



# IMI Work Package 5: Supplement 2 to Wave 1

## Case Study Report 1:b:iii: Efalizumab

### Multi-Criteria Decision Analysis: Decision Conference

### 01/12/2011

Dr Larry Phillips and Mr Nikolaos Zafiroopoulos, European Medicines Agency  
On behalf of PROTECT Work Package 5 participants

Version one dates 23 Jan 2013 Date of any subsequent amendments below	Person making amendments	Brief description of amendments
17/06/2013	Shahrul Mt-Isa	Updated trade name to generic, plus other editorial changes.

[https://eroombayer.de/eRoomReq/Files/PH-GDC-PI-SID/IMI-PROTECT/0\\_f9082/PROTECT WP5 report template.docx](https://eroombayer.de/eRoomReq/Files/PH-GDC-PI-SID/IMI-PROTECT/0_f9082/PROTECT WP5 report template.docx)

Disclaimer: The processes described and conclusions drawn from the work presented herein relate solely to the testing of methodologies and representations for the evaluation of benefit and risk of medicines. This report neither replaces nor is intended to replace or comment on any regulatory decisions made by national regulatory agencies, nor the European Medicines Agency

Acknowledgements: The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium, [www.imi-protect.eu](http://www.imi-protect.eu)) which is a public-private partnership coordinated by the European Medicines Agency. The PROTECT project has received support from the Innovative Medicines Initiative Joint Undertaking ([www.imi.europa.eu](http://www.imi.europa.eu)) under Grant Agreement n° 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution

## EXECUTIVE SUMMARY

During a decision conference at the EMA on 1 December 2011, seven members of the efalizumab Case Study Team of the PROTECT project developed a decision-theory-based model<sup>1</sup> for evaluating the benefit-risk balance of efalizumab, a drug for the treatment of adult patients with moderate to severe chronic plaque psoriasis who have failed to respond to other systemic therapies, compared to a placebo. The decision conference took the view of regulators in early 2009, when they were assessing the benefit-risk balance in light of new information received post-authorisation. This report summarises the process and results of the decision conference.

The group considered five favourable effects and ten unfavourable effects, the latter representing five effects from the clinical trials, and five from post-marketing observational data (p. 7, Effects Tree, Figure 1). Each criterion was carefully defined to enable meaningful evaluations of the drugs (pp. 8-9, Effects Table, Table 1). Measurement scales used in the clinical studies were identified for all the criteria.

Pooled data from 5 (five) phase III studies provided measures on the five favourable effects and the five unfavourable effects criteria observed in the clinical trials. Data for the five observational criteria were taken from the Merck Serono PSUR 10 document. Measures for each criterion were converted to preference values on 0-100 scales that were defined as encompassing the range of data, plus possible uncertainties, for each criterion (p. 8, Table 1 gives the ranges; page 11, Figure 3 provides an example). All conversions of measures to preferences employed direct or inverse linear transformations, except for PML, for which an inverse convex value function was judged by participants' to capture the clinical relevance of this effect (p. 10, Figure 2). All input scores are shown in the Effects Table, while their associated preference values are shown in APPENDIX B—THE MODEL.

The group also assessed relative weights for all the criteria. These weights equate the units of preference value across all the criteria. The method of swing weighting, which requires comparative judgments about the ranges of effects and clinical judgements about how much they matter relative to each other, made it possible to assign meaningful relative weights to all scales (p. 14, Figure 7). These weights reflect both the range from the least to most preferred effects on each scale, a matter of fact, and how much those effect differences matter, a consideration of clinical relevance that takes the context for decision making into account. The model's separation of facts from judgements ensures that swing-weights are scale constants, whereas the more commonly-asked question "how important is this effect compared to that one", does not yield meaningful scale constants.

Weighted averages of the scores, calculated by a computer and projected on-the-spot for the group as the model was constructed, provided a single, overall score for each treatment, with efalizumab scoring 51 (out of a possible 100—which would indicate maximum scores on all the favourable effects and no unfavourable effects), and the placebo 31, showing that the drug is overall most preferred.

---

<sup>1</sup> The model represents value preferences, as in multi-criteria decision analysis (MCDA), and their uncertainties (as in decision tree analysis), so can be considered a mixed model.

Those scores are broken down into their favourable and unfavourable effect contributions (p. 16, Figure 9) or by the contributions of the individual criteria (p. 17, Figure 10). Comparisons of the drug with the placebo showed that the main advantages of the drug are the PGA and the PASI75, while the main disadvantage is its potential for PML (p.18, Figure 11). It is this latter display that is perhaps the most useful to regulators and assessors as it shows the differences between drug and placebo based on both the measured data, whatever its form (percentages, scores, change scores, etc.) and the clinical relevance of the data.

Sensitivity analyses showed that the model is very robust to very substantial changes in individual weights for all criteria except PML. The key trade-off is between the 0-60% range on the PASI75 scale and the 0-5 range on PML, which was initially judged to be in the ratio of 2 to 1 (p. 13, Figure 6). Changing that ratio to be about equal, i.e., 60% of patients experiencing a 75% reduction in baseline PASI judged to be as clinically desirable as 5 cases of PML is undesirable, causes the overall benefits to be just balanced by the overall risks. Further increasing the weight on the PML scale causes the risks to exceed the benefits.

Modelling efalizumab at this point in time, two years after the drug was withdrawn, proved to be difficult because the judgements made in 2009 by the assessors and regulators are not recoverable. It is not even possible to know precisely what data led regulators to their decision, for none of the public documents, from 2004 onward, are clear about which criteria the assessors considered relevant to the benefit-risk balance, and which were not. So, though it was possible to model Raptive retrospectively, the model developed here may well be an incomplete representation of the all the explicit and implicit considerations assessors brought to bear at the time the assessment reports were written.

## PARTICIPANTS

Alain Micallef, Merck Serono

Diana Hughes, Pfizer (by teleconference)

Kimberley Hockley, Imperial College

Nan Wang, Imperial College

Torbjörn Callréus, Danish Medicines Agency

Dorina Bischof, Merck Serono (MD responsible for the safety of Efalizumab at the time of the MA suspension)

Dr Larry Phillips facilitated the decision conference, assisted by Mr Nikolaos Zafiropoulos (from the EMA's Benefit-Risk Project).

## Table of Contents

1	EFALIZUMAB BENEFIT-RISK APPRAISAL.....	6
2	MODEL STRUCTURE.....	6
2.1	The Options.....	6
2.2	The Criteria.....	6
2.3	Scoring the Options.....	11
2.4	Weighting.....	11
3	RESULTS.....	15
3.1	Overall.....	15
3.2	Comparative Analyses.....	17
3.3	Sensitivity Analyses.....	18
3.4	DISCUSSION AND CONCLUSIONS.....	21
4	APPENDIX A—DECISION CONFERENCING.....	23
5	APPENDIX B—THE MODEL.....	24
5.1	Node results.....	24
5.1.1	Overall Favourable-Unfavourable Effects Balance.....	24
5.1.2	Favourable Effects.....	24
5.1.3	Unfavourable Effects.....	25

## 1 EFALIZUMAB BENEFIT-RISK APPRAISAL

This report documents the process and results of a decision conference (a group modelling process described in APPENDIX A—DECISION CONFERENCING) on 1 December 2011 whose purpose was to create and explore a model of the benefit-risk balance for the drug efalizumab. The drug received marketing authorisation on 20 September 2004 for the treatment of adult patients with moderate to severe chronic plaque psoriasis who have failed to respond to other systemic therapies. By January 2009 the margin of benefits over risks had narrowed since approval, so the European Commission requested the CHMP to assess the concerns and its impact on the benefit/risk balance for efalizumab, to give its opinion on measures necessary to ensure the safe and effective use of efalizumab, and on whether the marketing authorisation for this product should be maintained, varied, suspended or withdrawn. The Marketing Authorisation Holder (MAH) did not wish to conduct further clinical trials, as the CHMP had required to lift the suspension recommended in February, so in June the European Commission withdrew the marketing authorisation for efalizumab.

This decision conference took the view of regulators in early 2009, when they were assessing the benefit-risk balance in light of the new information received post-authorisation. Two sources of data contributed to the benefit-risk model: the original 2004 EPAR and the PSUR 10 document provided by Merck Serono<sup>2</sup>. This report summarises the structure of the model developed at the decision conference and the results.

## 2 MODEL STRUCTURE

After a brief overview by Larry Phillips of the nature and purpose of a decision conference, he reminded participants of the primary task for the day: to develop a benefit-risk model of efalizumab, assuming a regulator's perspective in early 2009. Alain Micaleff and Kimberley Hockley had assembled the relevant data from the EPAR and PSUR into an extended Effects Table, which summarised the benefit and risk criteria as favourable and unfavourable effects, with their definitions, the relevant patient population from which the data were drawn, the measurement scales associated with the criteria, the units of measurement and the data. The Effects Table was created during the application of the ProACT-URL framework to the modelling of efalizumab. This pre-work expedited the work of the group in building a model.

### 2.1 The Options

The group recognised that data were available only for two options:

1. Efalizumab in 2009 (pre and post-marketing data)
2. Placebo in 2004 (premarketing data)

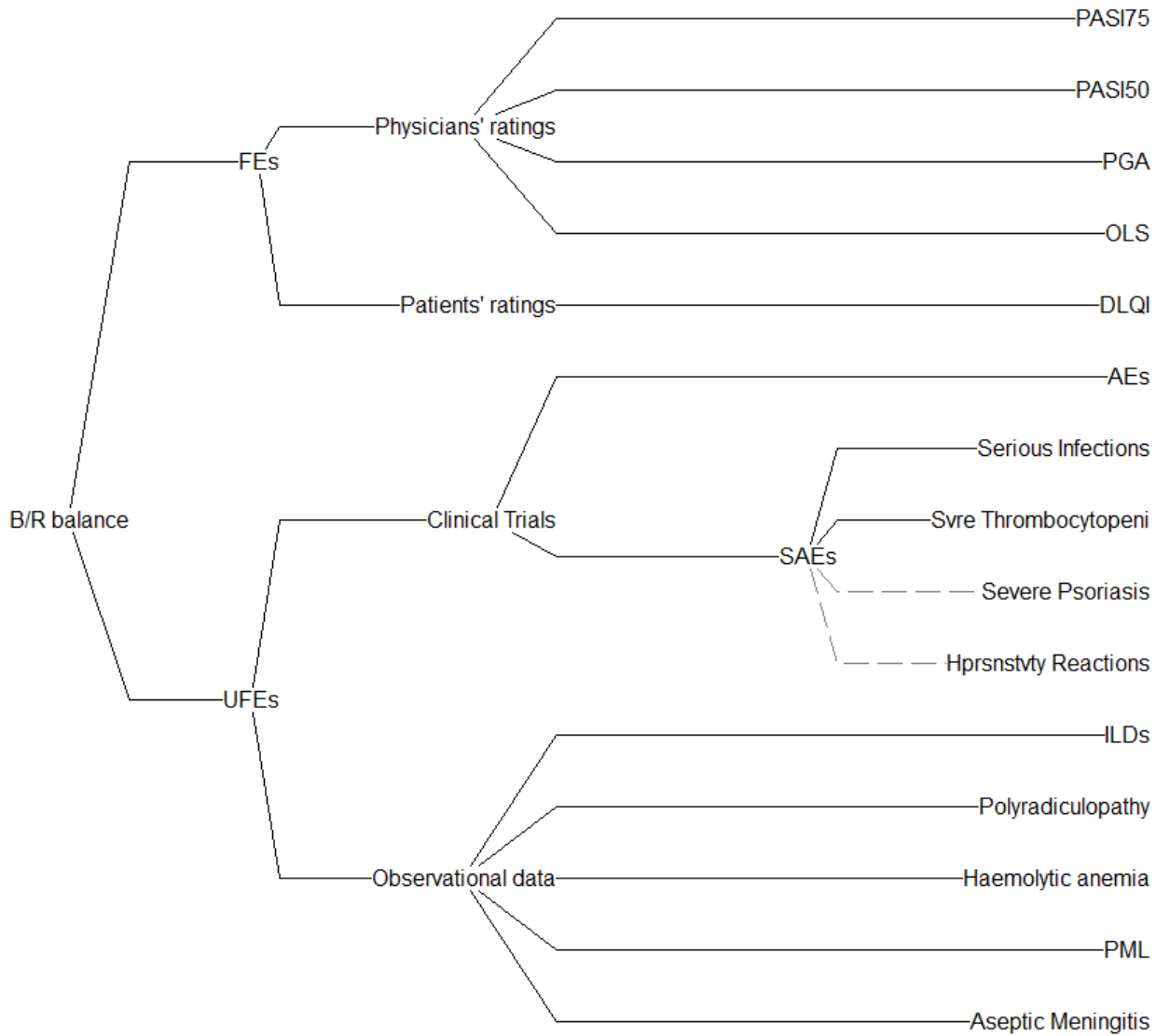
No data were available for an option discussed at the time both by Regulators and Company, resulting in a limitation of treatment to 2 years.

### 2.2 The Criteria

Five favourable effects and ten unfavourable effects characterise the final model. The clinical trials conducted prior to approval provided data for the five favourable effects and for five of the unfavourable effects, while the Merck Serono PSUR 10 document provided data for the other five unfavourable effects. Although the available documentation reports many effects, the group chose to model only those effects that might affect the benefit-risk balance; thus, many unfavourable effects were not included in the model. The Effects Tree, Figure 1, shows favourable and unfavourable effects at the nodes, and criteria against which the drugs are evaluated at the extreme right.

---

<sup>2</sup> PSUR 10 was the last Periodic Safety Update Report submitted to EMA in November 2008 before Market Authorisation suspension in February 2009.



**Figure 1: The evaluation criteria organised by Favourable Effects (FE) and Unfavourable Effects (UFE). The weights assigned to Severe Psoriasis and Hypersensitivity Reactions were so small that their cumulative weights are effectively zero, indicated by the dashed lines.**

An analysis of the data after the decision conference showed that although Serious Infections and Severe Thrombocytopenia were reported in the PSUR, they were less prevalent than in the clinical trials, where the model showed they had no effect on the benefit-risk balance, so they were not included as relevant criteria for the Observational Data.

Definitions of the criteria are given in Effects Table, Table 1. The table shows the short name given in Figure 1, the description of the effect, which in some cases are further explained in the footnotes, fixed upper and lower values that define a plausible range for the data, the units of measurement, and, finally, the data for efalizumab and the placebo. Data from more than one clinical trial were pooled to give the values shown in the Effects Table.

Table 1: Effects Table for efalizumab.

	Name	Description	Fixed Upper	Fixed Lower	Units	Efalizumab	Placebo
Favourable Effects	PASI75	Percentage of patients achieving 75% reduction in baseline PASI <sup>1</sup> at week 12.	60.0	0.0	%	29.5	2.7
	PASI50	Percentage of patients achieving 50% reduction in baseline PASI <sup>1</sup> at week 12.	60.0	0.0	%	54.9	16.7
	PGA	Percentage of patients achieving Physician's Global Assessment <sup>2</sup> clear/almost clear at week12.	40.0	0.0	%	29.5	5.1
	OLS	Percentage of patients with Overall Lesion Severity rating of minimal or clear at FT (day 84).	40.0	0.0	%	32.1	2.9
	DLQI	Dermatology Life Quality Index <sup>3</sup> . Mean percentage of patients showing an improvement.	10.0	0.0	Change score	5.8	2.1
Unfavourable Effects	AEs	Percentage of patients exhibiting injection site reactions, mild to moderate dose-related acute flu like symptoms.	50.0	20.0	%/100ptyrs	41.0	24.0
	Severe infections	Proportion of patients experiencing infections serious enough to require hospitalisation.	3.00	0.00	%/100ptyrs	2.83	1.4
	Severe Thrombocytopenia	Number of cases exhibiting severe (grade 3 and above) thrombocytopenia <sup>4</sup> .	10	0	number	9	0
	Psoriasis Severe Forms	Percentage of patients developing severe forms of psoriasis (erythrodermic, pustular).	4.0	0.0	%	3.2	1.4
	Common AEs as per SPC	Percentage of patients exhibiting hypersensitivity reactions, arthralgia, psoriatic arthritis, flares, back pain, asthenia, ALT and Ph. Alk increase.	10.0	0.0	%	5.0	0



**Pharmacoepidemiological Research on  
Outcomes of Therapeutics by a European Consortium**

Interstitial Lung Disease (ILD)	Number of cases of interstitial lung disease.	20	0	number	18	0
Inflammatory Polyradiculopathy	Number of cases of inflammatory polyradiculopathy.	5	0	Data	4	0
Haemolytic Anaemia	Number of cases of haemolytic anemia.	25	0	number	24	0
PML	Number of cases of progressive multifocal leukoencephalopathy.	5	0	number	3	0
Aseptic Meningitis	Number of cases of aseptic meningitis.	30	0	number	29	0

<sup>1</sup>PASI is a measure of the average redness, thickness and scaliness of the lesions (each graded on a 0-4 scale), weighted by the body region and the area affected. PASI range is from 0 to 72.

<sup>2</sup>PGA is a seven point scale with 7 being clear, 6 almost clear, 5 mild, 4 mild to moderate, 3 moderate, 2 moderately severe and 1 severe psoriasis.

<sup>3</sup>DLQI is a 10-item quality of life index scored by the patient on a four point scale.

<sup>4</sup>As shown in laboratory test results that indicate a decrease in number of platelets in a blood specimen.

The Hiview<sup>3</sup> computer program converted the scores of the drug and placebo on those measurement scales into 0-100 preference value scales. Either direct linear transformations (higher measures are more preferred) or inverse linear (lower measures are more preferred, as for mean change in PGA score). An exception was PML, for which a non-linear value function was deemed more appropriate over the whole range from 0 to 5 cases per patient year. Participants assessed the value function shown in Figure 2; this effectively captures the non-linear clinical relevance of the number of PML cases.

Weights later assigned to the criteria ensured the equality of units of the preference values on all scales. It is this conversion from different input measures into preference values, whose criterion scales are later weighted, that enable quantitative comparisons of benefits and risks.

It is apparent that some double-counting exists in the favourable effects. The proportion of patients achieving PASI75 is included in the proportion of patients PASI50. The subsequent weighting process took this into account by ensuring that the sum of weights on these two scales considered together was in the desired proportion to the other scales.

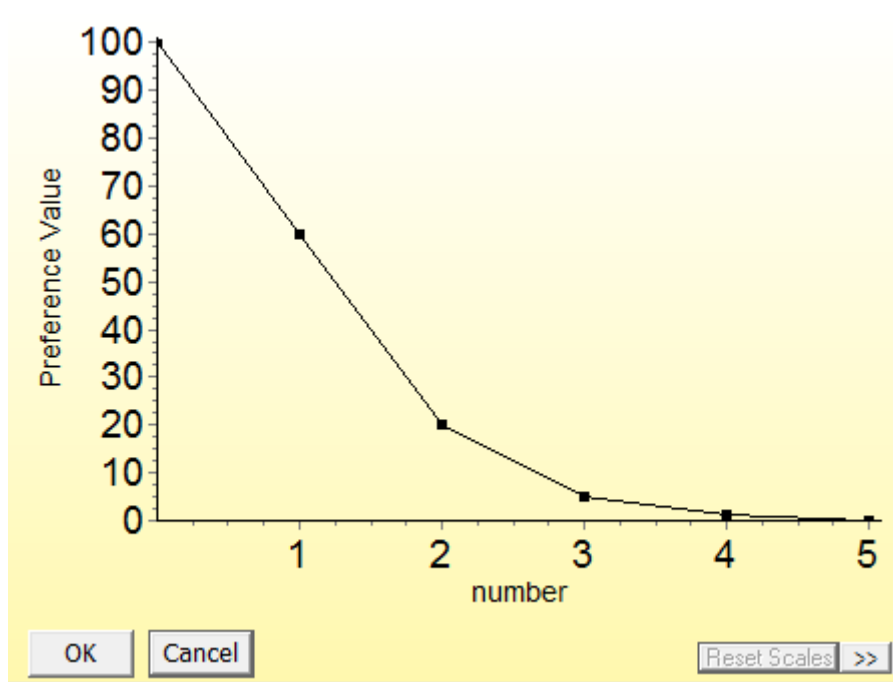
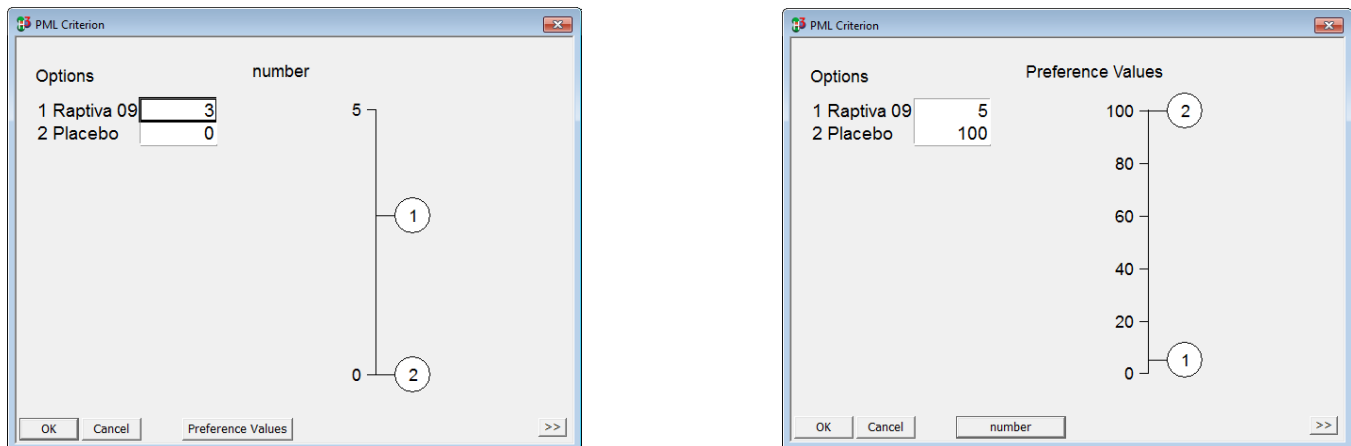


Figure 2: The group's assessed value function for number of PML cases.

<sup>3</sup> Hiview was originally developed at the London School of Economics & Political Science, and is now developed and available from Catalyze Limited, [www.catalyze.co.uk](http://www.catalyze.co.uk).

### 2.3 Scoring the Options

Measures expressing the performance of the options on each criterion were determined by the group on the basis of the pooled data, and entered into the computer. An example, PML, is shown in Figure 3. Input data on the left are displayed on the thermometer scale, whose range from 0 to 5 cases encompasses the entire range of uncertainty about this effect. The right panel shows the computer's inverse linear conversion of those scores onto a 0-100 preference scale.



**Figure 3:** Input data for the two options on the PML criterion, left panel, and their conversion into preference values, right panel, showing that lower proportions of the AE are more preferred, and that the non-linear value function, shown in Figure 2, substantially increases the difference between the drug and placebo.

At this stage in the analysis, all input data had been converted into 0-100 preference-value scales. As there are 10 such scales, the next task was to ensure that the units of preference value were equivalent across all the scales. That is the purpose of weighting.

### 2.4 Weighting

Some criteria are more clinically relevant expressions of preference value than others. Although that is an intuitively appealing statement, more precision is needed to enable the assessment of weights for the criteria. To ensure that assessed weights are meaningful, the concept of 'swing weighting' was applied. As an analogy, both Fahrenheit and Celsius scales contain 0 to 100 portions, but the swing in temperature from 0 to 100 on the Fahrenheit scale is, of course, a smaller swing in temperature than 0 to 100 on a Celsius scale; it takes 5 Celsius units to equal 9 Fahrenheit units. The purpose of weighting in decision theory is to ensure that the units of preference value on the different scales are equivalent, thus enabling weighted scores to be compared and combined across the criteria. Weights are, in essence, scale factors.

It follows, then, that to judge preference value, two steps in thinking must be separated. First, it is necessary to think about the difference in the measured effect represented by a preference value of 0, compared to the level of effect represented by a preference value score of 100. That is a straightforward assessment of a difference in effect, from the least preferred effect to the most preferred effect on that criterion. The next step is to think about how much that difference in effect matters; this is essentially a judgement of the clinical relevance of the difference in effect size. "How big is the difference and how much do you care about that difference?" This is the question that was posed in comparing the 0-to-100 swing in effect on one scale with the 0-to-100 swing on another scale.

During the decision conference participants first assessed weights within each right-most grouping of favourable effects, the four Physicians' ratings criteria first. Figure 4 shows the weights for those that grouping. The group agreed that the swing from 0% to 100% on the PASI75 scale was better than any of the other three 0% to 100% improvements, so the PASI75 was assigned a weight of 100. Compared to that, the group judged the swing on the PGA scale to be nearly as good, and agreed a weight of 80.

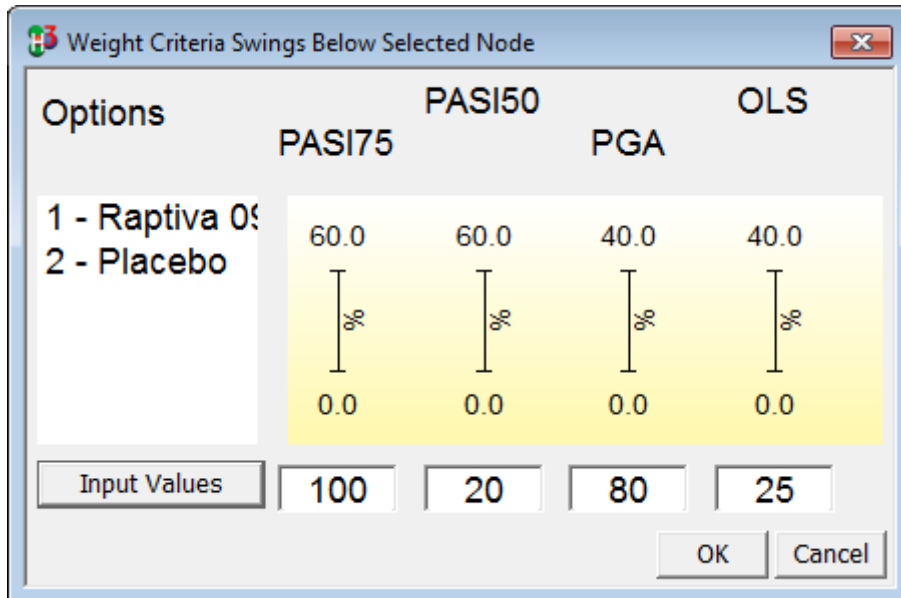


Figure 4: The swing-weights assigned to the four Physicians' ratings scales.

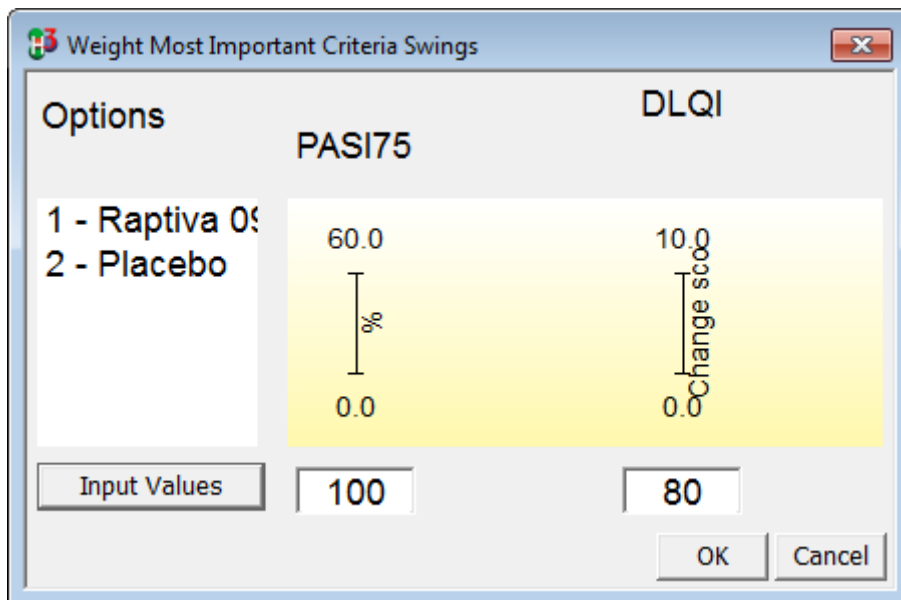


Figure 5: Swing weights assigned to the 100-weighted criteria for PASI75 and DLQI

In the next step, the group compared the PASI75 scale with the DLQI scale, assigning the latter a weight of 80 compared to the PASI75, as shown in Figure 5.

The group then turned to weighting the Unfavourable Effect criteria, starting with the SAE criteria; the largest swing weight was judged to be for Serious Infections, so that criterion was given a weight of 100. Next, that criterion was

compared to AE, which was assigned a weight of 20. Then, moving to the criteria under Observational data, the group quickly agreed that the 0-to-5 swing for PML was the most important, so it was given a weight of 100, and the other swings were judged relative to that 100. Comparing the 100-rated swing under Clinical Trials, Serious Infections, with the 100-rated swing under Observational data, PML, resulted in an assessed weight of 20 for Serious Infections compared to the 100 for PML.

The final, and most difficult comparison, is shown in Figure 6: PASI75 versus PML. After considerable debate, the group agreed that the PML swing, from 5 cases down to none, was half the clinical relevance of PASI75, from 0% to 60% of patients achieving PASI75. But sensitivity analysis on that weight was promised, for not everybody agreed that 2 to 1 was the final answer.

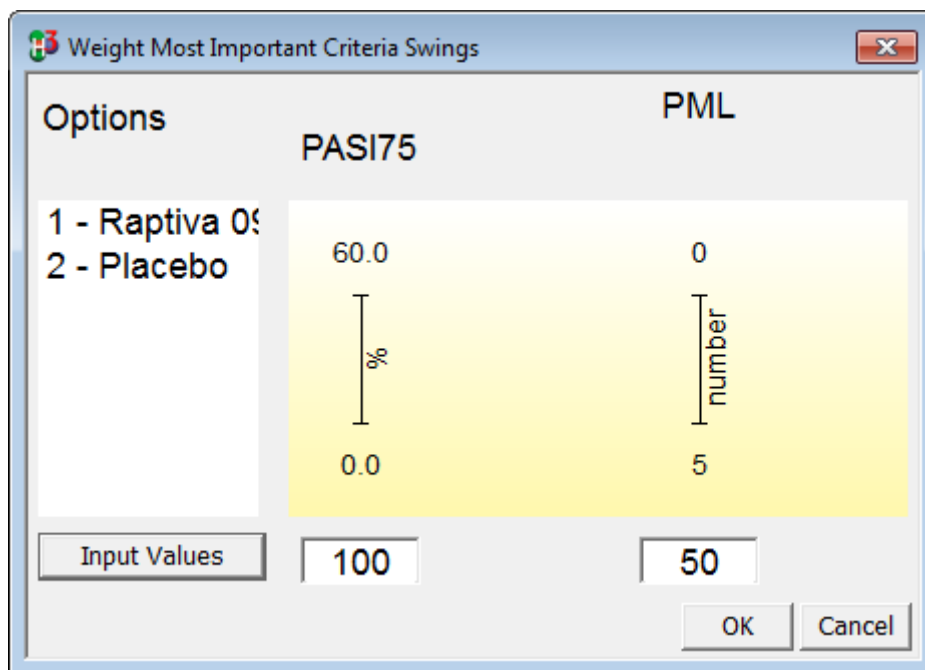


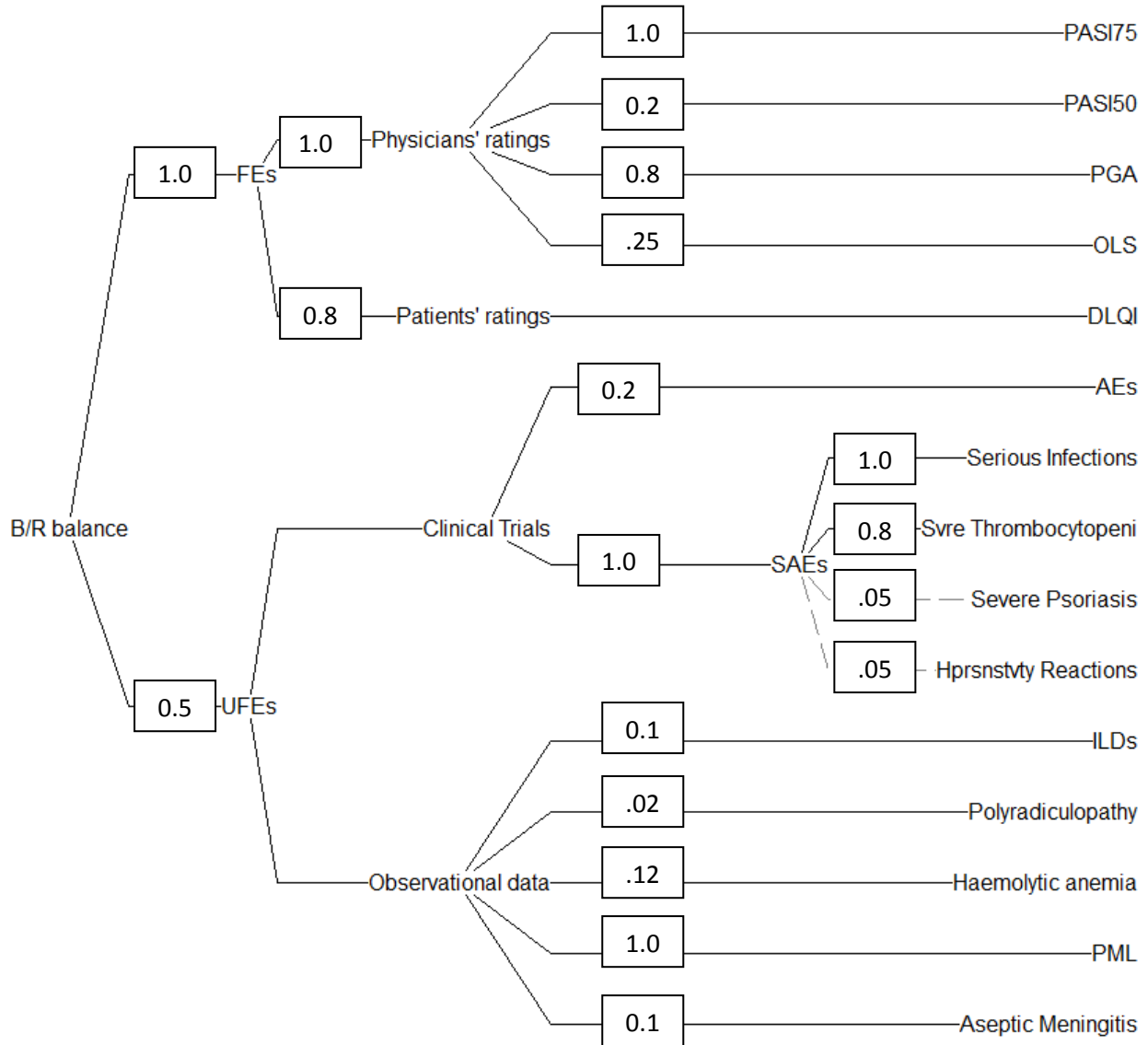
Figure 6: Swing weights comparing PASI75 to PML.

It is this process of comparing swings from least to most preferred positions on the criteria associated with a node, assigning one criterion swing a weight of 100, then comparing the 100-weighted criteria across the nodes, which ensures the comparability of the units of preference values across all the criteria.

It is easy to become lost in attempting to understand the weighting process by reading about it, so Figure 7 shows all the originally-assessed weights, each divided by 100, on the value tree. Hiview multiplies these weights along each path through the tree, sums the products for all 11 criteria and divides each product by the sum. This gives the cumulative weights shown in Figure 10, re-normalised to 100, with the criteria sorted in order of the cumulative weights.

It is important to keep in mind that a cumulative weight represents the total added preference value in moving from the least to most preferred positions on a scale. These weights represent the relative importance of the 0-100

preference value ranges on the scales, not the relative importance of favourable and unfavourable effects, and particularly not the relative importance of those effects for the drug and placebo. By summing cumulative weights, it is possible to see the weights at each node. For example, the sum of all the favourable effects weights is 78 with 22 for the unfavourable effects. In other words, the total range of 0-100 differences in preference values on the favourable effects three-and-a-half times the range of that on the unfavourable effects.



**Figure 7: The originally-assessed swing-weights, divided by 100, assigned at all the nodes.**

### 3 RESULTS

#### 3.1 Overall

With scoring and weighting completed, it was possible to calculate sums of weighted preference values and show preliminary results at any node. Figure 9 shows the relative scores at the FE/UF Balance node of Figure 1 as stacked bar graphs. Each section of each bar graph shows the contribution of favourable effects and unfavourable effects to the overall score, which is shown at the bottom of the bar. Note that longer green bars represent *more* benefit, while longer red bars represent *more* safety. Efalizumab shows a 20-point advantage over the placebo.

The stacked bar graphs can also be shown for their separate contributions from the criteria, as seen in Figure 10. This instructive display shows the three main advantages of Raptive: PASI75

PGA and DLQI. Collectively, they far outweigh the advantages of the placebo: its modest side effects and absence of PML. However, as the group learned, this result depends on the relative weights between the favourable and unfavourable effects, explored below.

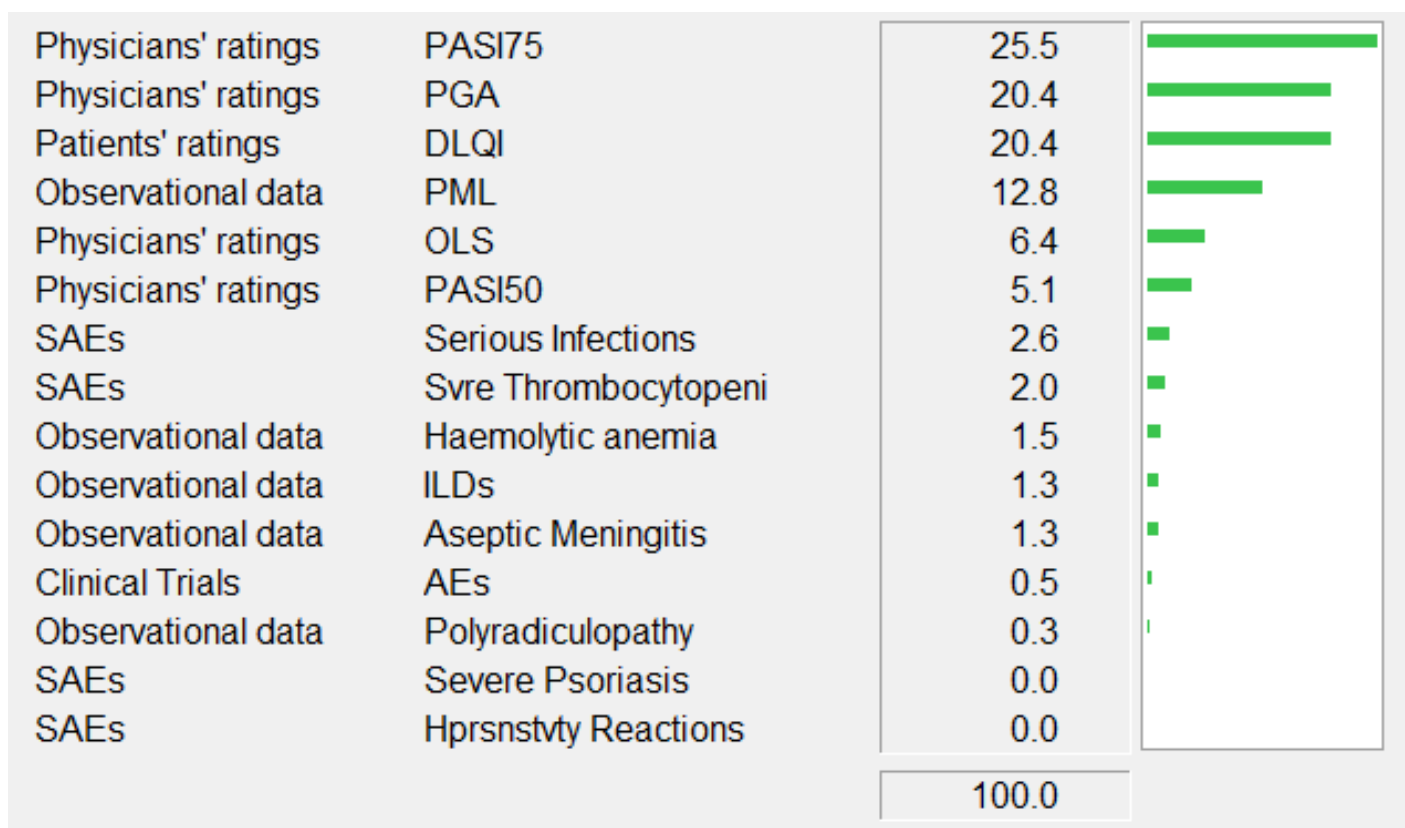


Figure 8: Cumulative weights of all the criteria, with the criteria ordered by the size of their cumulative weights, which represent the swings in preference from the least to the most preferred positions on the scales.

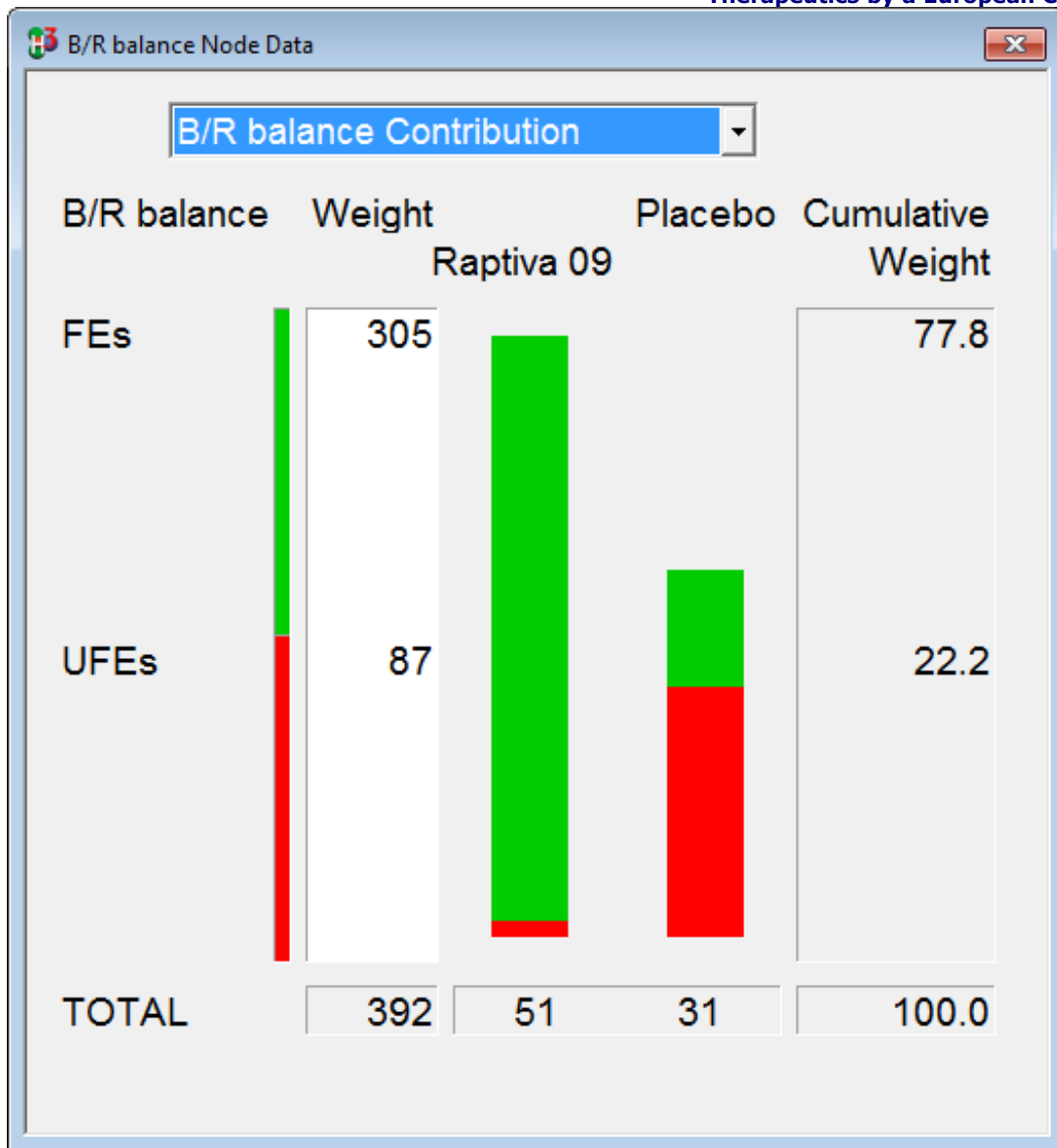


Figure 9: Overall Benefit-Risk balance for efalizumab. Longer green bars represent more benefit, while longer red bars show more safety. The Cumulative Weight column shows the normalised weight on the FE and UFE nodes, favourable effects weighted more than three times as much as for unfavourable effects.



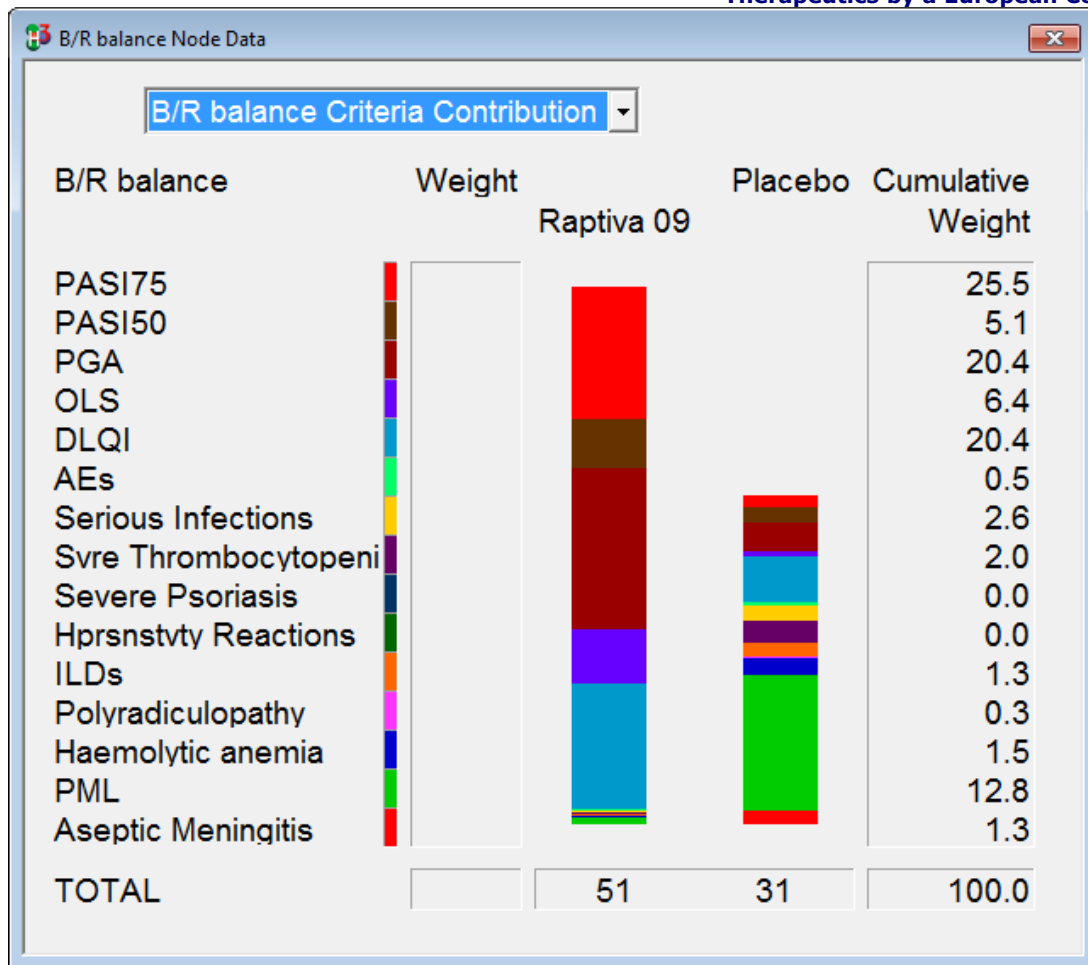


Figure 10: The drugs ordered by their overall weighted preference scores, with the stacked bar graphs showing the contribution to the overall score of the criteria. The right column shows the cumulative weights, normalised to 100, of each of the criteria. Flare rate, for example, is 20.2.

### 3.2 Comparative Analyses

A more clear display of the differences between efalizumab and the placebo can be seen in Figure 11. The Diff column in each display shows the difference in the preference scores, while the Wtd Diff column multiplies that difference by the cumulative weight on the criterion. It is this weighted difference that reveals the true advantages and disadvantages of the comparisons, criterion-by-criterion. They are the ‘part scores’, whose sum, 19.8, represents the overall weighted difference of preference values for the two options.

The two main advantages of efalizumab are PGA and PASI75. Note that the PASI50, the primary endpoint, is in fifth position. It shows a large preference-value difference of 60 compared to the placebo, but the weight on that criterion is a quarter as large as the weight on PASI75. For the latter, the difference score of 45 is smaller, but that criterion is more heavily weighted, so the weighted difference score on PASI75 of 11.4 is nearly four times as large as the weighted preference score on PASI50.

Although the efalizumab-Placebo difference for patients’ ratings, DLQI, is the smallest of the favourable effects at 37 points, it is on a heavily-weighted criterion, with the result that the weighted difference score is more than twice that of the primary endpoint.

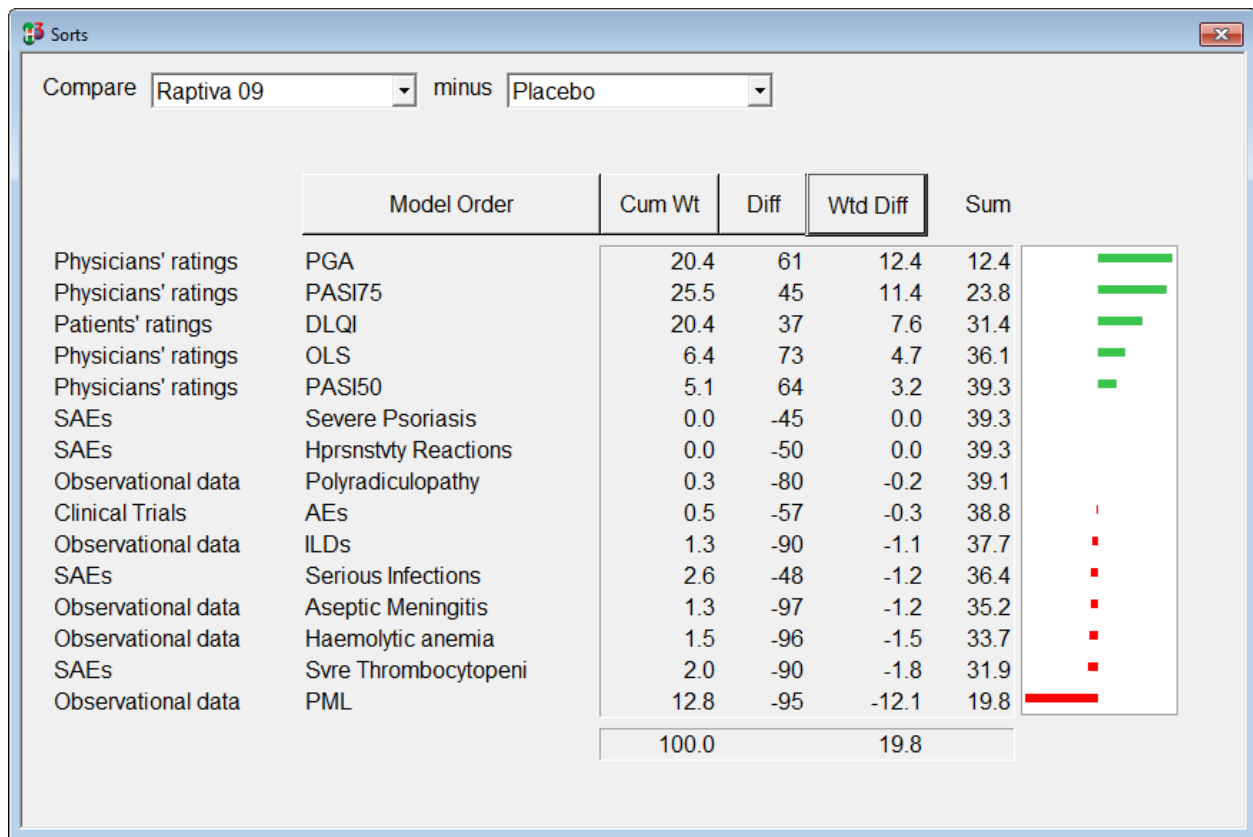


Figure 11: Efalizumab compared to the placebo. The sum of the five favourable effects, 39.3, outweighs the sum of the unfavourable effects, 19.5, to give an overall weighted preference value of 19.8 in favour of efalizumab over the placebo.

### 3.3 Sensitivity Analyses

These analyses explore the sensitivity of the overall results to changes in weights on the criteria, which were the source of much of the debate about the balance of benefits and risks. The first analysis examined the weight on the unfavourable effects to see if increasing that weight, and thereby decreasing the weight on the favourable effects (so that the total cumulative weights continue to sum to 100) would tip the benefit-risk balance in favour of the placebo. The normalised weight on the Unfavourable Effects node was 22.2, as shown in the right column of Figure 9. The computer varied that weight over its entire feasible range, 0 to 100, with the result shown in Figure 12.

The vertical red line intersects the horizontal axis at 22.2, and its intersections with the red and green lines give the overall scores for the efalizumab doses and the placebo, 31 and 51. Increasing the weight on the UFEs node increases the overall preference scores for the placebo and decreases the score for the drug. Increasing the cumulative weight to about 37 changes the most preferred option from efalizumab to the placebo, at the intersection of the two lines and indicated by the transition in background colour.

Brief experimentation with the relative swing weights on PASI75 compared to PML reveals that the two overall weighted sorts on the two options are 43 for efalizumab and 42 for the placebo when the weights shown in Figure 5 are 100-100, i.e., 60% of patients experiencing a 75% reduction in baseline PASI is as clinically desirable as 5 cases of PML is undesirable. This can be seen graphically in Figure 13.

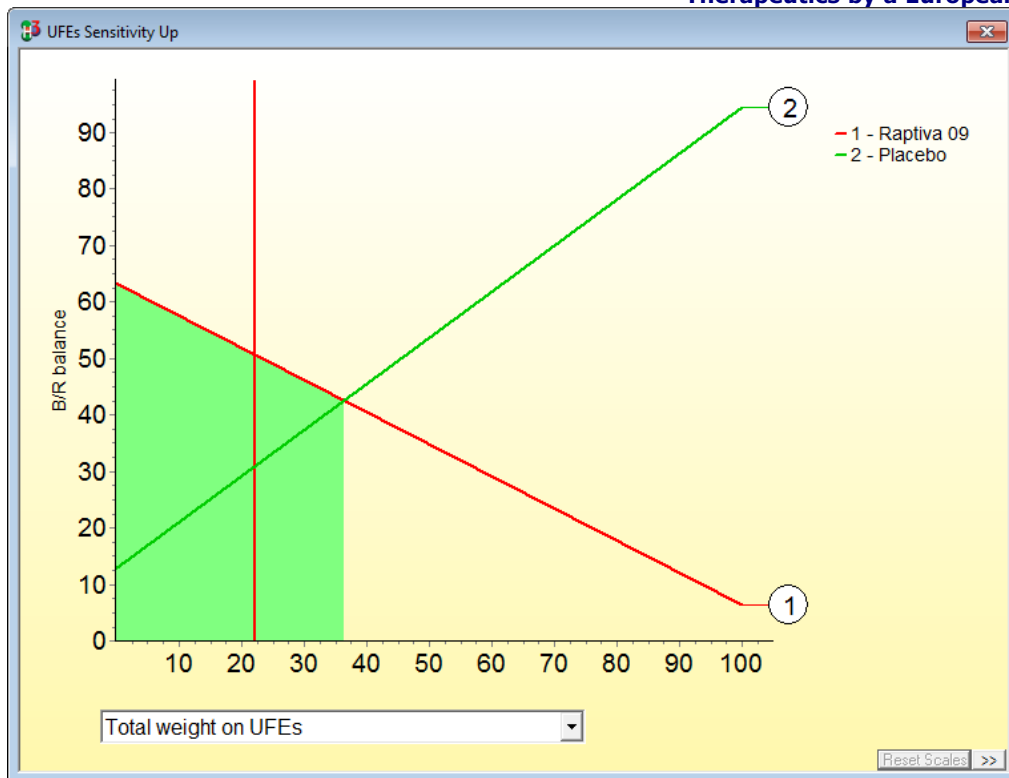


Figure 12: Increasing the weight on the UFE node from its current value of 24.1 shows that the weight would have to more than double for the placebo to be preferred.

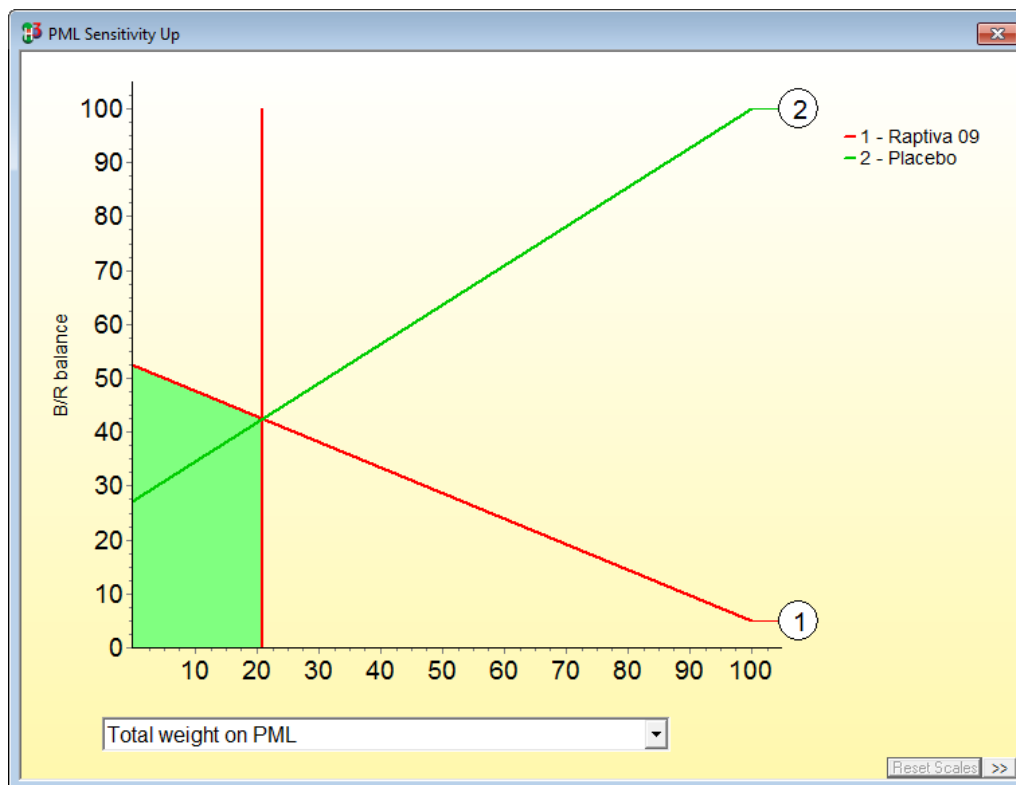


Figure 13: Increasing the weight on PML to equal that on PASI75 shows that equal clinical concern for these two effects results in equal overall weighted scores for efalizumab and the placebo.

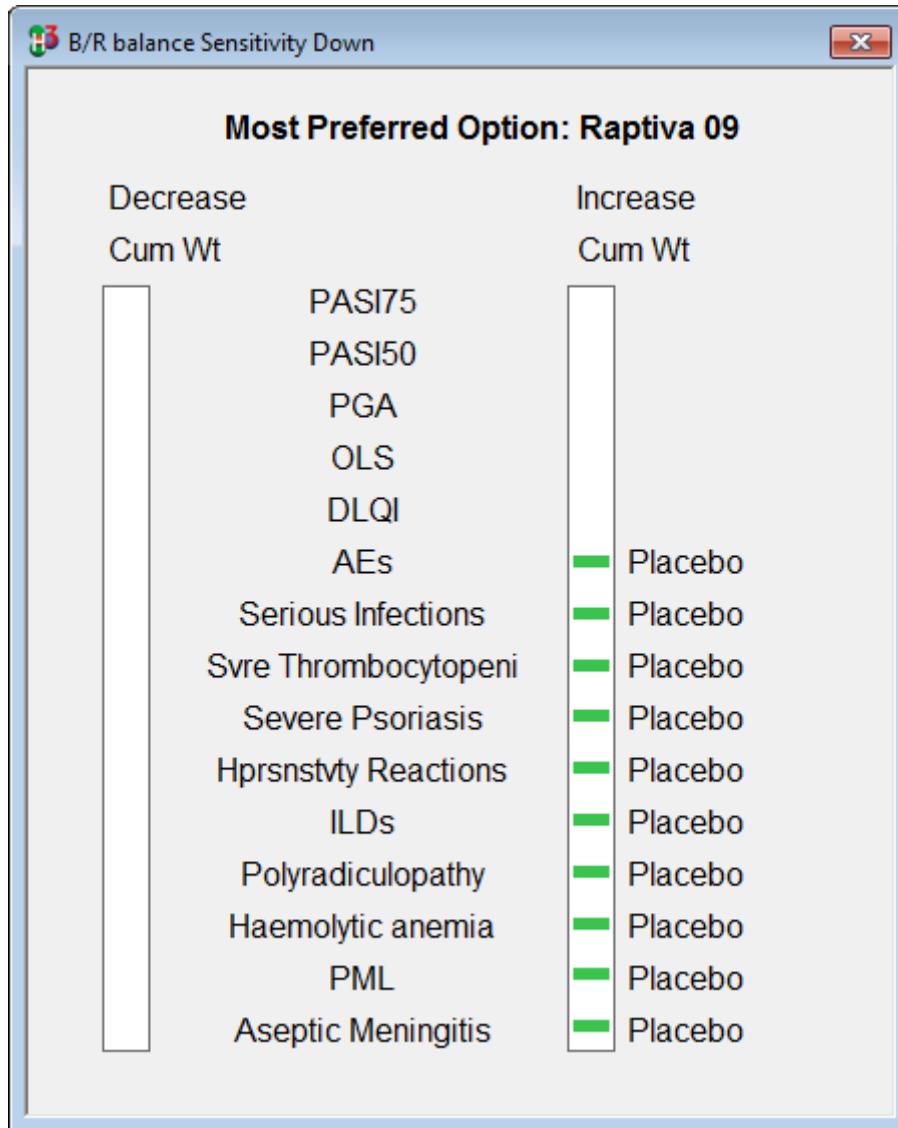


Figure 14: Separate sensitivity analyses on each of the criteria shows how the most preferred option, the efalizumab, would change as the cumulative weight on a criterion is decreased or increased. Green bars show cumulative weight changes greater than 15 points are needed to shift the overall preferences. Had a yellow bar appeared, it would signal a change of 5 to 15 points would change the result, while a red bar would indicate that a small change in a weight, less than 5 points, would change the most preferred option. Here, the absence of any bars for the five favourable effects, and no yellow and red bars indicates a robust model.

After returning the relative weights on PASI75 and PML to their base-case values of 100 and 50, the group explored whether or not there were any more crucial judgements that could shift the results. A simultaneous sensitivity analysis on all the criteria indicates which criterion weights make a difference. Figure 14 shows the summary display, with efalizumab at the top as the most preferred option. The middle column lists the criteria, while the right column shows the results of increasing the cumulative weight on each criterion independently, and the left column the result of decreasing the cumulative weight.

As noted in the previous analysis, PML just barely missed a yellow bar, but that really is the only sensitive criterion. The weight on any single unfavourable effect has to be increased substantially to change the overall result, while changing the weight on any single favourable effect, increasing or decreasing it, will not push the placebo into first place.

### 3.4 DISCUSSION AND CONCLUSIONS

The overall result of the modelling showed that the benefit-risk of efalizumab is substantially better than that of the placebo, even taking into account the three PML cases. This conclusion is robust to substantial differences of opinion about the individual weights on the criteria. Indeed, orders of magnitude increases would be required for the unfavourable effects, except for PML, to tip the balance. Only when more weight is given to 5 cases of PML compared to 60% of patients achieving a 75% reduction in baseline PASI would the model favour the placebo over efalizumab.

So, why did the CHMP recommend in February 2009 that marketing authorisation for efalizumab should be suspended? The official public statement explains that “its benefits in the treatment of psoriasis were modest, while there was a risk of serious side effects, including the occurrence of progressive multifocal leukoencephalopathy (PML)”. The suspension could be lifted if a sub-population could be identified for whom the benefits would outweigh the risks. The Marketing Authorisation Holder declined to conduct the necessary clinical trials, so the European Commission withdrew marketing authorisation for Raptive in June.

Is there a conflict between the decision of the CHMP and the model results reported here? The answer is “not necessarily”. Models don’t make decisions; people do. Models simply reflect back, in changed form, the information given to them. For the efalizumab model, the information provided includes the criteria shown in the Effects Tree, the measured data from the clinical trials and the incidences of unfavourable effects from the post-authorisation period, the judgement of the value function for PML and the assessments of swing-weights for the criteria. The pooled information on which the model results are based does not necessarily reflect all the available information, for the data are not always reported fully in the publicly-available reports. It is difficult to reconstruct today what was in the minds of assessors in 2004, 2008 and 2009, what data they used and how they pooled the available information. Modelling is best done at the time when a recommendation is required and the issues are ‘hot’. Thus, a shortcoming of the model reported here is that it may not adequately reflect the situation experienced by assessors in early 2009.

By 2009, information in addition to the clinical studies had become available, but it is difficult to determine from the public assessment reports what new information led to the view that efalizumab’s benefits were “modest”. Indeed, the April 2008 Assessment Report for Raptiva® (EMA/112794/2009) notes that for Study 25300 “the response rate [PASI75] in patients (n=232) who were refractory to all three major systemic treatments (i.e. cyclosporin, methotrexate, and PUVA) was 61% versus 69% in patients not refractory for any of these (p=0.03)”. From the perspective of the patient who was unresponsive to the other treatments, this is not a modest effect.

But the reporting raises the issue of what is meant by a ‘modest effect’. That phrase first appears in the EPAR, on page 36, as a summary of the finding that 27% of patients achieved PASI 75 (the primary endpoint). Data reported in the Effects Table in this report show similar percentages of patients achieving some sort of improvement, judged by physicians or patients. All the percentages shown there are around 30% (except for the PASI 50, which is generally disregarded in the Assessment Reports as being of little clinical significance). It would appear that ‘modest’ is more a public health interpretation, in that less than one third of psoriasis sufferers would be helped, than it is an indication

that the efficacy itself will be modest. In other words, a psoriasis patient reading the EPAR might conclude that he or she would only experience modest relief, when in fact the data show that for responders the efficacy could be considerable.

In short, the public health perspective of regulators can lead to potential communication problems for failing to distinguish between the magnitude of an effect from an individual's uncertainty that they will benefit from the effect. It might have been clearer to report that "27% of patients can expect to experience a 75% reduction in their condition".

Returning now to the question of whether the efalizumab model conflicts with the CHMP's final recommendation to withdraw the product, it is important to recognise that the function of a decision model is to serve as a 'tool for thinking', a decision aid that provides as many answers as there are judgements and assumptions provided as inputs. Many answers arise from disagreement about inputs. Experience of modelling five drugs during the EMA's Benefit-Risk Project, and more generally of working with teams of stakeholders and key players, shows that experts and assessors frequently disagree. Bringing them together in groups allows them to share their differing perspectives and experience so that informed assumptions, judgments and assessments can be tested for their effects on the overall benefit-risk balance, as described in APPENDIX A—DECISION CONFERENCING. Thus, a model gives as many different results as there are different inputs, but the process will enable the assessors to achieve a shared understanding of the important factors that affect the benefit-risk balance, to develop a sense of whether the benefit-risk balance is favourable or unfavourable, and, finally, agreement about what recommendations to make. Consensus about inputs is not required to achieve this level of agreement about the way forward.

## 4 APPENDIX A—DECISION CONFERENCING

The approach taken to constructing a benefit-risk model is based on *decision conferencing*<sup>4</sup>. This is a socio-technical process that combines working in groups helped by an impartial facilitator, on-the-spot computer-based modelling of data and participants' judgments, and continuous visual display of the model and its results. The 'socio' aspect of the process relies on mobilizing the right people at the right time to give the right inputs to the model. The 'technical' part refers to the model itself. This is based on decision analysis, first introduced in 1968 by Howard Raiffa<sup>5</sup>, and extended in 1976 by Ralph Keeney and Howard Raiffa<sup>6</sup> to cover decisions with multiple objectives, now an accepted methodology for dealing with decisions that are characterized by uncertainty and multiple objectives<sup>7</sup>.

The generic purposes of decision conferencing are to achieve a shared understanding of the issues (though not necessarily consensus), a sense of common purpose (while preserving individual differences of opinion) and a commitment to the way forward (though allowing individual differences in the paths). The idea is to encourage individual creativity, and to use differences of perspective to find ways forward that will gain support from those implementing the actions. A key assumption of decision conferencing is the notion of 'requisite modelling'<sup>8</sup>: that a model should be just sufficient in form and content to resolve the issues at hand. For benefit-risk analysis of drugs, the model need not be more complex than is sufficient to determine if the benefits outweigh the risks and to determine what additional information might be necessary. The model is a 'tool for thinking' enabling participants to see the logical consequences of differing viewpoints, and the effects of uncertainty on the benefit-risk balance.

A decision conference typically moves through four stages. The first stage is a broad exploration of the issues. In the second stage, a model is constructed of the favourable and unfavourable effects, incorporating available data and participants' judgements about clinical relevance of the effects. In the third stage, the model combines the effects and shows the benefit-risk balance. Extensive sensitivity analyses examine the effects on the balance of imprecision in the data, uncertainties, and differences in participants' risk tolerance. Discrepancies between model results and members' judgements are examined, causing new intuitions to emerge, new insights to be generated and new perspectives to be revealed. Revisions are made and further discrepancies explored; after several iterations the new results and changed intuitions are more in harmony. Then the group moves on to the fourth stage summarising key issues and conclusions, formulating next steps and, if desired, agreeing recommendations. The facilitator prepares a report of the event's products after the meeting and circulates it to all participants.

<sup>4</sup> Information about decision conferencing can be found on the website maintained by the London School of Economics and Political Science, at <http://www.lse.ac.uk/collections/decisionconferencing/>.

<sup>5</sup> Raiffa, H. (1968). *Decision Analysis*. Reading, MA: Addison-Wesley.

<sup>6</sup> Keeney, R. L. and H. Raiffa (1976). *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*. New York, John Wiley. The only book describing multi-criteria modeling across proposals, the approach used in PEP, is in Chapter 12, "Resource allocation and negotiation problems" of Goodwin, P. and G. Wright (1998). *Decision Analysis for Management Judgment, 2nd edition*. Chichester, John Wiley. That chapter is better understood by first reading Chapter 2, "Decisions involving multiple objectives."

<sup>7</sup> For additional information about benefit-risk methodologies for regulators see the WP2 report on the EMA public website. Click on the Special Topics tab, then on Benefit-Risk Methodology in the left column, and choose the pdf file "Benefit-risk methodology project work package 2 report".

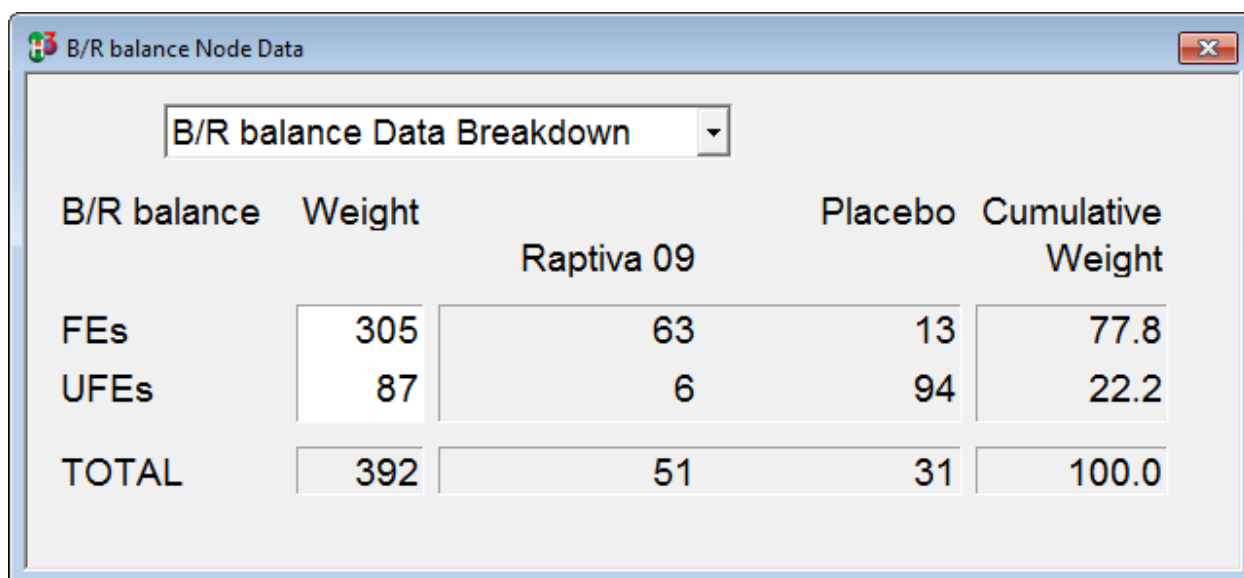
<sup>8</sup> Phillips, L. D. (1984). A theory of requisite decision models. *Acta Psychologica* 56: 29-48.

## 5 APPENDIX B—THE MODEL

### 5.1 Node results

Each of the following matrices corresponds to a node in the value tree of Figure 1. The scores shown are the linear conversion of the input scores onto 0-100 scales. The weights shown in the left column are the sums of the original weights at lower nodes. The final cumulative weights, obtained after a further normalisation to ensure all criterion weights sum to one, are shown in the right column. Asterisks identify criteria at the extreme right of the Effects Tree.

#### 5.1.1 Overall Favourable-Unfavourable Effects Balance

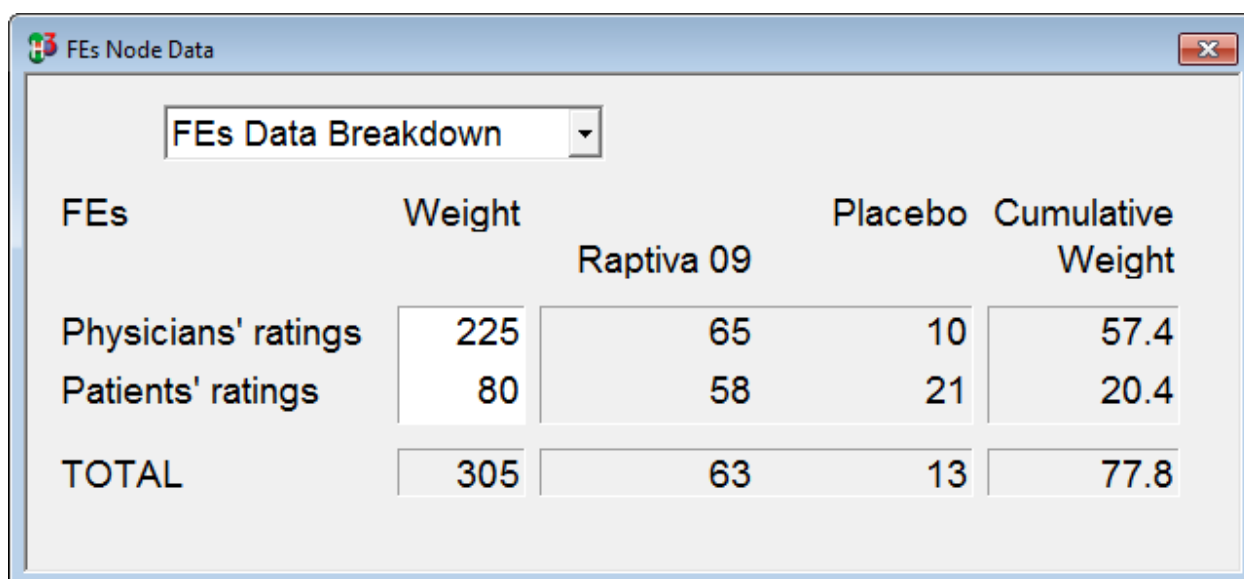


B/R balance Node Data

B/R balance Data Breakdown

B/R balance	Weight	Raptiva 09		Placebo	Cumulative Weight
FEs	305	63	13	77.8	
UFEs	87	6	94	22.2	
TOTAL	392	51	31	100.0	

#### 5.1.2 Favourable Effects



FEs Node Data

FEs Data Breakdown

FEs	Weight	Raptiva 09		Placebo	Cumulative Weight
Physicians' ratings	225	65	10	57.4	
Patients' ratings	80	58	21	20.4	
TOTAL	305	63	13	77.8	



Physicians' ratings Node Data

Physicians' ratings Data Breakdown

Physicians' ratings	Weight	Raptiva 09		Placebo	Cumulative Weight
PASI75*	100	49	5		25.5
PASI50*	20	92	28		5.1
PGA*	80	74	13		20.4
OLS*	25	80	7		6.4
<b>TOTAL</b>	<b>225</b>	<b>65</b>	<b>10</b>		<b>57.4</b>

5.1.3 Unfavourable Effects

UFEs Node Data

UFEs Data Breakdown

UFEs	Weight	Raptiva 09		Placebo	Cumulative Weight
Clinical Trials	20	10	75		5.1
Observational data	67	5	100		17.1
<b>TOTAL</b>	<b>87</b>	<b>6</b>	<b>94</b>		<b>22.2</b>

Clinical Trials Node Data

Clinical Trials Data Breakdown

Clinical Trials	Weight	Raptiva 09		Placebo	Cumulative Weight
AEs*	2	30	87		0.5
SAEs	18	8	74		4.6
TOTAL	20	10	75		5.1

SAEs Node Data

SAEs Data Breakdown

SAEs	Weight	Raptiva 09		Placebo	Cumulative Weight
Serious Infections*	10	6	53		2.6
Svre Thrombocytopeni*	8	10	100		2.0
Severe Psoriasis*	0	20	65		0.0
Hprsnstvy Reactions*	0	50	100		0.0
TOTAL	18	8	74		4.6

Observational data Node Data

Observational data Data Breakdown

Observational data	Weight	Raptiva 09	Placebo	Cumulative Weight
ILDs*	5	10	100	1.3
Polyradiculopathy*	1	20	100	0.3
Haemolytic anemia*	6	4	100	1.5
PML*	50	5	100	12.8
Aseptic Meningitis*	5	3	100	1.3
<b>TOTAL</b>	<b>67</b>	<b>5</b>	<b>100</b>	<b>17.1</b>