Benefit-risk methodology project: work package 3 report: field tests

Report
(Published version)

31 August 2011
EMA/718294/2011
Human Medicines Development and Evaluation

Benefit-risk methodology project

Work package 3 report: Field tests

Revised version of the adopted report with any confidential information removed

Disclaimer

This report was sponsored by the European Medicines Agency in the context of the Benefit-risk methodology project and the views expressed are those of the authors. An opportunity for public consultation will be given in the future prior to the adoption of a final position. This report is the intellectual property of the European Medicines Agency.

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

Table of Contents

EXECUTIVE SUMMARY

This report summarises the experience of five field tests in European Agencies of quantitative benefit-risk modelling. Participating agencies included AFSSAPS in France, MHRA in the United Kingdom, MEB in the Netherlands, PEI in Germany and AEMPS in Spain. Each agency chose a drug that was currently under review by the CHMP. We engaged between four and six people in each agency in a facilitated one-day workshop to construct, on-the-spot, a benefit-risk model of the drug and its comparators. Participants included clinical assessors and experts who were knowledgeable about the drug being modelled. Questionnaires filled out by participants at the start and end of the workshop consisted of 20 questions that provided a comparison of current processes with how the modelling session could contribute to benefit-risk assessments.

The eight-stage PrOACT-URL framework (Problem formulation, Objectives, Alternatives, Consequences, Trade-Offs, Uncertainties, Risk Attitude and Linked Decisions), adapted for use in drug benefit-risk assessment, provided an informal agenda for each of the five meetings. Using assessment reports at different stages of writing and other data sources, participants were able to construct Effects Trees that showed the organisation of favourable and unfavourable effects. For each drug, an Effects Table provided definitions of the criteria, showed upper and lower limits of scoring scales, the units in which the data for each criterion were expressed and the type of value function. The latter relates the measured input data to the clinical relevance of the scale range, thus enabling the input data to be transformed into preference values.

Data were entered into the computer model, and judgements made about the relative utilities of the effects. These judgements were expressed as 'swing weights' that compared the clinical relevance of the ranges of the measurement scales. These weights acted as scale constants that equated the units of preference across all the preference scales in the model. With all input data converted into preference values, it was possible to examine the benefit-risk balance of each drug. Extensive sensitivity analyses enabled participants to see how that balance changed, if at all, with different judgements about clinical relevance, or with different input data (such as the pessimistic limits of a confidence interval).

Two graphical representations of the final results proved to be the most useful displays. The first was the Added-Value graph, a stacked bar graph that showed the overall added value for the drug, and another for each comparator. Added value was defined as the benefit from all the favourable effects, and the *absence* of severity of all the unfavourable effects. The second graph was the Difference Display, which showed the difference between the drug and a single comparator (e.g., the placebo) for each effect, favourable and unfavourable. This graph showed the weighted difference of the preference values, so it accommodated both the data and clinical judgements about all the effects.

Analysis of the comparison between pre- and post-questionnaires showed, on average, favourable views of the modelling session for all 20 questions. Overall, this report provides ample evidence for the success of the facilitated modelling sessions and it demonstrates the feasibility of quantitative modelling for determining the benefit-risk balance.

Note: At this writing, any detailed information of the five drugs is considered confidential, so no data about these drugs can be reported. A hypothetical rheumatoid arthritis drug, Drug X, will be used here to illustrate the graphical displays that were found useful in the modelling of the five drugs. This report will be revised to give specific details of each of the models when the confidentiality status will be changed.

These five field tests were done in parallel to the ongoing assessment process without affecting it.

Benefit-risk methodology project

Work package 3 report: Field tests

1. Introduction

The main objective of the EMA Benefit-Risk Methodology Project [1] is the development and testing of tools and processes for balancing multiple benefits and risks, which can be used as an aid to informed, science-based regulatory decisions about medicinal products. The project consists of five consecutive work packages. The first work package reported on the current practice of benefit-risk assessment in the centralised procedure for medicinal products in the EU regulatory network [2]. The report of that work package described processes at the six participating agencies, all of which effectively serve the centralised procedure, but in different ways. Interviews with 55 staff of the agencies revealed substantial differences in the meanings of 'benefit' and 'risk', particularly for the latter. Recognising that a science of benefit-risk decision making requires agreed definitions, we recommended that the EMA's assessment reports be structured to distinguish favourable from unfavourable effects of medicines, and to separate out the uncertainties associated with those effects. This was adopted in the autumn of 2009, with the result that the benefit-risk sections of Assessment Reports are more comprehensive and clear, and the justification for the assessment of the benefit-risk balance is made explicit.

The second work package examined the applicability of current tools and processes for assessing the benefit-risk balance [3]. We found that only decision theory can explicitly incorporate the three key ingredients of all benefit-risk assessments: data for the favourable and unfavourable effects, uncertainties about those effects, and clinical judgements about the desirability, severity and relevance of the effects. However, we recognised that support to decision theory modelling might be desirable from five other approaches: probabilistic simulation, Markov processes, Kaplan-Meier estimators, QALYs and conjoint analysis.

Support for our view about combining methodologies came in the summer of 2010 when four post-graduate students from the London School of Economics studying operational research and decision sciences were given the European Public Assessment Reports (EPARs) of 14 drugs that had already completed the centralised assessment procedure and were asked to use any models they found suitable for quantitative modelling of the benefit-risk balance. Their project reports showed that a variety of model combinations were used, but all were based on decision theory, as can be seen in Table 1.

The experiences of the students lent support to the next phase of our project: work package 3, whose purpose was to field test models for drugs that were currently under review by the CHMP. The present report gives our findings from this field research.

Table 1: The four drugs modelled by LSE students and the type of model for each indication.

Product	Indication	Method
Acomplia	Obesity	MCDA
Cimzia	Rheumatoid Arthritis	MCDA + probabilistic simulation
Sutent	Gastrointestinal cancer	Decision Tree + Markov model
Tyverb	Breast cancer	MCDA + probabilistic simulation

2. Method

Five European Agencies participated in the field research, as shown in Table 2. We asked each agency to choose a drug that is currently under scrutiny by the CHMP, and to schedule a time for our visit to suit their schedule and availability of assessors and experts for a one-day meeting. We held a teleconference with some or all of the participants to explain the purpose of our visit and the preparation that would be helpful. We indicated that the session would be conducted as a decision conference [4], a facilitated workshop to construct, on-the-spot, a benefit-risk model of the drug and its comparators. Questions and discussion followed, and we established that it was important for participants to represent a diversity of views about the drug. About a week before each meeting, we sent email reminders to all participants, which included an attachment describing the decision conferencing process.

The facilitators read the relevant assessment reports, or their drafts, prior to each meeting. It was apparent that a multi-criteria decision analysis (MCDA) model might be useful for each of the five cases. There are three key requirements for such models:

1. To enable comparisons of dissimilar favourable and unfavourable effects to be compared (e.g., percentages of responders vs. mean QTc prolongation) each measured quantity is converted to a preference value on a 0 to 100 scale.
2. The conversion for each effect can be accomplished by a linear or non-linear translation, called a 'value function', which is an assessment of the clinical relevance of various levels of the measured quantity.
3. The units for the preference value scales are equated through a process known as 'swing weighting', which requires judgements of the clinical relevance of scale differences. This enables weighted effects to be summed to give an overall benefit-risk balance.

Each meeting started at about 9am and continued with breaks and lunch until about 4 or 5pm. The meetings were attended by between four and six people, always including clinical assessors and experts in the disease state. The day began with introductions and a re-iteration of today's research purpose: to determine whether or not a quantitative model could help assessors in their task of preparing an assessment report. We handed out a questionnaire consisting of 20 questions about the Agency's existing processes for assessing the benefit-risk balance of a drug. Each question was accompanied by a 7-point scale on which the participant could indicate the extent of their disagreement or agreement with the question.

Table 2: The five agencies participating in the field research and the stage of assessment.

Agency	Date of visit	Stage of Assessment
AFSSAPS	29 Oct 10	pre-Day 80
MHRA	22 Nov 10	pre-Day 80
MEB	20 Dec 10	post-Day 80
PEI	20 Jan 11	*CHMP Overview and Request for Supple-mentary Information*
AEMPS	9 Feb 11	*Day 150 Joint Response Assessment Report*

To start the modelling process, the facilitator briefly explained the foundations of decision theory [5], whose axioms of coherent preference lead to the expected utility model (briefly explained here in the Results section). Recognising that people's preferences are not always coherent, particularly in situations of uncertainty [6], the facilitator explained that today we would use decision theory to help construct a coherent view of the drug's benefit-risk balance. Thus, coherent views are not assumed at the start, but should emerge by the end of the day.

2.1. Problem formulation

The PrOACT-URL framework [7] presented in Work Package 2 (and reproduced here in Appendix A) guided the process of developing the model, though it did not act as a rigid agenda—only those stages and questions that the facilitator felt would be helpful to the group were covered. Most of the questions in stage 1, PrOBLEM, were answered: some drugs are chemical entities while others are biological; some diseases are very serious, others less so; some involve potentially fatal side effects while others do not; one has already been approved for adults and an indication for children is being sought; one is an orphan drug; all differ in the extent of unmet medical need, and so forth. For all five groups, the problem was framed as involving both multiple objectives—favourable and unfavourable effects—and their uncertainties.

As part of the context, the group explained where they are now in assessing the drug. For one drug, only an incomplete draft of the Day 80 Assessment Report had been prepared, and for another the joint Day 150 was available. The assessment by the other three groups fell between the Day 80 and Day 150 reports. In all cases, information from the dossier could have been or was accessed.

At this point in the process, the computer, with Hiview[1] software loaded, was projected so all participants could see the model being created at each step.

2.2. Alternatives (options)

Alternatives were usually considered next. In three cases the drug was compared only to a placebo. In addition to the placebo, two comparators were included for one drug, and two possible doses were considered for another. As these alternatives were given in the dossier and assessment reports, this stage of the modelling was completed very quickly.

[1] Hiview was initially developed at the London School of Economics to enable application of decision theory to problems involving multiple objectives and uncertainty. It is now developed by Catalyze Ltd, and available at www.catalyze.co.uk.

2.3. Objectives and criteria

The same could not be said for identifying and defining the criteria in the OBJECTIVES stage. Part of the problem is that a great many effects, favourable and unfavourable, are mentioned in the reports, but not all of them are relevant to the task of balancing benefits and risks. For example, quality of sufficient standard was assumed for all the five cases, and tolerable adverse events were excluded. Thus, the process has to begin by eliminating criteria that are assumed to be satisfied, or can be assumed to have little or no appreciable effect on the benefit-risk balance. Once agreement was reached, the group constructed an 'Effects Tree'. An example, shown in Figure 1, gives an Effects Tree for 'Drug X', a hypothetical rheumatoid arthritis drug to be used in combination with methotrexate.

After a 'short list' of criteria had been agreed, the next steps were to define the criteria and operationalise measurement scales for them, with clearly defined units. This had, of course, already been done, or the dossier would be considered incomplete. However, finding precise definitions and then defining the measurement scale was usually not easy because this information does not appear in just one place; it is scattered throughout the available reports. In addition, the definitions are often assumed by the assessors, who are familiar with their disciplines, and whose understanding of these issues is often implicit. Precise definitions of measurement scales were often missing in the assessment reports, and required web searches to discover their exact meaning.

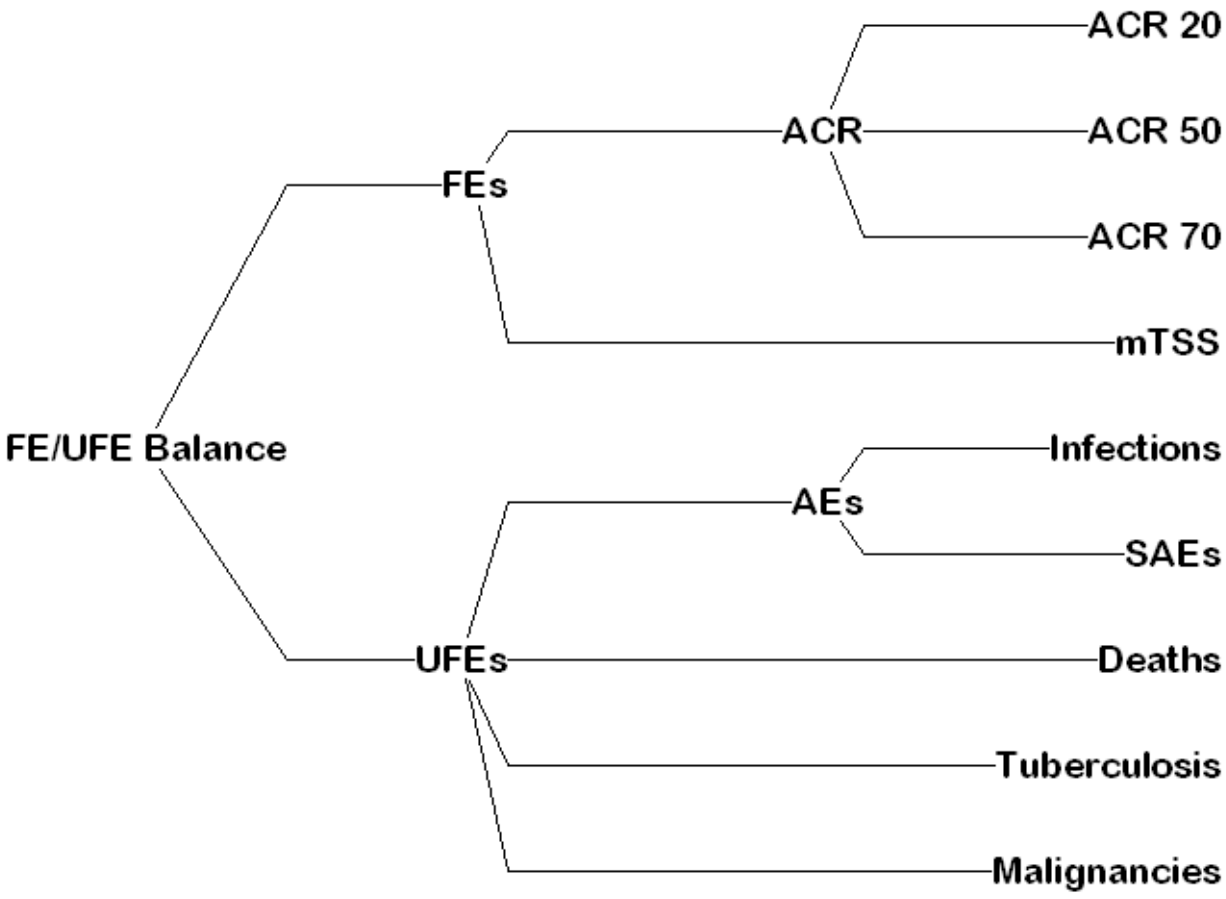

Figure 1: An Effects Tree for Drug X.

Next, the group established two points on each scale that encompassed the range of data, including uncertainties in the data. Thus, for the average response rate of favourable effects, some scales extended from 0 to 100 percent, while for incidence of adverse events that were rare, a scale might extend only from 0 to 20 percent. For scales of numbers of serious adverse events requiring hospitalisation, the range might be very small, say from 0 to 5 SAEs per person per year. These ranges were implicit in the available data, so some effort was needed to make them explicit. The reason for establishing these ranges was to provide meaningful limits for subsequent judgements of trade-offs between the criteria.

2.4. Consequences

Nowhere in any report did we find summaries such as those suggested by the CONSEQUENCES stage of PrOACT-URL, so, again, the group had to engage in considerable searching to provide both the qualitative and quantitative data. Time did not permit the creation of a consequence table, but measurable data for most of the criteria were identified. Occasionally, we suggested a criterion for which informed judgements could be made directly on preference scales, for example, the criterion "Potential for serious adverse events" was added for one of the five drugs even though no SAEs had as yet been observed.

Criterion definitions, measurement units and ranges were entered into the computer program as that information for each alternative was identified, followed by entering measured data obtained from the clinical studies. In preparing reports of each decision conference we summarised all this information in an Effects Table like the one shown here for Drug X, Table 3.

While the data for a criterion were being considered, the facilitator asked questions to determine how those data were considered clinically to ensure that the conversion from measured input data to preference values would capture the way regulators think about clinical relevance. Generally, direct linear scales were found to be appropriate for favourable effects (the higher the effect, the more the preference), and inverse linear scales for unfavourable effects (the more severe the effect, the less the preference). However, there were some exceptions. An example is shown in Figure 2, a value function for Drug X. No malignancies is most preferred, so it scores 100, but the function declines quickly, with 0.5% at a preference value of 50, 1% at 20, 1.5% at 5 and 2 or more at zero.

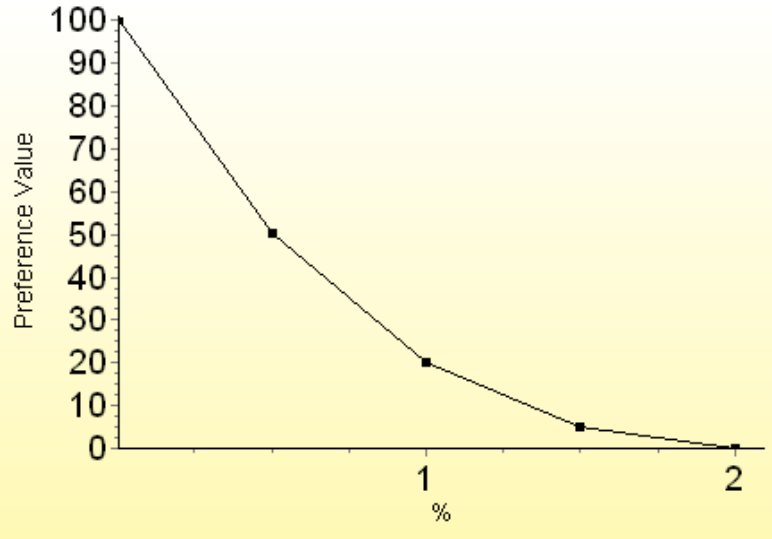

Figure 2: A value function for the proportion of patients developing at least one malignancy.

Table 3: The Effects Table for Drug X.

	Name	Description	Fixed Lower†	Fixed Upper†	Units	Placebo	Drug X 200 mg+MTX	Drug X 400mg+MTX
Favourable Effects	ACR 20	Proportion of patients achieving ACR* 20 at week 24	0	100	%	11.7	58.2	59.6
	ACR 50	Proportion of patients achieving ACR* 50 at week 24	0	100	%	5.8	34.8	36.6
	ACR 70	Proportion of patients achieving ACR* 70 at week 24	0	100	%	2.4	18.8	16.1
	mTSS	Mean amount of progression of joint damage in hands and feet at week 52**	0	10	Change Score±SD	2.8±7.8	0.4±5.7	0.0±4.8
Unfavourable Effects	Infections	Proportion of patients experiencing infections & infestations	70	80	No. per 100 pt-yrs	72.13	79.88	76.62
	SAEs	Proportion of patients experiencing musculoskeletal & connective tissue disorders	25	60	No. per 100 pt-yrs	57.05	28.39	25.88
	Deaths	Proportion of patient deaths	0	3	%	0.15	0.42	0.97
	Tuberculosis	Number of patients contracting tuberculosis	0	30	Number	0	5	28
	Malignancies	Proportion of patients developing at least one malignancy	0	2	%	0.9	1.9	1.4

2.5. Trade-offs

At this stage in the analysis, the model included sufficient information for all input data to be translated into preference value scales that extended from 0 to 100. The next step, swing-weighting, ensured that the units across all the preference scales were equivalent. As an analogy, both Celsius and Fahrenheit temperature scales include 0 to 100 portions. However, the differences in temperature are not equivalent: five units of measurement on a Celsius scale are equivalent to nine units on a Fahrenheit scale. To establish the comparability of preference among all the criteria for a given Effects Tree, we asked this question: "How large is the difference between the 0 and 100 points on this scale, and how much do you care about that difference, compared to the 0-to-100 difference on this other scale." The first part of the questions is a matter of fact, the second of clinical judgement. For example, for the proportion of malignancies, shown in Figure 2, the 0-to-100 difference corresponds to 2% of patients developing at least one malignancy, but how much that difference is considered to be clinically relevant is a matter of judgement. It is both the factual difference and its importance that are compared, criterion-to-criterion.

This swing-weighting process was first carried out within a cluster of criteria, with the largest swing that mattered assigned a weight of 100. The process was then repeated for the remaining clusters. Next, comparisons across clusters compared the 100-weighted criteria from each cluster. For example, the swings on the three ACR criteria were first compared, two at a time (paired-comparisons) to establish the largest swing, which in this case was the ACR 70 criterion, not surprisingly since that would be the best outcome of the three, even though ACR 20 was the primary end point. Swings on the other two ACR criteria were compared, one at a time to the ACR 70 swing, with weights assigned as percentages of that swing, ACR 50 at 70 and ACR 20 at 40. ACR 70's swing of 100 was then compared to the swing on mTSS, which was judged to be equivalent, so was also assigned a 100.

At this point, the three ACR scores summed to 210, but since a degree of double counting was evident in the definitions of the ACR scales, the facilitator asked if the 210 figure should be reduced compared to the 100 on the mtSS score. However, the assessors agreed that the approximate two-to-one ratio between ACR and mTSS appeared to them to be realistic, so the sum was retained. Ideally, the ACR criterion would be sub-divided into non-overlapping segments, but the facilitator decided it was better to accept the definitions of the favourable effects as they were to appear in the assessment report. By adding the swing weight of 100 to the ACR total, the final total for the favourable effects was 310.

The swings on all five unfavourable effects were compared by first identifying the largest swing (Deaths) and assigning it 100, followed by paired comparisons of remaining four criteria. Consistency checks on the sums of the swing weights, which indicate the added degrees of safety if the undesirable effect is absent, enabled the assessors to check the consistency of their assessments, and make changes if necessary to ensure the realism and consistency of their judgements. The final total swing weight sum on the unfavourable effects came to 231. Thus, the ratio of weights on the favourable to unfavourable effects was 310 to 231, which when normalised so the weights sum to 100 becomes 57.3 and 42.7.

Finally, the ACR 70 and Deaths criteria, each of which had been weighted at 100, were compared. The assessors found it difficult to make this judgement, but after discussion agreed they were about equal. This final judgement, along with the other judgements about criterion weights, ensured the equivalence of the units of weighted scores, thereby providing a common scale on which both favourable and unfavourable effects could be displayed.

At this point, sufficient information was available for the computer to make the necessary calculations and display the results. As will be seen in the Results section, various graphical displays showed the overall results and the contribution of each criterion to the overall balance of effects.

2.6. Uncertainty

Uncertainties in the data were explored with sensitivity analyses. Changing the relative weights on the favourable versus unfavourable effects was usually examined first. Changing weights on other criteria followed, and the computer carried out an analysis on all the weights to see which ones mattered most to the benefit-risk balance.

Sensitivity analyses were also carried out with input scores. Ranges of confidence intervals on key effects provided changes to the base-case model enabling exploration of optimistic and pessimistic perspectives about the data. In a few cases, changes in both input scores and weights provided a way for exploring possible future scenarios. For example, it was possible to see how easy or difficult it was to tip the effects balance with different judgements; this gave some indication of the model's robustness. When differences in judgements arose from participants (usually about weights), sensitivity analyses provided a way of determining whether or not the overall balance of effects remained positive or negative. In some cases, the group engaged in 'what-if' analyses to see how the benefit-risk balance might change in light of new information or alternative perspectives about the clinical relevance of some criteria, i.e., scenarios representing possible future developments. These were explored by changing various combinations of revised scores and weights.

2.7. Risk attitude

Many of the judgemental inputs could be considered as expressions of participants' risk attitudes. This could, of course, be different depending on the perspective being taken: regulator, clinician or patient, for example. By assuming different roles, participants for some drugs explored how different perspectives could affect the overall balance. This could involve changing a non-linear value function, or altering weights assigned to criteria. These changes could simulate different levels of risk tolerance as perceived by others.

2.8. Linked decisions

Although this is a separate stage, the group discussed at various times during the construction of the model and in exploring results how consistent their thinking and judgements had been with past decisions, and what the implications of being so explicit might be for future decisions.

By the end of the day, all five groups completed building and exploring the model. We then administered the same 20-item questionnaire, but this time they were asked to "reflect on the modelling session you have just experienced and on how it has impacted the way you can evaluate favourable and unfavourable effects and deal with uncertainty when working on an assessment report". Later, we looked at the changes before and after the workshop to determine its possible added value to participants.

Finally, we wrote a report of about 30 pages for each workshop and circulated it to all the participants as a full record of the session.

3. Results

Most of the results are shown graphically, which enhances exploration of the results and helps to deepen understanding. We begin with a graphical representation of the effects themselves, followed by stacked bar graphs of the benefit-risk balance and difference displays comparing the drug to the placebo. Sensitivity analyses show how varying weights on specific criteria over their full range from 0 to 100 can affect the overall benefit-risk balance, and how combinations of changes in scores and weights can simulate possible future scenarios that could tip the benefit-risk balance.

3.1. Effects Tree

Figure 1, above, gives the Drug X effects tree, four favourable effects and five unfavourable effects. However each of the trees developed for the five drugs we modelled is different; there is no standard tree, nor or there any criteria that are common to all trees, though Infections appears on four of the five trees. Also, serious adverse effects appears as a criterion in only two cases, though many of the unfavourable effects could be classed as SAEs.

Deciding what criteria to include required careful discussion in each group because all Assessment Reports included more criteria than were eventually represented in the Effects Tree. Groups appeared to use an 'Elimination-by-Aspects" (EBA) [8] strategy: criteria were rejected if they failed to satisfy certain minimum requirements for inclusion, such as no data available, effect not clinically important or significantly different statistically from the placebo, effect double-counts one already included, etc.

3.2. Criteria Definitions

Writing a report following each workshop helped us to realise after the early sessions shortcomings in our understanding of the criteria that were included in each model. As a result, we included more information in the Effects Table in the later reports than in the earlier ones. We increasingly appreciated that providing clear and agreed definitions for all the criteria was a non-trivial step.

Assessment reports often failed to provide either complete definitions or references to sources. This problem was particularly evident for scoring systems that were the province of the therapeutic area, and in a couple of cases it applied to the major endpoint. For example, web searches were required to find the exact definition of three favourable effects for two of the drugs. The List of Abbreviations in the assessment report for these three did no more than spell out the acronym and definitions were not given in the assessment reports.

That said, we appreciated that the clinical assessors in each group were well aware of these scoring systems, and their current state of knowledge was quite sufficient to understand the clinical relevance of different levels of the scores and to assess preference scores or value functions. The precision required to express a precise definition remained a part of the assessors' tacit knowledge.

3.3 Added-Value Bars

Results are calculated by applying two simple equations: expected utility and weighted utility. The expected utility rule, which we mentioned earlier is derived from the principles of coherent preference, specifies that the utility and probability should be multiplied for each possible consequence of an action, then summed over all those products to provide a guide to action:

$$\text{Expected utility} = \Sigma\, p \times u$$

As applied to the benefit-risk assessment of a drug, first consider all the effects, favourable and unfavourable. Second, judge the relative value (or utility) of each effect. Third, assess the probability that the effect will occur. Fourth, apply the expected utility rule—multiply effect utility by its probability—for each effect and sum the products separately for the favourable effects and for the unfavourable effects. Finally, if the sum is positive, the favourable effects outweigh the unfavourable effects.

That calculation assumes all effects can be assessed on a common value (or utility) scale. To achieve that commonality for all favourable and unfavourable effects, it is necessary to apply the weighted utility rule:

$$\text{Weighted utility} = \Sigma\, w \times u$$

The form is the same as the expected utility rule, but this time a weight multiplies the utility for each effect to ensure the comparability of the resulting weighted utilities. Since a weighted utility is a utility, it can then be multiplied by the probability associated with the effect. In this way, both types of equation are used to effect the benefit-risk balance.

All five of the models reported here used both equations. In situations of no uncertainty, the weighted utility rule is the only calculation required; it forms the basis for multi-criteria decision analysis (MCDA). When uncertainty is involved and the outcomes of decisions can all be measured in a common unit (money, for financial decisions; QALYS for health outcomes), then only the expected utility calculation is required. But for drugs, balancing benefits with risks requires both equations, so these five models are a mixture of MCDA and expected utility.

The first graphical result we looked at in each decision conference was a stacked bar graph, like the example in Figure 3. This shows the overall weighted preference values for Drug X at 200mg and 400mg plus methotrexate, and the placebo.

These graphs show the added value associated with the presence of favourable effects and the *absence* of unfavourable effects. That is, longer green bars indicate more benefit, and longer red bars show more safety.

If a drug scored at the upper limit of the data range on all the favourable effects (therefore achieving preference values of 100 on all the favourable criteria) and exhibited no unfavourable effects at all (thus obtaining preference values of 100 on all the unfavourable effects), its overall score, however the criteria were weighted, would be 100. A similar line of reasoning for a 'diabolical' drug, one with no favourable effects and displaying the upper limit of the range for all unfavourable effects, would show an overall score of zero. With these extremes of 100 and zero defined, the drugs and placebos in our studies must score somewhere between these limits, depending on their added value from favourable effects and from their safety.

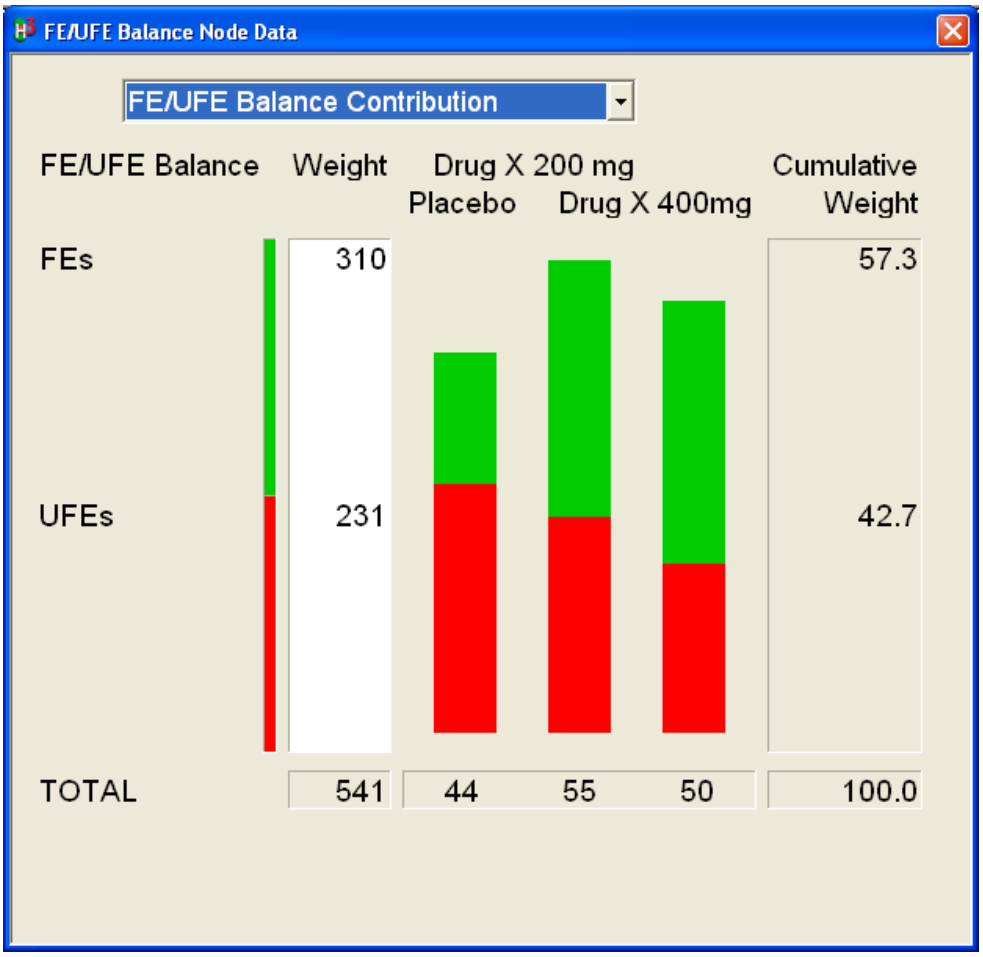

Figure 3: Added-value bar graphs for the favourable and unfavourable effects of Drug X 200mg+MTX, Drug X 400mg + MTX, and for the placebo. Longer green bars indicate more benefit, longer red bars indicate more safety.

In Figure 3, both doses show more added values for the favourable effects, with the 200mg dose better than the 400mg and both better than the placebo. Note that the beneficial effects for the 200mg dose are only slightly less than those for the 400mg dose, but the safety of the lower dose is better.

A more detailed added-value bar graph for Drug X is shown in Figure 4. The favourable and unfavourable effects components are shown for each criterion. It is just possible to compare the segments with each other. For example, fewer deaths and tuberculosis cases are evident for the lower dose than the higher dose.

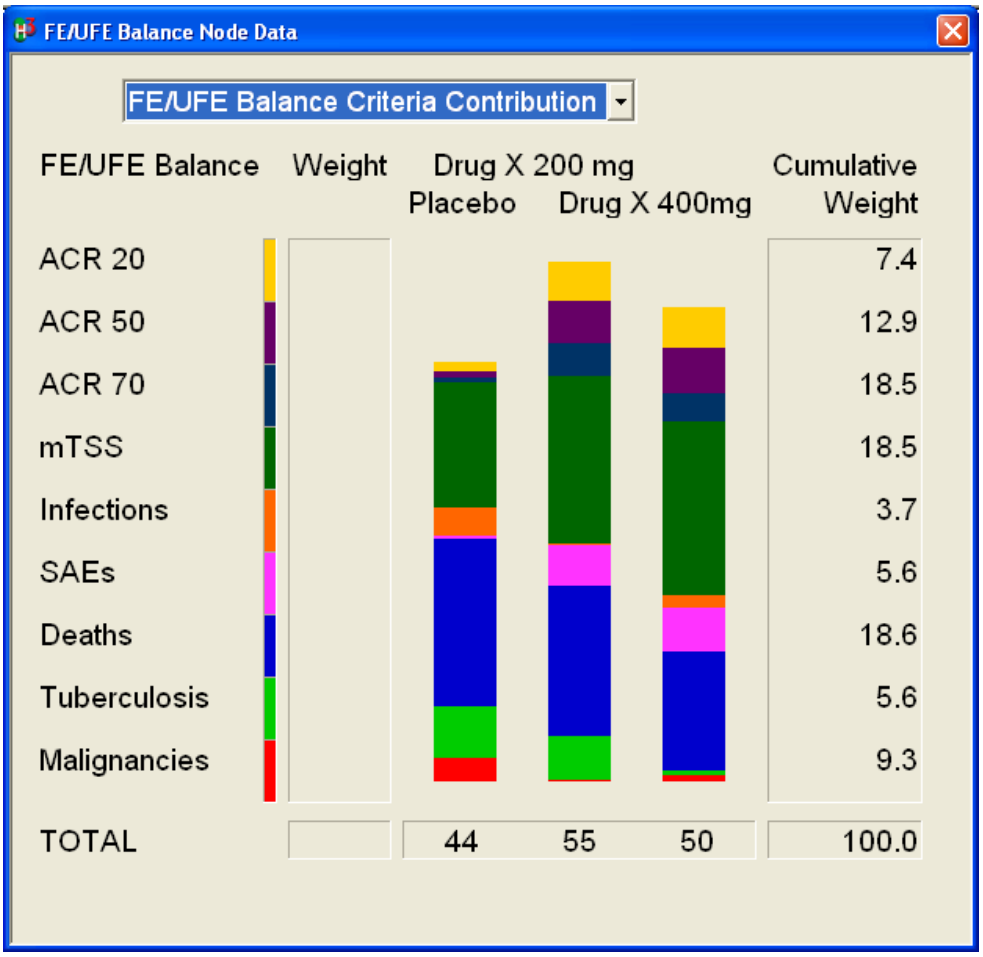

Figure 4: Added-value bar graphs for all effects of Drug X 200mg+MTX, Drug X 400mg + MTX, and for the placebo.

3.3. Difference Display

While the added-value bar graphs give a good overall comparison of the options, it is easier to see the difference in Figure 5. The criteria have been re-ordered on the basis of the weighted differences between the preference scores on the 400mg dose and the placebo.

The first column of figures shows the cumulative weights, which are the products of the weights along each branch of the value tree, normalised so the products sum to one. The next column shows the difference in the original preference values assigned to the options on each criterion. The third column shows each difference multiplied by that criterion's cumulative weight. It is this weighted difference that is based on the data, its clinical relevance expressed as a preference value, and the relative clinical importance of the criterion, as judged in the swing weights. *In short the weighted differences show the clinical relevance of the evidence-based difference between the drug and the placebo.*

The right-extending green bars show the relative magnitude of those weighted differences that favour the drug, while the left-extending red bars favour the placebo. The cumulative sum of the green bars minus the lengths of the red bars comes to 6.0, the difference in overall scores shown in Figures 3 and 4, 50 minus 44. Curiously, one of the unfavourable effects, SAEs, is better for the drug than the placebo.

Figure 6 shows the differences between the two doses. Clearly, the main advantages of the lower dose are Tuberculosis and Deaths.

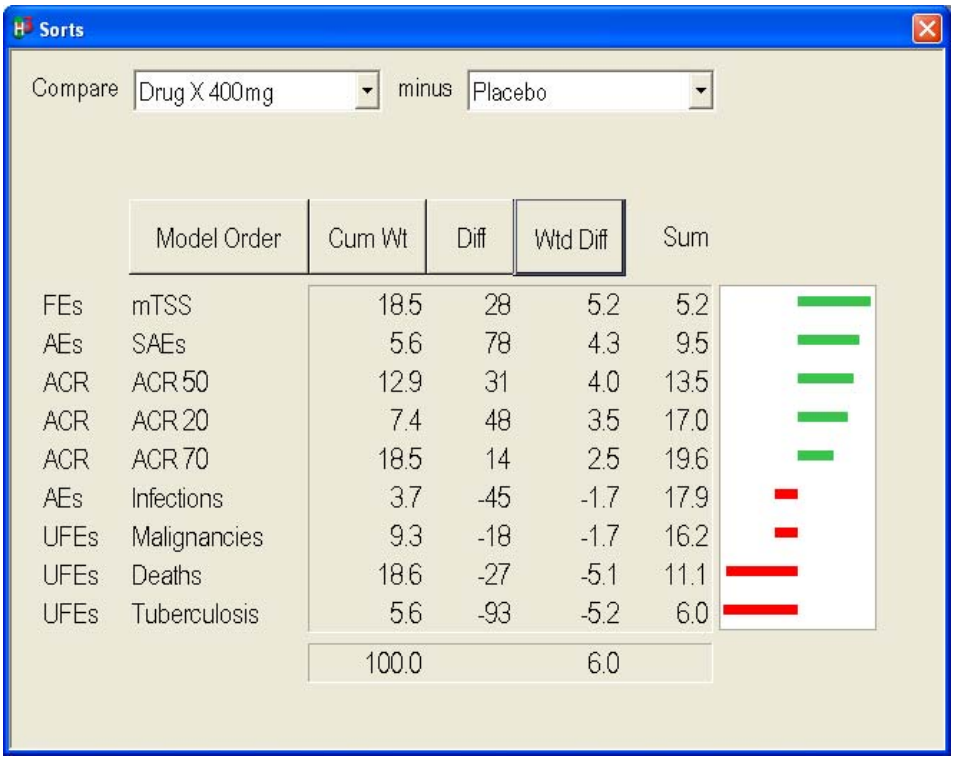

Figure 5: Difference display comparing Drug X 400mg + MTX with the placebo.

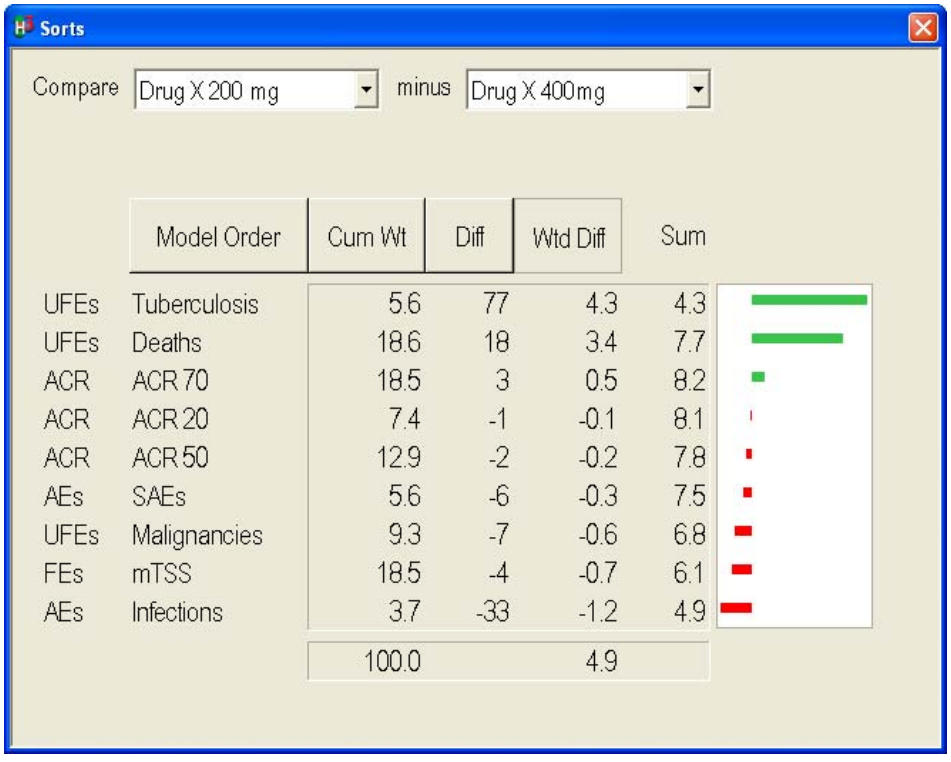

Figure 6: Difference display comparing Drug X 400mg + MTX with Drug X 200mg + MTX.

3.4. Sensitivity Analyses

After seeing initial results, the groups examined the trade-off between favourable and unfavourable effects. Recall from Section 2.5, that the normalised weight on the unfavourable effects was 42.7, which is shown in the Cumulative Weight column of Figure 3, reproduced below in the left graph, and is also the position of the vertical red line in the right graph of Figure 7. The red line intersects the three lines above at the total scores of 44, 55 and 50 for the options, as displayed in the left graph. Those cumulative weights appear in the right column of the Added Value Bars in Figure 7. Increasing the weight on the Unfavourable Effects node would move the red line to the right, with the overall preference value for the 200mg drug increasing slightly but the overall value for the placebo increasing rather more, so that the placebo overtakes the drug at a cumulative weight of about 74. Decreasing the weight on the Unfavourable Effects results in the 400mg becoming most preferred (but only very slightly) when the cumulative weight reaches about 7.

Thus, over a substantial range of trade-off judgements between the favourable and unfavourable effects, the 200mg dose remains most preferred. The model is quite robust to differences of opinion about the relative importance of these two classes of effects.

Figure 7: Left, the Added-Value Bars for Drug X, as shown in Figure 3. Right, a sensitivity analysis showing the effects on the bars of changing the cumulative weight on the Unfavourable Effects node. The transition from yellow to the green background colour shows the point at which the drug would no longer be most preferred.

Next, we usually showed sensitivity analysis graphs for weights on specific criteria, particularly those that were judged to be most important. Sometimes the graph showed that very large changes in weights would be required before the results would change in favour of a different option, but at other times only a small change would be required.

Hiview includes a facility for carrying out sensitivity analyses all at once on the criteria. We always used this facility to gain an overall impression of the robustness of the model. Figure 8 shows the

analysis for Drug X. One participant was concerned about tuberculosis, for the cases had occurred mainly in high-incidence countries, so he felt that the weight on this criterion should be reduced. As can be seen in Figure 8, the weight would have to be increased by 5 to 15 points, but as the cumulative weight on Tuberculosis is already just 5.6, decreasing it by 5 points would effectively eliminate this criterion altogether, which was not deemed realistic by the other assessors.

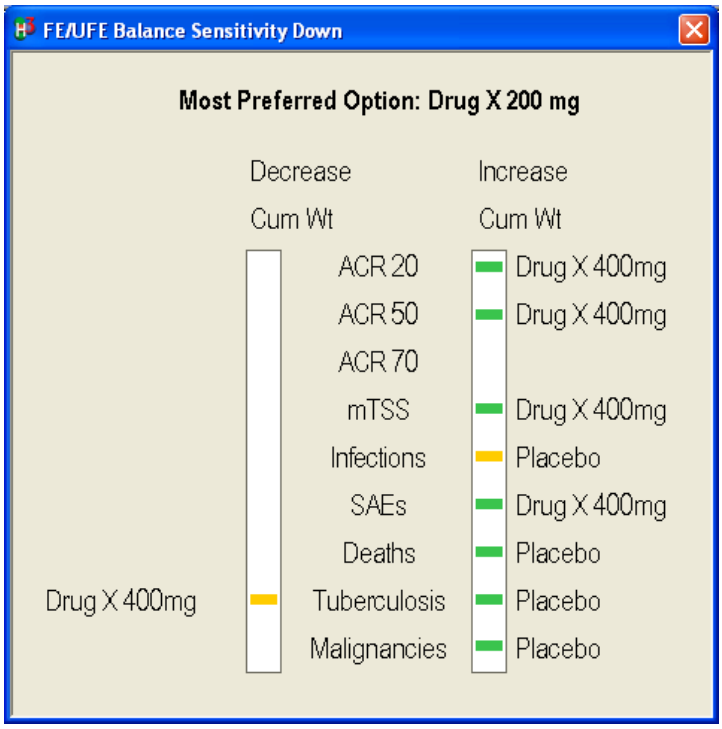

Figure 8: Sensitivity analyses on the cumulative weights separately for each of the effects for Drug X. The coloured bars indicate by how much the cumulative weight must change for a different option to become most preferred: green—more than 15 points, yellow—between 5 and 15 points, red—less than 5 points. With no red bars, and only two yellow ones, substantial changes in weights would be required to change the overall most preferred option from the 200mg dose.

3.5. Scenario Analyses

Although sensitivity analyses are one way of exploring the effects of uncertainty in the input data about weights on the overall benefit-risk balance, it is also possible to deal with uncertainty about the future by engaging in scenario analysis. For example, one drug posed a new challenge: its sole clinical study was based on very few patients because it was an orphan drug. The substantial lead in the drug's overall score over the placebo, the largest difference of the five drugs studied, might not be so clear-cut if several of the mean scores for the favourable effects were overestimates, and if the percentage of patients experiencing infections were underestimated. So, we constructed a pessimistic future scenario by inputting to the computer all the lower limits of the 95% confidence intervals for the mean scores on the favourable effects, and the upper limit of percentages of infections, the only unfavourable effect of particular clinical concern. The consequence of these changes was to reduce the substantial lead to 5 points, though the model remained fairly robust to changes in weights. The only exception was that less than a five-point change in the cumulative weight for infections could tip the balance in favour of the placebo.

Drug X provides another example of scenario analysis based on the input data. Note from Table 3 the range of plus or minus one standard deviation on the mTSS data. One standard deviation from the

mean in the optimistic direction for the placebo and in the pessimistic direction for the drug gives a clear overall result favouring the placebo over either drug dose. Clearly, the uncertainty in this one end point matters to the benefit-risk balance and would be worth further exploration, perhaps using probabilistic simulation.

For another of the five drugs we studied, the consequences of two changes were explored by the group. First, the weight on Potential for SAEs was dropped to zero because none had been observed in the clinical studies. This substantially increased the preference values for the higher dose of the drug over the placebo, which demonstrated the potential effectiveness of the drug.

These analyses showed the ease with which other scores and weights can be used to explore the effects of uncertainty about the future performance of the drug.

3.6. Questionnaire Results

Nineteen participants completed both the pre- and post-questionnaires. The median number of positive differences for each participant at the end of the workshop was 14 (out of 20 questions). The average change scores were positive for all 20 questions and the overall average change was 1.12 (on a 7-point scale from strongly disagree to strongly agree).

The seven highest differences, and their average change scores in parentheses, were that the modelling process:

1. can easily test different perspectives for their impact on the results (2.95),
2. helps us explore how the overall balance is affected by a reduction or increase in uncertainty (2.47),
3. helps me to see the impact of uncertainties on the benefit-risk balance (1.84),
4. has an overt and clear structure (1.42),
5. helps us combine data about value and uncertainty into an overall balance between favourable and unfavourable events (1.42),
6. helps us make our assumptions, multiple objectives and trade-offs explicit (1.37), and
7. helps us anticipate the outcome of possible issues of concern (e.g. less efficacy or more severe side effects) (1.21).

4. Discussion

The positive change scores show that all the participants who filled out both questionnaires found much that is useful in the modelling process. It is important to recognise the three elements that contributed to the favourable experiences: impartial facilitation, a structured quantitative model developed on-the-spot, and a group process that enabled everyone to contribute in a context of peer review. This is very different from the back-room modelling that typically attends statistical analyses: data turned over to an individual statistician, who develops a model that expresses results in summary form, which are then given in assessment reports as measures of central tendency, confidence intervals and significance levels. The goal of that type of modelling is to develop as far as possible objective statements in the form of valid statistical inferences about effects.

The goal in decision modelling is to extend the statistical analysis to include the effects of uncertainty and value judgements on the overall benefit-risk effectiveness of drugs. Favourable and unfavourable effects are based on facts; benefits and risks are based in part on facts, but also on considerations of uncertainties about the facts, and on value judgements, including clinical relevance, of the facts.

A key feature of decision modelling is its ability to distinguish facts from value judgements, and to combine both features into an overall assessment. Because value judgements are necessarily subjective, it is important to make them explicit and subject to discussion, debate and peer review. Even if agreement cannot be reached, a quantitative model can be used to explore whether or not disagreements matter to the final result. If they do, then further information or data might be required, or further exploration of the reasons for the disagreement surfaced. Always, whose values are to be considered is a key question: industry, regulator, prescriber or patient. Decision theory provides no answer other than a reminder that it is the decision maker whose values matter. Since decisions about drugs are made by all four of these decision makers, it is likely that they do not wholly share the same value judgements. Our focus is on regulators who are acting on behalf of an authority; presumably, then, it is the authority's values that are being brought to bear on the benefit-risk balance.

4.1. Process features

The process described in this report helps to create a context that enables useful and constructive interaction between key players in the decision making process. We see this as the result of eight features:

First, the *PrOACT-URL framework* provides a process for developing amongst a group of key players a shared understanding of the issues, a sense of common purpose, and a commitment to the best way forward.

Second, *making the structure of a benefit-risk problem explicit*, in the form of agreed options, an effects tree of objectives and criteria, and criteria definitions, ensures that everyone understands the meaning of the key elements of a benefit-risk assessment. The effects tree is a structure that provides information by the way the criteria are clustered under higher-level objectives. Criteria within a cluster are more similar than those between clusters, usually because all the criteria within a cluster are different ways of realising the higher-level objective represented by the node. Just seeing the overall organisation provides information to decision makers even before data are input.

Third, *clustering criteria* assists in the assessment of weights. It is easier to compare similar criteria to each other first, then to take the most heavily-weighted criteria, one from each cluster, and compare them to each other. The structure also facilitates paired-comparisons rather than having to think of many criteria at the same time.

Fourth, clustering makes it easier to carry out *consistency checks*. Swing weights represent added value, so sums of swing weights should be meaningful to participants. For example, if criteria A, B and C are assigned swing weights of 100, 70 and 30, then the added value from least to most preferred positions on scales B and C should together equal the added vale on criterion A. We frequently asked these questions, which often resulted in scaling the non-100 weights downward. Without those consistency checks, there is a tendency to think everything is important, and thus to over-weight the criteria.

Fifth, the goal of *realism and consistency* in scoring, developing value functions and assessing weights, pervades the process. The facilitator's questions to the group are often in aid of this goal, whether it is seeking agreed definitions about effects, or making consistency checks in weighting.

Sixth, using *computer technology* to reassemble the individual constituent parts into an overall result, and to explore differences of opinion and the effects of uncertainty, with instant playback of results in easily interpretable graphic displays, extends the capability of assessors to construct an informed judgement about the benefit-risk balance.

Seventh, the *facilitator's impartiality* provides a climate in which information is seen as a neutral commodity so that values can be safely explored [9]. In particular, the facilitator guides the model-building, attends to the group process and maintains a task orientation for the group, but does not contribute to the content of the discussion, thereby maintaining his or her impartiality.

Eighth, the *decision conference format*, in which the model is constructed on-the-spot using specialised software that utilises a variety of graphical displays of results, immerses participants in every phase of the work. The result is that they gain a deep understanding and ownership of the results, which can be more nuanced even for simple models than any one individual's understanding of the benefit-risk issues.

4.2. Observations

In this section we report several observations that are indirectly relevant to benefit-risk assessment of drugs. However, the observations are limited by the potential unrepresentativeness of the five drugs we modelled, so we can at best only propose some working hypotheses.

For some of the five drugs, participants reflected on the consistency of their judgements with previous cases. When value functions were elicited, participants sometimes wondered whether the same value function would apply in the future for new drugs. We do not know the answer to this question, but it should emerge as drug-specific value functions are created in the future.

Do we think that generic models could emerge to be plucked off the shelf for a new drug? Surely there is no single generic model, for the five models reported here have little in common, other than the infections criterion. But we anticipate that models for a particular disease state may emerge as being so similar that a single model, consisting mostly of common criteria, but with provision for new criteria, could be helpful.

It is interesting to see that the robustness of a model depends mainly on weights for unfavourable effects, but not for favourable effects. This is not surprising, as clinical studies are powered for the favourable effects, so the major uncertainties are in the unfavourable effects. At this stage, the dependence of robustness exclusively on unfavourable effects is no more than a working hypothesis, to be confirmed, or not, in subsequent research.

Another working hypothesis is that the reason participants liked the Difference Display better than the Added-Value graph is that the former provides a better match to the mental model of assessors. Several assessors whom we interviewed in the first year of the Benefit-Risk Project explained that their

mental model starts with an expectation about a new drug that is determined by the therapeutic area, higher for areas in which available drugs are already successful, but lower for drugs dealing with very difficult conditions. That starting point serves as an initial anchor, which is then adjusted as assessors compare the drug with the placebo one effect at a time, first building up a reasonably positive picture from the primary and secondary endpoints, then feeling the balance become less good as the unfavourable effects are considered. This 'anchoring-and-adjustment' strategy has been noted in several psychological studies, often with the finding that people's adjustments as more data become available are insufficient [10-11].

We realised only at the fourth session the importance of exploring scenarios in which the impact of possible future events is explored. Past experience suggests that combinations of changes involving input scores, value functions and criterion weights can simulate possible future developments, such as unanticipated effects and their consequences, or the effects of a risk management plan. Thus, the model does not have to just represent a drug as it is now for the intended population; it can also be used to anticipate possible futures.

These five cases only required application of decision theory, guided by the PrOACT-URL framework, to deal with the uncertainty and utilities of favourable and unfavourable effects. They did not require application of any of the other methodologies, probabilistic simulation, Markov processes, Kaplan-Meier estimators, QALYS or conjoint measurement, which we suggested in our Work Package 2 report might be useful. We still believe there is scope for these, but none of the five live cases we have looked at so far would have benefitted from these additional approaches.

One exception is probabilistic simulation, for which we see an important long-term role. In each of the five cases, statistical summaries of central tendency formed the inputs for some of the criteria, and we used limits of confidence intervals as inputs in sensitivity analyses to test the robustness of model results. We believe that could become routine, but confidence intervals are only statements about standard errors of the statistic in question; they do not encompass the entire range of data in clinical studies. By treating the difference between favourable and unfavourable effects as an uncertain quantity, probabilistic simulation could show the distribution of people in the population over that uncertain quantity. That distribution would provide the simple statistic of the proportion of people for whom the benefits of the favourable effects exceed the severity of the unfavourable effects.

We commented earlier on the many specialised scoring systems that have been developed for particular disease states. These systems typically consist of multiple criteria, sometimes combined in complex ways to provide a single score for a patient. These are, in effect, multi-criteria systems, developed without necessarily applying the formalities of multi-criteria decision analysis, as formulated in 1976 by Keeney and Raiffa [12], with the result that some are flawed. The same could be said of QALYs, which were mainly developed by health economists seeking a common unit of benefit so that cost-effectiveness could be compared across different disease states. The models developed in our five field tests are entirely consistent internally, in the sense that the weighting process ensures the equality of preference values across all the effects, but the scales are not comparable across the five models. It is our belief that regulators will need to continue using scoring systems developed by experts in disease states, but that these will continue to be incompatible with the QALY-based approach of health economists.

However, we see a possible bridge between regulators and HTAs. A quantitative model developed by a regulator could be passed on to an HTA, who could add cost and QALY criteria. Once a trade-off is judged between a unit of benefit in the regulator's model and a QALY, all the benefits in the regulator's model can be expressed in units of QALYs and then the cost effectiveness calculation can proceed. In practice, this will be more complex than we suggest, but at least the outline of an improved working relationship between regulators and HTAs is worth exploring.

4.3. Limitations

All decision making, whatever the topic, is characterised by a trade-off between quality and effort. We devote more resources, intellectual, physical and emotional, to decisions whose consequences really matter to us. It is clear that this trade-off is exercised in regulatory decision making: more effort is typically expended in assessing the benefit-risk balance for new, innovative drugs that deal with life-threatening conditions than for 'me-too' drugs for non-life-threatening diseases. Quantitative modelling is certainly not required for every new drug application—any improvement in quality might be so small that it would not justify the effort required to develop a model. Our study has not identified when quantitative modelling would be appropriate; the five cases were in part chosen under the judgement by an Agency's assessors that these were suitable candidates for quantitative modelling, so we feel it would be premature to suggest when quantitative modelling would in general be appropriate. Guidelines for choosing to model or not will only emerge with experience and with follow-through of cases in which modelling was carried out.

Neither does this study of five drugs answer the related question of how much modelling, simple or complex, would be useful to regulators. Again, the answer will emerge with experience, as we indicate in items 1 through 3 of our recommendations in the next section. It is clear to us that the PrOACT-URL framework would be helpful in assessing any new drug without developing a quantitative model.

Criticisms occasionally arise that quantitative modelling is resource intensive. The assumption behind this criticism appears to arise from several previous attempts to apply MCDA to benefit-risk assessment with groups of assessors [13-14]. In all these cases, the assessors either created a case on-the-spot, or were given a write-up of a case study which they had not seen before. For our five models, all participants were well acquainted with the drug in question, and had in some way participated in writing at least a draft of the Day-80 Assessment Report. Thus, they were already clear about the options, and had a good sense of which effects were most important. That enabled us to structure the problem very quickly, and enabled a complete model to be thoroughly explored in just a single day.

Previous experience using decision-analytic modelling with various pharmaceutical companies confirms that with experience, the level of effort required to complete a benefit-risk model approximately halves once participants know what to expect. In advance of a facilitated modelling session the participants can ensure that the options are agreed, that only criteria likely to affect the benefit-risk balance are included and that definitions and measurement scales are identified. Even entering the objective data could be done before the group meets. That leaves the task of briefing the group on work to date and then devoting the meeting to creating value functions, expressing clinical judgements in weighting and conducting sensitivity analyses.

Another limitation of our study is that we did not record the assessors' justifications for their clinical judgements. We have only recorded the value functions and weights in numerical form. Although the Hiview software provides a facility for annotating any inputs with text explanations, we did not have time to do this with our five groups. When this has been done in the past, the facilitator pauses after each set of judgements (e.g., the weights on all criteria below a chosen node) and asks the group to provide the reasons for those judgements. That can add an hour or two to sessions in which 10 to 15 criteria and three to five options are included in the model. This approach could save time subsequently as model results are transferred into text (e.g. for the EPAR). It would also decrease subsequent challenges because the reasoning is now apparent, although in some cases could increase debate as others disagree with the original assessors' opinions. But that at least would open up discussion on important issues that might otherwise have remained unaddressed. Here, again, the quality-effort trade-off has to be judged separately for each case.

A final limitation is inherent in the available software. Decision-analytic software generally falls into two categories: those modelling uncertainty will use decision trees and influence diagrams, with limited ability to represent multiple objectives, while those modelling multiple objectives use value trees with limited capabilities for incorporating uncertainty. Hiview falls into the latter category, and we managed uncertainty in three ways: by using measures of central tendency as inputs for some criteria (so distributions of uncertain quantities have been summarised by measures such as a mean, which accommodates the uncertainty), by carrying out sensitivity analyses on uncertain quantities (e.g., varying a weight over its entire range, or using the upper/lower limits of confidence intervals as input scores) and by exploring scenarios (e.g., combinations of different scores and weights) to anticipate the results of uncertain future developments. We believe that further development of software would make benefit-risk modelling easier and more routine. Indeed, software could be developed for use by assessors who have no knowledge of decision theory or MCDA, though it would be important to design the software in such a way that it can be interrogated easily to see why results were obtained, thus avoiding the 'black box' syndrome.

5. Conclusions

We return, now, to the purpose of the Benefit-Risk Project: the development and testing of tools and processes for balancing multiple benefits and risks, which can be used as an aid to informed, science-based regulatory decisions about medicinal products. The requirement to be 'science-based' suggests that the regulatory decisions should be guided by explicit problem-solving and decision-making processes. We field tested a combination of three processes, the PrOACT-URL framework, decision-analytic modelling and facilitated workshops, and found that their collective application could indeed aid regulatory decision making. Deliverables for Work Package 3 consisted of a report of each field test "summarising the exercise and identifying those processes, tools and organisational structures that were considered to add value to the process of benefit-risk assessment by regulatory authorities". Those reports have been completed and approved.

To identify the features that added value, the pre- and post-session questionnaires included most of the questions we had proposed for Work Package 4, whose purpose is to develop benefit-risk tools and processes. The WP4 deliverables include an operational decision-aid, and a consultative workshop "to explore the acceptability and the potential implementation of the tools and processes". The further development of both could easily take place over the next few months in two stages: develop an operational decision aid by the end of 2011, then hold a public consultation and workshop in the first half of 2012.

The success of the five field tests raises issues for the EMA/CHMP about endorsing the three processes we applied in the facilitated group modelling as a useful adjunct to its current processes. These include:

1. *Adopting the PrOACT-URL framework to guide the process of evaluating the benefit-risk balance.* It could be left to the Rapporteur and Co-Rapporteur to decide how the framework should be applied to any specific drug. In some cases, once the nature of the problem and its context, the options to be assessed and all effects -- favourable and unfavourable -- have been identified, the solution may be so obvious that no further steps are required. If not, then providing a Consequence Table, with the options in columns, the criteria as rows, and quantitative and qualitative outcomes in the cells, may be sufficient to see a solution. If further work is required, then successive stages might benefit from a quantitative model.

2. *Developing a simple quantitative model.* The simplest quantitative model replaces the cells in the Consequence Table with preference values that are linearly related to the measures shown in the cells of each row. Weights are then assigned to the scale ranges in each row. Weighted scores are calculated for each cell, and the weighted scores summed in each column, with the totals showing the overall added value for the options. Simple and effective, anybody can learn it. It is clearly explained in Chapter 6 of *Multi-Criteria Analysis: A Manual* [15] (http://www.communities.gov.uk/documents/corporate/pdf/1132618.pdf).

3. *Developing a more complex model.* If a simple model does not adequately represent the complexity of the problem (e.g., the assumption of linearity between preference values and measured data is violated) a more comprehensive model, like that for Drug X, could be developed with the support of appropriate decision-analytic software.

4. *Applying quantitative modelling as a once-off exercise, or as a continuous process.* Modelling could be applied at any stage of the 210-day process, though not before assessors and other experts have at least gained a degree of familiarity with the dossier. A model could assist regulators in preparing the Day-80 report, or in consolidating the subsequent assessment reports as more information is obtained. It could even be used in post-approval monitoring.

5. *Encouraging applicants to develop quantitative models to clarify the benefit-risk balance*. Companies could be encouraged to submit benefit-risk models in support of their application, as some are already doing. Assessors would then be in a better position to judge the benefit-risk balance because a model makes explicit the many pieces of the benefit-risk puzzle, providing a clear view of how adequately each piece contributes to the overall result. Of course, this also imposes a duty of care on the assessors to understand the model so they can make valid comments on it. Lacking that expertise, as must have been the case in the early days when statistical analyses were not well understood by assessors, special training programmes will be required for assessors to provide them with sufficient information about benefit-risk models. Of course the purpose of statistical modelling is to obtain a view that is as objective as possible. Benefit-risk assessment additionally requires and makes value judgements explicit, and the ideal application would show that the benefit-risk balance is robust to differences in those value judgements. This would enable the CHMP to challenge, if necessary, those value judgements that are the crucial ones in making the final recommendation and not spend time on those judgements that change the benefit-risk balance insufficiently to make a difference to the recommendation.

6. *Using quantitative models during CHMP meetings*. This would provide a very rapid means for the Rapporteur to brief the meeting on the benefit-risk issues, and it could help members resolve differences of opinion about matters that have been modelled. The model could facilitate the process of turning quantitative results back into words that provide a clear justification for the benefit-risk assessment in the EPAR and other documents. Indeed, the development of requisite decision models [16] was motivated by the finding that final decisions did not always align to model results. In fact, it was when they did not align that decision makers found the modelling most helpful. This was because the model provided a framework for clear and rational thinking and a structure for constructive dialogue. As a result the modelling process helped to develop new insights, realigning intuitions and creating a sense of common purpose among participants. In short, it is the process of modelling, not necessarily the result, from which decision makers mainly derive benefit.

7. *Providing training in the PrOACT-URL framework and in decision-analytic modelling.* Two levels of training would be appropriate: an overview and introduction whose purpose is only to acquaint participants with the framework and the models, sufficient to enable them to understand results, and intensive training for those who wish to be able to guide others in using the framework and in doing the modelling.

8. *Using the modelling to facilitate justification of the benefit-risk assessment.* The modelling process of peer review and interacting with the model can deepen insight about the benefit-risk balance, which helps participants to form clear preferences about a drug's effects. This should facilitate the preparation of a clear justification for the benefit-risk assessment, especially in the EPAR. It should be possible to monitor and compare EPARs that have benefitted from the modelling process with those that have not to establish whether or not modelling has improved the communicability of benefit-risk assessment.

It is worth once again stressing that the role of a framework is simply to provide a guide, rather like the check-list that is now widely used in surgical operations. And the role of a model is to provide a tool to aid thinking; it does not do the thinking and it does not give the 'right' answer. It simply feeds back to participants the logical consequences of the data and clinical judgements that have been given to it. The model is rather like another participant, but one that does not talk back; it simply reflects back what it has been told, but in changed form. The computer requires human judgement to tell it the favourable and unfavourable effects, the input data, the judgements in the form of value functions and weights, then, no matter how large the problem, it unfailingly provides the logical consequences of

those inputs, thereby overcoming human limitations on the amount of information we can process [17] in our heads.

But that is not the end of the process. Participants may well feel uneasy about the displayed results. In a facilitated workshop, the next step is to explore the discrepancy between intuitions and computer results. By doing so, new insights inevitably develop, demonstrating that the computer is wrong, or that intuitions need to be changed or informed. With new insights, changes are made to the model or to intuitions, results examined, intuitions challenged again. After several iterations of this process a 'requisite model' [16] emerges—sufficient in form and content to resolve the benefit-risk issues. At this point, the group is much clearer about what recommendations to make, and they can turn the results of the requisite model into words, sometimes supported by graphs from the model output. (For reporting purposes, four displays seem to be sufficient: the Effects Tree, the Effects Table, the Added-Value Bar Graph and the Difference Display.) It is even possible for the group to reject the model's results; often, these are the models a group has found to be most helpful, because it gave results that challenged thinking, which led to new insights and perspectives that allowed members of the group to agree on the way forward.

In short, the work reported here is to serve as an aid to informed, science-based regulatory decisions about medicinal products. It is intended to supplement and enhance human capability to deal with complex issues.

6. References

1. European Medicines Agency. Benefit-Risk Methodology Project. London: European Medicines Agency; 2009.
2. European Medicines Agency. Work Package 1 Report: Description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network. London: European Medicines Agency; 2009.
3. European Medicines Agency. Work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. London: Access at www.ema.europa.eu, Special topics, Benefit-risk methodology; 2010.
4. Phillips LD. Decision Conferencing. In: Edwards W, Miles RF, von Winterfeldt D, editors. Advances in Decision Analysis: From Foundations to Applications. Cambridge: Cambridge University Press; 2007.
5. Raiffa H. Decision Analysis. Reading, MA: Addison-Wesley; 1968.
6. Kahneman D, Slovic P, Tversky A, editors. Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press; 1982.
7. Hammond JS, Keeney RL, Raiffa H. Smart Choices: A Practical Guide to Making Better Decisions. Boston, MA: Harvard University Press; 1999.
8. Tversky A. Elimination by aspects: A theory of choice. In: Shafir E, editor. Preference, belief, and similarity: selected writings: The MIT Press; 2003.
9. Phillips LD, Phillips MC. Facilitated work groups: Theory and practice. Journal of the Operational Research Society. 1993;44(6):533-49.
10. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. Science. 1974;185:1124-31.
11. Epley N, Gilovich T. The anchoring-and-adjustment heuristic: Why adjustments are insufficient. Psychological Science. 2006;17(4):311-8.
12. Keeney RL, Raiffa H. Decisions With Multiple Objectives: Preferences and Value Tradeoffs. New York: John Wiley, republished in 1993 by Cambridge University Press; 1976.
13. 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.
14. Mussen F, Salek S, Walker S. Benefit-Risk Appraisal of Medicines: A Systematic Approach to Decision-Making. Chichester: John Wiley & Sons; 2009.
15. 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.
16. Phillips LD. A theory of requisite decision models. Acta Psychologica. 1984;56(1-3):29-48.
17. Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 1956;63(2):81-97.

Appendix—The PrOACT-URL process

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, life-threatening or not, 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**. (Steps 2 and 3 are interchangeable in order.) 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?). Establish measurement scales for all criteria.
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, dosages; post-approval: do nothing, limit duration, change dosage, restrict indication, suspend).
4. **CONSEQUENCES**. Based on available data, describe how the alternatives perform on the criteria, i.e., describe the magnitude of possible favourable and unfavourable effects. It may be helpful to consider intermediate outcomes, such as safety and efficacy effects. Describe the clinical relevance of the effects, effectiveness and frequency of favourable effects, and severity and incidence of unfavourable effects. Create an 'Effects Table' with alternatives in columns and criteria in rows. Write descriptions of the effects in each cell, qualitative and quantitative (including statistical summaries with confidence intervals). It may at this stage be helpful to record the basis for uncertainties about the consequences in preparation for step 6 (e.g., possible biases in the data, soundness and representativeness of the clinical trials, potential for unobserved adverse events), 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 changes by taking account of the uncertainty associated with the consequences.
7. **RISK ATTITIDE**. 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, life-threatening or not, affected population 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).