred because the adverse effect caused by the drug is so rare. Thus, a small NNT means fewer people have to be treated to see one favourable effect, while a large NNH shows that only by treating many people will just one person show the unfavourable effect. Of course, that is true only on average, since the proportions are uncertain.

Our view

NNT and NNH might seem practical because of their simplicity but this simplicity is deceiving. Their main problem is that they cannot be combined to determine if benefits outweigh risks because neither statistic takes account of clinical relevance [66]. For example[1], suppose a drug is found to reduce the incidence of death in vCJD sufferers from 100% to

[1] Thanks to Rob Hemmings for this example, and for bringing to our attention other problems about NNT and NNH.

90%. The NNT is 10. Now imagine a drug that reduces pneumonia deaths from 50% to 40%. The NNT is also 10, but it seems unlikely that anyone would consider these two cases of equal value, and that is just for comparing NNT for two different disease states. A comparison of NNT with NNH for the same disease state requires considering the clinical significance of the favourable event with the unfavourable event: if NNT=NNH, that doesn't mean that the benefit-risk ratio is one. A further problem arises in attempting to apply these statistics to outcomes over time—different values of the statistics would be obtained at different time periods, as for example shown in Kaplan-Meier curves. This difficulty led Hildebrandt et al [67] to conclude "there is much room for improvement in the application of the number needed to treat to present results of randomised controlled trials, especially where the outcome is time to an event."

The underlying problem here is that preference judgements based solely on differences in probabilities violate the criterion of logical soundness, and no amount of 'fixing' the statistics can overcome this problem. In decision theory, probabilities multiply by utilities and it is the probability-weighted utilities, i.e., expected utilities, that are compared, not the probabilities themselves. It is expected utilities that express clinical relevance as well as the probability of realising the effects.

More generally, preferences cannot be well informed by proportions, where both numerator and denominator can take on different ranges of values, with the result that the same proportion could result from very different base conditions. Doubling a survival rate from one month to two months is surely not equivalent in preference to a doubling from one year to two years. Thus, relative risk, p_t/p_c, by itself also violates logical soundness. As Fahey et al have shown empirically [68], different measures yielding the same results can lead to different preferences.

Finally, we should add that Holden's suggestions [65] to incorporate relative utility values (RVs) into the NNH calculations, and to apply minimum clinical efficacy (MCE) analysis for comparing therapeutic options, use probability differences in these models, which can lead to failures of logical soundness.

Note that our critique here is aimed only at the usefulness of NNT/NNH for decisions by drug regulators. We have not formulated a view about applications for other purposes, such as communication by physicians to their patients.

3.4. Measurement methods

The section includes methods that have been proposed for measuring favourable and unfavourable effects and their uncertainties. All the methods result in quantitative expressions of benefits and risks, and some make it possible to compare benefits with risks.

3.4.1. Conjoint analysis

Conjoint analysis (CA) is a measurement method that forces respondents to think about trade-offs. Real or hypothetical products described by various attribute levels are compared, and from many comparisons it is possible to calculate overall preferences and preference functions. Typically, a more preferred level on attribute X is combined with a less preferred level on attribute Y, and this combination is compared with a less preferred level on X combined with a more preferred level on Y, where, *a priori*, at least an ordinal sense of preference is assumed by the experimenters. When several attributes are to be compared, mixtures of more and less preferred levels are presented to respondents, who are asked for their overall preferences, either by stating a preference for Treatment

option A (combination of low X and high Y) or Treatment Option B (combination of high X and low Y) over the other, or by rating their strength of preference for one over the other.

As an example of the latter approach, consider the example shown in Table 2, which concerns a hypothetical treatment for gastroesophageal reflux disease (GERD) [69]. Different combinations of the attribute levels are chosen and presented to respondents, like the example in Table 3. In a choice format, a box would be ticked under either Treatment A or Treatment B to indicate preference for one or the other treatment, or, in the strength-of-preference format, a point would be chosen from a five-point rating scale extending from "Strongly Prefer A" to "Strongly Prefer B".

By repeatedly asking for many comparisons between many pairs of different treatments, it is possible to apply statistical multiple regression analysis, which provides 'utility weights' associated with the attribute levels. Each 'weight' in CA could be interpreted from the perspective of MCDA as a combination of a score on a criterion and the weight assigned to the criterion. However, the concept of 'weight' is different in MCDA and CA; a weighted score in MCDA is simply a weight in CA. Also, the concept of 'utility' and 'utility function' is different from the interpretations in decision theory [70]. From a decision theory perspective, the 'utility weights' could be interpreted as values on an interval scale.

The origins of conjoint analysis can be found in psychometrics and measurement theory. The idea of decomposing a person's overall judgement about a multi-dimensional entity into its component parts became feasible with the development of an axiomatically-based theory, which showed how to do it[71]. Hammond applied CA to capture the 'policy' of an individual in making judgements in situations characterised by multiple 'cues'[72], and Green exploited the ability of CA to capture tradeoffs among competing features of products to assist marketers to achieve a better understanding of how people perceive products[73] and how they trade-off product attributes in making choices.

Table 2. Example of attributes, attribute levels and health-state descriptions for GERD treatments

Attribute	Level	Description
Response time	Fast	Immediate
	Medium	Within one hour
	Slow	Within three days
Relief duration	High	24 hours
	Medium	8 hours
	Low	2 hours
Reversibility	Yes	Effect stops immediately after last dose
	No	Effect lasts for several days after last dose
Dose frequency	Once	Once a day
	Twice	Morning and evening
	Three times	Before each meal
Cost	$100	Medication cost is $100 per month
	$50	Medication cost is $50 per month
	$25	Medication cost is $10 per month

Source: F. Reed Johnson [69].

Table 3. A combination of attribute levels for hypothetical treatments A and B

Feature	Treatment A	Treatment B
Response time	Immediate	Within three days
Relief duration	8 hours	24 hours
Reversibility	Yes	No
Dose	Twice a day	Once a day
Cost	$100	$25

Source: F. Reed Johnson [69].

Our view

It would appear that CA could be useful to regulators in judging the trade-off between favourable and unfavourable effects. This raises the question of whose trade-offs would be modelled. Most applications to date have engaged patients in health economics studies to determine the relative values associated with different health states. Studies of hypothetical treatments, as in the example above, have also been carried out on members of the public. So, it is clear that CA could be useful in making explicit patients' perspectives. It is less clear how CA could be used directly with regulators, for the factorial designs used in CA experiments are time-consuming, and even the non-factorial designs are cognitively demanding because of the complexity of the stimuli, which leads to cognitive simplification strategies, such as ignoring attributes judged to be of lesser importance [74].

A possible way to circumvent the cognitive burden problem is to apply the 'self-explicated method' of CA described by Hauser [75] as an approach that makes it possible to "compose preferences by asking respondents questions about the features themselves." This is a two-stage process of, first, directly valuing the levels within an attribute, then judging the relative importance of the attributes themselves. This is, of course, precisely the process used in MCDA, so with this approach, CA and MCDA come together. The advantage for regulators is that the approach can be used quickly for a unique medicinal product to produce the scores and weights used in an MCDA model.

3.4.2. Contingent valuation

This is the favoured approach of cost-benefit analysis, in which all benefits are translated into monetary values through questions that elicit an amount of money an individual is willing to pay for a good[76]. This approach is not reviewed here as monetary valuations are not deemed relevant to the benefit/risk task facing drug regulators.

3.4.3. Stated preferences

These constitute a collection of methods for assessing an individual's utility functions from their preference statements. There is as substantial literature on these methods in health economics, which is beyond the scope of this paper. A review is provided by Torrance [77], who describes rating scale, standard gamble and time trade-off techniques in detail. Any of these might be useful to regulators. Probability assessment is well covered by Morgan and Henrion [78].

4. Discussion

The intention when we began this paper was to discover the main approaches to benefit-risk evaluation that might be useful to regulators. We soon discovered that 'approaches' is a term whose breadth encompasses nearly everything from a framework to a model to a measurement method, and several other things in between. We soon found that there is potential in nearly every approach, save for a few that we have dismissed, mainly because they are overly simple or too-prescriptive versions of more comprehensive and flexible approaches.

We have formulated suggestions for approaches that could be explored in WP3, in full recognition that no regulator of medicinal products is using any quantitative model or fully-structured, explicit process to support their benefit-risk judgements. Our summary views of the relevance of each approach or method are summarised in Table 4, along with an indication of the potential usefulness to regulators, pre- and post-approval. In general, if an approach or method accommodates both the value or utility of favourable and unfavourable effects, and their uncertainty, the greater the relevance to regulatory decisions. In addition, if the approach lends transparency and aids communication, it is judged to be useful.

Table 4. Suggestions for WP3. Usefulness to regulators is indicated as high, medium or low

Approach/method	Relevance to regulators	Usefulness
Qualitative approach	Essential to follow a structured set of steps for any regulatory decision and to develop a quantitative model.	High
Discrete event simulation	Complex models such as Archimedes could be relevant post-approval to understand actual drug usage. Lack of transparency restricts understanding of its results.	Low
Probabilistic simulation	Can illuminate the risk/benefit trade-off when uncertainty is a major feature of a regulatory decision.	Medium
System dynamics	No use of this approach has yet appeared, but it may be relevant to post-approval decisions about drug usage.	Low
Bayesian belief networks (BBNs)	No use yet reported, but may be relevant to modelling the conditional uncertainties about safety and efficacy.	Low
Bayesian statistics	Can integrate evidence and its uncertainty, both pre- and post-approval, with multiple criteria in decision models.	High
Decision trees and influence/relevance diagrams	Many applications in the medical decision making literature demonstrate the relevance of this approach. Can be integrated with BBNs and MCDA.	High
Evidence-based benefit and risk model	A simple multi-criteria model that is too prescriptive and restricted in scope. Can be subsumed under MCDA.	None
Incremental net health benefit	A simple multi-criteria model, restricted in scope, that can be extended to accommodate uncertainty.	Low
Markov processes	Extends a decision tree to include the movement between health states over time. May be most relevant for post-approval decisions.	Medium
Multi-criteria analysis (esp. MCDA)	Multi-criteria decision analysis extends decision theory to accommodate multiple, conflicting objectives. Provides common units of value for both benefits and risks.	High
Principle of threes	Too restricted and simplistic to be relevant to regulators.	None

Approach/method	Relevance to regulators	Usefulness
QALYs/DALYs	Current focus on health outcomes restricts their relevance, but as they are multi-criteria models, they could be developed for both regulators and health technology assessors.	Medium
TURBO	Too restricted; it is a simplistic multi-criteria model.	None
Kaplan-Meier estimator	Relevant to displaying changes in health states over time, these can be used in Markov models or decision trees, and can incorporate the utilities of the health states.	Medium
NNT/NNH	These statistics don't provide a means for balancing benefits against risks, and their focus on probability differences can lead to logically unsound decisions.	None
Conjoint analysis	Focuses on trade-offs between favourable and unfavourable effects, particularly relevant to eliciting patients' preferences but doesn't consider uncertainty.	Medium
Contingent valuation	Converts all effects into monetary values, so not relevant to regulators.	None
Stated preferences	These methods for assessing utilities and probabilities, can be relevant to any of the above models.	Low

5. Conclusions

Our survey of methods and approaches recognises that human judgement plays an important role in regulatory decision making. Research findings in cognitive psychology show that models can assist, though not replace, the complex process of translating data into useable evidence, and combining favourable and unfavourable effects and their uncertainties into an overall judgement about the balance between benefits and risks. Our evaluations of 18 quantitative approaches, guided by criteria of logical soundness, comprehensiveness, acceptability of results, practicality and generativeness, lead us to these conclusions:

1. Any quantitative method or approach requires a qualitative framework within which the model can be effectively developed. Indeed, the qualitative approach may be sufficient by itself for simpler benefit/risk decisions.

2. Only three quantitative approaches are sufficiently comprehensive to enable the benefit-risk balance to be represented numerically (as a difference or a ratio) by incorporating the value or utilities of favourable and unfavourable effects, along with probabilities representing the uncertainties of those effects: Bayesian statistics, decision trees and influence/relevance diagrams, and multi-criteria decision analysis (MCDA).

3. Five other approaches, while more restricted in scope, may well prove useful for particular cases: probabilistic simulation when the focus is on uncertainty of effects; Markov processes and Kaplan-Meier estimators for changes in health states over time; QALYS for modelling multiple health outcomes; and conjoint analysis to explicate trade-offs among effects, especially for eliciting patient preferences.

4. Combinations of approaches will prove useful in situations characterised by more than one of the following issues: the magnitude of favourable effects, the seriousness of unfavourable effects, uncertainty about the effects, transitions in health states and the time spent in each state, and trade-offs between effects.

Sufficient examples and case studies in the literature reinforce our belief that structured processes, both qualitative and quantitative, could further improve the transparency, communicability, auditability, quality and speed of decision making.

6. References

1. Phillips LD, Fasolo B, Zafiropoulos N, Eichler H-G, Ehmann F, Jekerle V, et al. Modelling the risk-benefit impact of H1N1 influenza vaccines submitted.
2. Meehl PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Northvale, NJ: Jason Aronson; 1996.
3. Edwards W. Conservatism in human information processing. In: Kleinmuntz B, editor. Formal representations of human judgment. New York: John Wiley & Sons; 1968.
4. Grove WM, Lloyd M. Meehl's contributiion to clinical versus statistical prediction. Journal of Abnormal Psychology. 2006;115(2):192-4.
5. Regan-Cirincione P. Improving the accuracy of group judgment: A process intervention combining group facilitation, social judgment analysis, and information technology. Organizational Behavior and Human Decision Processes. 1994;58:246-70.
6. CIOMS_IV. Benefit-Risk Balance for Marketed Drugs. Evaluating Safety Signals. Geneva: Council for International Organizations of Medical Sciences; 1999.
7. Guidance Document. Day 80 Critical Assessment Report: OVERVIEW AND LIST OF QUESTIONS. London: European Medicines Agency; 2009.
8. Dodgson J, Spackman M, Pearman A, Phillips LD. Multi-criteria analysis: A manual. London: Department for Communities and Local Government, First published in 2000 by the Department for Environment, Transport and the Regions; 2009.
9. Rosenhead J, Mingers J. Rational analysis for a problematic world revisited. Chichester: John Wiley & Sons; 2001.
10. Hammond JS, Keeney RL, Raiffa H. Smart Choices: A Practical Guide to making Better Decisions. Boston, MA: Harvard Business School Press; 1999.
11. Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, et al. Decision Making in Health and Medicine. Cambridge: Cambridge University Press; 2001.
12. Walker S, McAuslane N, Liberti L, Salek S. Measuring benefit and balancing risk: strategies for the benefit-risk assessment of new medicines in a risk-averse environment. Clinical Pharmacology & Therapeutics. 2009;85(3):241-6.
13. Liberti L, McAuslane N, Walker S. Progress on the development of a benefit/risk framework for evaluating medicines. Regulatory Focus. 2010:1-6.
14. Jenkins J. A United States Regulator's Perspective on Risk-Benefit Considerations. Risk-Benefit Considerations in Drug Regulatory Decision-Making. Shady Grove Conference Center, Rockville, MD: New York Academy of Sciences; 2010.
15. Eichler H-G, Pignatti F, Flamion B, Leufkens H, Breckenridge A. Balancing early market access to new drugs with the need for benefit/risk data: a mounting dilemma. Nature Reviews Drug Discovery. 2008;7(October 2008):818-26.
16. Cross JT, Garrison LP. Challenges and Opportunities for Improving Benefit-Risk Assessment of Pharmaceuticals from an Economic Perspective. London: Office of Health Economics, reprinted in Appendix 2 of Mussen et al; 2008. p. 1-16.
17. Mussen F, Salek S, Walker S. Benefit-Risk Appraisal of Medicines: A Systematic Approach to Decision-Making. Chichester: John Wiley & Sons; 2009.
18. Eddy DM, Schlessinger L. Validation of the Archimedes diabetes model. Diabetes Care. 2003;26(11):3102-10.
19. Brandeau ML. Modeling complex medical decision problems with the Archimedes model. Annals of Internal Medicine. 2005;143(4):303-4.
20. Eddy DM, Schlessinger L. Archimedes: a trial-validated model of diabetes. Diabetes Care. 2003;26(11):3093-101.

21. Krishna R. Model-based benefit-risk assessment: Can Archimedes help? Clinical Pharmacology & Therapeutics. 2009;85(3):239-40.

22. Kahn J. Modeling human drug trials--Without the human. Wired. 2009 November 15, 2009.

23. Lynd LD, O'Brien BJ. Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis. Journal of Clinical Epidemiology. 2004;57:795-803.

24. Winston WL. Simulation Modeling using @RISK: Cengage Learning; 2001.

25. Lynd L, Najafzadeh M, Colley L, Byrne MF, Willan AR, Sculpher MJ, et al. Using the inremental net benefit framework for quantittive benefit-risk analysis in regulatory decision-making--A case study of alosetron in irritable bowel syndrome. Value in Health. 2009.

26. Savage SL. The Flaw of Averages. Hoboken, NJ: John Wiley & Sons; 2009.

27. Nenov IP, Fylstra DH. Interval methods for accelerated global search in the Microsoft Excel Solver. Reliable Computing. 2003;9:143-59.

28. System Dynamics 2010 [cited; Available from: http://en.wikipedia.org/wiki/System_dynamics

29. Koski T, Noble JM. Bayesian Networks: An Introduction. Chichester: John Wiley & Sons; 2009.

30. Savage LJ. The Foundations of Statistical Inference. London: Methuen; 1962.

31. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain Judgements: Eliciting Expert Probabilities. Chichester: John Wiley & Sons; 2006.

32. Raiffa H. Decision Analysis: Introductory Lectures on Choices Under Uncertainty: McGraw Hill; 1968.

33. Ashby D, Smith AFM. Evidence-based medicine as Bayesian decision-making. Statistics in Medicine. 2000;19:3291-305.

34. Goodwin P, Wright G. Decision Analysis for Management Judgment. Chichester: John Wiley & SOns; 2009.

35. Clemen RT, Reilly T. Making Hard Decisions with Decision Tools Suite: Update: Brooks/Cole; 2005.

36. Stonebraker JS. How Bayer makes decisions to develop new drugs. Interfaces. 2002;32(6):77-90.

37. Ramsey FP. Truth and probability. In: Braithwaite RB, editor. The Foundations of Mathematics and other Logical Essays. London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company. 1999 electronic edition; 1926.

38. Beckmann J. Basic aspects of risk-benefit analysis. Seminars in Thrombosis and Hemostasis. 1999;25(1):89-95.

39. Sonnenberg FZ, Beck JR. Markov models in medical decision making: A practical guide. Medical Decision Making. 1993;13(4):322-38.

40. Thompson JP, Noyes K, Dorsey ER, Schwid SR, Holloway RG. Quantitative risk-benefit analysis of natalizumab. Neurology. 2008;71:357-64.

41. Belton V, Stewart TJ. Multiple Criteria Decision Analysis: An Integrated Approach: Springer; 2001.

42. Keeney RL, Raiffa H. Decisions With Multiple Objectives: Preferences and Value Tradeoffs. New York: John Wiley; 1976.

43. Nunnally JC, Bernstein IH. Psychometric Theory, 3rd Ed. New York: McGraw-Hill; 1994, 1978.

44. Phillips LD, Bana e Costa CA. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. Annals of Operations Research. 2007;154(1):51-68.

45. Felli JCN, Rebecca A; Cavazzoni, Patrizia A. A Multiattribute Model for Evaluating the Benefit-Risk Profiles of Treatment Alternatives. Medical Decision Making. 2009;29:104-15.

46. Chuang-Stein C. A new proposal for benefit-less-risk analysis in clinical trials. Controlled Clinical Trials. 1994;15:30-43.

47. Phillips LD. Decision theory and its relevance to pharmaceutical medicine. In: Mann RD, Rawlins MD, Auty RM, editors. Textbook of Pharmaceutical Medicine. Carnforth, Lancashire: Parthenon Publishing Group; 1993. p. 247-55.

48. Walker S, Cone M. Benefit-Risk Assessment: Summary Report of a Model for Benefit-Risk Assessment of Medicines Based on Multi-Criteria Decision Analoysis. Epsom, Surrey: CMR International Institute for Regulatory Science; 2004.

49. Walker S, Phillips L, Cone M. Benefit-Risk Assessment Model for Medicines: Developing a Structured Approach to Decision Making. Epsom, Surrey: CMR International Institute for Regulatory Science; 2006.

50. Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines - part 1: The development of a new model using multi-criteria decision analysis. Pharmacoepidemiology and Drug Safety. 2007;16(S2-S15).

51. Edwards W. Social utilities. Engineering Economist. 1971;Summer Symposium Series, 6:119-29.

52. Goodwin P, Wright G. Decision Analysis for Management Judgment, 4th edition. Chichester: John Wiley; 2009.

53. Abadie E, Alvan G, Breckenridge A, Flamion B, Jefferys D. Commentaries on 'A quantitative approach to benefit-risk assessment of medicines'. Pharmacoepidemiology and Drug Safety. 2007;16:S42-S6, reprinted in Appendix 4, F. Mussen, S. Salek and S. Walker, Benefit-Risk Appraisal of Medicines, Wiley-Blackwell. 2009.

54. Edwards IR, Wilholm B-E, Martinez C. Concepts in risk-benefit assessment: A simple merit analysis of a medicine. Drug Safety. 1996;15(1):1-7.

55. Garrison Jr. LP, Towse A, Bresnahan BW. Assessing a strucured, quantitative health outcomes approach to drug risk-benefit analysis. Health Affairs. 2007;26(3):684-95

56. Weinstein MC, Torrance G, McGuire A. QALYs: The Basics. Value in Health. 2009;12(1):S5-S9.

57. O'Donnell JC, Pham SV, Pashos CL, Miller DW, Smith MD. Health technology assessment: Lessons learned from around the world--An Overview. Value in Health. 2009;12(S2):S1-S5.

58. Airoldi M, Morton A. Adjusting life for quality or disability: Stylistic difference or substantial dispute? Health Economics. 2009;18:1237-47.

59. Hawthorne G, Richardson J, Day NA. A comparison of the assessment of Quality of Life (AQoL) with four other generic utility instruments. Annals of Medicine. 2001;33(5):358-70.

60. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics. 2004;13:873-84.

61. Dolan P, Gudex C, Kind P, Williams A. A Social Tariff for EuroQol: Results from a UK General Population Survey. York: Centre for Health Economics, University of York; 1995.

62. Johnson FR, Hauber AB, Osoba D, Hsu M-A, Coombs J, Copley-Merriman C. Are chemotherapy patients' HRQoL importance weights consistent with linear scoring rules? A stated-choice approach. Quality of Life Research. 2006;15:285-98.

63. Gelber RD, Cole BF, Gelber S, Aron G. Comparing Treatments Using Quality-Adjusted Survival: The Q-Twist Method. The American Statistician. 1995;49(2):161-9.

64. Gelber RD, Gelman RS. A quality-of-life-oriented endpoint for comparing therapies. Biometrics. 1989;45:781-95.

65. Holden WL. Benefit-Risk analysis: A brief review and proposed quantitative approaches. Drug Safety. 2003;26(12):853-62.

66. Fang C-T, Sau W-Y, Chang S-C. From effect size into number needed to treat. The Lancet. 1999;354(9178):597-8.

67. Hildebrandt M, Vervolgyi E, Bender R. Calculation of NNTs in RCTs with time-to-event outocmes: A literature review. BMC Medical Research Methodology. 2010;9(21).

68. Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. British Medical Journal. 1995;311:1056-9.

69. Johnson FR. Measuring conjoint stated preferences for pharmaceuticals: A brief introduction. Research Triangle Park, NC: RTI Health Solutions; 2006.

70. von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton, NJ: Princeton University Press; 1947.

71. Luce DR, Tukey JW. Simultaneous conjoint measurement: A new type of fundamental measurment. Journal of Mathematical Psychology. 1964;1:1-27.

72. Hammond KR, McClelland GH, Mumpower J. Human Judgment and Decision Making: Theories, Methods, and Procedures. New York: Praeger Publishers; 1980.

73. Green P, Rao VR. Conjoint measurement for quantifying judgmental data. Journal of Marketing Research. 1971;8:355-63.

74. Lloyd AJ. Threats to the estimations of benefit: Are preference elicitation methods accurate? Health Economics. 2003;12(5):393-402.

75. Hauser JR, Rao VR. Conjoint analysis, related modeling, and applications. In: Wind YJ, Green PE, editors. Advances in Marketing Research: Progress and Prospects. Norwell, MA: Kluwer Academic Publishers; 2004. p. 141-68.

76. Treasury H. The Green Book: Appraisal and Evaluation in Central Government. London: The Stationery Office; 2003.

77. Torrance GW. Measurement of health state utilities for economic appraisal: A review. Journal of Health Economics. 2002;5(1):1-30.

78. Morgan MG, Henrion M. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge: Cambridge University Press; 1990.