# IMI-PROTECT Benefit-Risk Group

# RECOMMENDATIONS REPORT

# Recommendations for the methodology and visualisation techniques to be used in the assessment of benefit and risk of medicines

Prepared on behalf of the PROTECT Consortium by:   Diana Hughes[1] Diana.Hughes@pfizer.com, Ed A J Waddingham e.waddingham@imperial.ac.uk, Shahrul Mt-Isa s.mt-isa@imperial.ac.uk, Alesia Goginsky alesia.goginsky@pfizer.com, Edmond Chan kkchan@doctors.org.uk, Gerald Downey gdowney@amgen.com, Christine E. Hallgreen c.hallgreen@imperial.ac.uk, Kimberley S. Hockley kimberley.hockley08@imperial.ac.uk, Juhaeri Juhaeri Juhaeri.Juhaeri@sanofi.com, Alfons Lieftucht alfons.x.lieftucht@gsk.com, Marilyn A. Metcalfe marilyn.a.metcalf@gsk.com, Rebecca A. Noel Noelrn@lilly.com, Larry Phillips larry_phillips@msn.com, Deborah Ashby[2] deborah.ashby@imperial.ac.uk, and Alain Micaleff[2] alain.micaleff@merckgroup.com,

on behalf of IMI-PROTECT Work Package 5.

[1] Team leader of IMI-PROTECT Work Package 5 Recommendation Report
[2] Co-leaders of IMI-PROTECT Work Package 5

# Executive Summary

In support of the goal to strengthen the monitoring of the benefit-risk balance of medicines in Europe, four years of research was conducted by a public-private partnership under the auspices of the EU and coordinated by the EMA and representatives of the pharmaceutical industry. The results of this work confirmed the added value of using more formal and structured approaches to the benefit-risk assessment of medicines to improve the transparency and communicability of this process.

Under the umbrella of the IMI PROTECT consortium, members have advanced the understanding of both the integration and visual representation of benefit and risk data. Following a robust review of the literature, selected methodologies and visualisation techniques were applied in several case studies, each one constructed from publicly available data and representative of the more challenging benefit-risk assessments encountered throughout the life cycle of a drug.

The experience of the case study teams has been distilled into a clear set of practical recommendations for benefit-risk decision processes and supporting tools, and these are organised around the five stages of a generic benefit-risk assessment roadmap:

I. Planning: This stage encourages stakeholders to focus on critical issues related to benefit-risk assessment, including the purpose and context of the assessment. Clear documentation of discussions allows future analyses and updates to utilise the same foundations.

   Useful methodologies included frameworks, such as the Benefit-Risk Action Team (BRAT) and Problem, Objectives, Alternatives, Consequences, Trade-offs, Uncertainty, Risk and Linked decisions (PrOACT-URL) frameworks that organise data, with tree diagrams and structured tables providing useful means of visualisation.

II. Evidence gathering and data preparation: This stage identifies data sources and extracts evidence relevant to the benefit-risk assessment, and may include aggregation of multiple sources of evidence, which may require the use of estimation techniques. It encourages the systematic handling of missing data and requires engagement of clinical, statistical, epidemiological, and database expertise.

   Useful methodologies include Indirect/Mixed Treatment Comparison (ITC/MTC) and Probabilistic Simulation Method (PSM), and visualisation techniques such as structured and colour-coded tables, and network graphs to enhance the communication of data.

III. Analysis: In this stage, the data are evaluated, quantifying the magnitudes of benefits and risks, and perhaps weighing and/or integrating favourable and unfavourable effects as required by a given approach.

   Useful methodologies for analysis include metric indices which provide numerical representations of benefits and risks (Number Needed to Treat / Number Needed to Harm (NNT/NNH), Impact numbers),

quantitative frameworks which model benefit-risk trade-off and balance benefits and risks (Multi-Criteria Decision Analysis (MCDA), Stochastic Multi-criteria Acceptability Analysis (SMAA)), and utility survey techniques which elicit stakeholders' preference information (Discrete Choice Experiment (DCE)).

Visualisations recommended for the analysis stage include visualisation techniques specific for eliciting value preferences (tree diagram, method-specific visualisations such as MACBETH grid, Analytic Hierarchy Process (AHP) table, swing-weighting 'thermometer' scale, drop-down list), and visualisations for presenting analysis results (tables, forest/interval plots for qualitative or partially quantitative analyses; 'Difference display' (MCDA), and stacked or grouped bar charts for quantitative analyses).

IV. Exploration: This stage assesses the robustness and sensitivity of the main results to various assumptions and sources of uncertainties, considers impact or added value of risk minimisation measures, and likely requires both statistical and clinical input.

Useful methodologies include ITC/MTC, utility survey techniques (DCE, AHP, Swing-weighting, MACBETH), PSM, and SMAA. Preferred visualisation techniques include the box, distribution, scatter, and forest/interval plots; tornado diagram; and most importantly, techniques that are interactive with the user.

V. Conclusion and Dissemination: This is the point at which, after considering all the information in the previous four stages, a conclusion is reached. The results and consensus from the benefit-risk assessment are then explicitly communicated to a wider audience, providing a transparent audit trail of the whole assessment process and bringing all aspects together in a holistic fashion. The content of the communication and visualisation methods used should match the needs of the intended audience.

While no single benefit-risk methodology can fully capture all aspects of a benefit-risk assessment, the choice of a single approach or combination of methodologies should be matched to the complexity of the problem. Application of a simple descriptive framework can provide a clear and easily communicable benefit-risk assessment, could be sufficient for the majority of benefit-risk problems, and can be enhanced for clarity with varying degrees of quantification. For more complex problems, a framework supplemented by quantitative models can facilitate consideration of trade-offs amongst the benefits and risks, address uncertainty, and potentially lead to a more comprehensive overall assessment. To understand the perspective of a particular stakeholder, elicitation of preference values for weighing benefits and risks may be required.

Several ongoing initiatives are examining the role of formal benefit-risk assessment methods, and PROTECT's experience makes what we believe is a unique contribution which complements and builds on these other efforts. The recommendations provided should serve as a valuable guide for readers who are new to the world of benefit-risk assessment as they highlight key issues and considerations that are common to many approaches.

# Contents

# List of Figures

# Section 1   Introduction

## Part 1:  Background: IMI-PROTECT and WP5

Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) is a project set up under the Innovative Medicines initiative (IMI). Its goal is to strengthen the monitoring of the benefit-risk balance of medicines in Europe. This will be achieved by:

- Developing a set of innovative tools and methods for:
  - early detection and assessment of adverse drug reactions from different data sources, and
  - integration and presentation of data on benefits and risks.

- Applying these methods to real-life situations in order to provide all stakeholders (patients, prescribers, public health authorities, regulators, and pharmaceutical companies) with accurate and useful information supporting risk management and continuous benefit-risk assessment.

PROTECT is a collaboration amongst 33 private and public sector partners.  It receives funding from the Innovative Medicines Initiative (IMI) and the European Federation of Pharmaceutical Industry Association (EFPIA), and is coordinated by the European Medicines Agency (EMA).  For more information, refer to: http://www.imi-protect.eu/.

One of the PROTECT working groups is Work Package 5 (PROTECT WP5), whose focus is on the integration and visual presentation of benefit and risk data. PROTECT WP5 aims to provide practical recommendations for benefit-risk decision processes and supporting tools to various stakeholders, particularly the regulators. PROTECT WP5 advocates for increased transparency and robust decision making by making explicit and effectively communicating the methodologies, assumptions, and outcomes utilised in the assessment of benefit-risk in medicine.

PROTECT WP5 executed a review of currently available Benefit-Risk Methodologies, consisting of a comprehensive literature review of various methodologies for conducting benefit-risk assessment, and made recommendations on 13 potentially competent approaches for the assessment of the benefit-risk balance in medicine (Mt-Isa et al., 2012; also refer to Appendix 3).  These 13 recommendations were based on the criteria set out in Part 5 of this Introduction.

Four case studies tested some of the approaches during the first wave (Wave 1) of application in historical real-life drug decision making problems. These were followed by four further case studies in Wave 2, two of which were extensions from Wave 1.  A summary of the eight case studies is presented in Table 1.  The criteria for selecting the case studies were as follows:

- **Wave 1 Case Study Criteria**
  - Scenario is within the last 5 years.
  - There is uncertainty around benefits and risks OR nature of disease (not both).
  - Sufficient reliable data are available.

- Same timeframe is used for both benefits and risks.
- Benefit-risk assessment may be applied at the population, individual, and/or sub group level.
- There is no restriction on age of drug.
- Data on comparators used must be drawn from comparable populations at the same stage of disease.
- There is no restriction on indications.

- **Wave 2 Case Study Selection Criteria**
  - Scenario can be any time.
  - There may be uncertainty around benefits, risks, and the nature of the disease.
  - Multiple data sources of variable reliability may be used.
  - Benefits and risks may have different timeframes.
  - Benefit-risk assessment may be applied at the population, individual, and/or sub group level.
  - There is no restriction on age of drug.
  - Data on comparators does not have to be drawn from comparable populations or at the same stage of a disease.
  - There is no restriction on indications.

All of these case studies are historical examples where previous regulatory assessments have been made leading to decisions on their Benefit-Risk balance. Therefore, it is important to stress that the processes described and conclusions drawn from the work presented in these case studies 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 or the EMA.

**Table 1 PROTECT WP5's case studies**

|  | Case study | Questions addressed | Perspectives | Notes |
|---|---|---|---|---|
| WAVE 1 | Efalizumab | Given the emergence in the post-marketing setting of PML (Progressive Multifocal Leucoencephalopathy) in addition to other serious risks (cardiotoxicity, neurotoxicity, serious infections including tuberculosis), are there in January 2009 any risk minimisation measures which could be rapidly implemented, thus maintaining the benefit-risk balance of the drug as positive? If not, should the Market Authorisation be suspended / revoked? | Regulator | Initial Market authorisation was controversial; later withdrawn due to serious safety concerns. |
|  | Telithromycin | Based on data available from the EMA/CHMP EPAR product information and scientific discussion, 2007: | Regulator | Key risks emerged in post-marketing setting, resulting in labelling |

| | Case study | Questions addressed | Perspectives | Notes |
|---|---|---|---|---|
| | | 1. Should telithromycin be given marketing approval at the time of first registration?<br>2. Is FDA justified in removing the indications ABS (acute bacterial sinusitis) and AECB (acute exacerbation of chronic bronchitis) from the labelling in 2007? | | changes relating to populations appropriate for use and increased risk of muscle injury due to interaction with some statins. |
| | Natalizumab | Given the evidence of PML (Progressive Multifocal Leucoencephalopathy),<br>1. Should natalizumab be given marketing approval at the time of first registration given the evidence of PML in clinical trials?<br>2. Should natalizumab be kept on the market given that increased episodes of PML were observed in the post-marketing setting? | Regulator, using the values and weights of patients | An effective treatment for a serious disease, with a rare but very serious side effect. Withdrawn from market but then reintroduced including consideration of patient perspective. |
| | Rimonabant | 1. To compare different benefit-risk methods using rimonabant as a test case<br>2. To evaluate benefit-risk profile of rimonabant based on data available during submission and around withdrawal period | Regulator, physician, layman | Risk of psychiatric disorders emerged post-marketing, resulting first in label changes and then voluntary withdrawal. |
| WAVE 2 | Natalizumab | To apply visualisations and probabilistic uncertainty modelling to the problems addressed in the Wave 1 natalizumab case study. | Regulator, using patient values and weights | An extension of the Wave 1 natalizumab case study to address more complex issues relating to uncertainty in benefit-risk data and novel visualisation techniques. |
| | Rimonabant | To apply visualisations, Stochastic Multi-criteria Acceptability Analysis (SMAA), and elicitation of patient preferences to the problems addressed in the Wave 1 rimonabant case study. | Patient | An extension of the Wave 1 rimonabant case study to address more complex issues relating to uncertainty, preference values, and novel visualisation techniques. |
| | Rosiglitazone | To test the suitability of a group-oriented approach to developing a quantitative benefit-risk model and to explore the value to incorporating uncertainty about all the effects using probabilistic simulation. | Regulator | Adverse effects on cardiac function recognised. Labelling restrictions in USA; market authorisation suspended in EU. |
| | Warfarin | To assess the benefit-risk balance of warfarin for the treatment of atrial fibrillation, examining in particular changes in data and comparators over time, and to use individual patient demographics to predict the variability of the benefit-risk balance. | General | To investigate the difficulties that may be encountered with undertaking a benefit-risk assessment with visualisation for a mature product with well-established use and accepted safety profile. |

In addition to the benefit-risk methodologies review of information, a second review looking at the available methods for visual representations of benefit and risk was conducted. This was a two stage process and conducted in parallel with the case studies to identify and assess the common types of graphics used to present benefits and risks. The Stage 1 Visualisation Review (Mt-Isa et al., 2013a) performed a systematic appraisal on the suitability of graphics encountered in the PROTECT Methodology Review (Mt-Isa et al., 2012) using previously published criteria. However, PROTECT WP5 quickly learned that the breadth of visual representation technologies in the Stage 1 review was limited due to its narrow scope so they aimed to remedy this issue in the second stage of the Visualisation Review (Mt-Isa et al., 2013b) by (a) conducting a formal literature review in the area of benefit-risk visualisation to capture other innovative visual display technologies (e.g., dashboards, visualisation creation technologies), and (b) working closely with the Wave 2 case studies, which offered an opportunity to test/apply additional visualisation methods beyond those used in the Wave 1 case studies.

Additionally, the need to gather stakeholders' perspective on weighing benefit and risk for appropriate quantitative modelling and a strong interest in patient and public involvement (PPI) from the PROTECT case study task forces motivated the creation of the PPI workstream. This workstream is currently working to develop a toolbox for those who wish to incorporate PPI into medical benefit-risk decision making (such as when eliciting preference information or communicating the results of the decision making process).

## Part 2: Scope and target audience

The evaluation of the balance between benefits and risks of drugs is fundamental to all stakeholders involved in the development, registration, and use of drugs including patients, health care providers, regulators, payers, and pharmaceutical companies.

The scope of this Recommendations document is to distil the experience of the case study teams into a clear set of recommendations for those interested in carrying out or reviewing a benefit-risk analysis. This will be done by leading the reader through the benefit-risk assessment process and, at each stage, discussing the applicable methodologies and visualisations that were reviewed by PROTECT WP5, providing examples of how they were tested in the case studies, and making recommendations regarding their use.

A more formal, systematic approach to benefit-risk assessment improves the transparency of the process. This, in turn, facilitates the involvement of key stakeholders in decisions that affect them at the various stages of a medicinal product's life cycle.

It must be emphasised that benefit-risk decisions are subjective in nature. The reason for this is that moving from information about safety and efficacy to an assessment of benefit-risk balance requires judgement about the relevance of the information to the decision or recommendation to be made. Thus, the "right answer" to a

benefit-risk problem, if it even exists, depends on the perspective that is adopted for the decision and the processes used to reach that decision.

Moreover, it should be reminded that a benefit-risk assessment never comes "ex nihilo." In addition to the available evidence, it takes into account a larger context, including the nature and severity of the condition the drug is intended to treat or to prevent, the benefits and risks of alternative therapies for the condition, and any risk management tool which can positively influence the benefit-risk balance of the drug.

Therefore, none of the methodologies reviewed by PROTECT WP5 should be treated as a blind mechanistic process for making decisions; rather, they can merely present supporting evidence to elucidate the decision making process.

## Part 3: Classification of methodologies

Forty-seven approaches were identified during the PROTECT WP5 Methodology Review (Mt-Isa et al., 2012) (Figure 1) and grouped into the following categories:

- *frameworks* which are stepwise structured approaches;
- *metrics* which are measures for benefits and risks (usually endpoint-specific);
- *estimation techniques* such as simulation techniques and meta-analysis; and
- *utility survey techniques* to elicit stakeholders' preferences (utilities)

*Frameworks* were further divided into those that are *descriptive frameworks* (qualitative or semi-quantitative) and those that also provide comprehensive *quantitative* trade-off approaches. *Metrics* were subdivided into the *threshold indices*, the *health indices*, and those that explicitly allow *trade-offs*.

**Figure 1 Classification of benefit-risk methodologies**



| | | |
|---|---|---|
| All B-R assessment approaches | | Approaches excluded and not appraised |

**Benefit-risk assessment framework**

**Metric indices for B-R assessment**

**Estimation techniques**

**Utility survey technique**

**Descriptive framework**

**Quantitative framework**

**Threshold indices**

**Health indices**

**Trade-off indices**

| Descriptive framework | Quantitative framework | Threshold indices | Health indices | Trade-off indices | Estimation techniques | Utility survey technique |
|---|---|---|---|---|---|---|
| PROACT-URL | BLRA | NNT | QALY | UT-NNT | DAGs | SPM |
| ASF | NCB | NNH | DALY | INHB | PSM | CV |
| BRAT | Decision tree | AE-NNT | HALE | BRR | CPM | CA |
| FDA BRF | MDP | RV-NNH | Q-TWiST | GBR | ITC | DCE |
| CMR-CASS | MCDA | Impact numbers | | Principle of 3 | MTC | |
| COBRA | SMAA | MCE | | TURBO | CDS | |
| SABRE | SBRAM | RV-MCE | | Beckmann | | |
| UMBRA | CUI | MAR | | | | |
| | DI | NEAR | | | | |

Semi – quantitative

Legend:

| Main categories |
|---|

| Sub-categories |
|---|

# Part 4: The need for combinations of methodologies

In modelling the six drugs considered in PROTECT WP5, we found that no single methodology listed in Figure 1 could fully capture all aspects of a benefit-risk assessment. We recognise that a fully-comprehensive analysis may not be required for a decision at a particular stage in a drug's life. Sometimes, the application of a non-quantitative framework would be sufficient to develop a clear and easily communicable benefit-risk assessment. At other times, supplementing the framework with some quantification would add clarity. For drugs presenting many effects, a framework supplemented by quantitative models can facilitate consideration of trade-offs amongst the effects and deal with uncertainty, leading to a more comprehensive benefit-risk assessment. In short, depending on what makes a case complex, multiple methodologies might usefully be applied.

The PROTECT WP5 Methodology Review (Mt-Isa et al., 2012) argued that *descriptive frameworks* do not necessarily perform an integrated benefit-risk assessment per se but do frame the decision problems through a structured approach so as to ensure a better definition of the decision context and transparency in its communication. A descriptive framework has the capacity to bring in quantitative approaches toward prioritising identified favourable

and unfavourable effects, weighing up benefits against risks* or assessing uncertainty more in depth, but it should be noted that the balancing, or trading off, of benefit versus risk is often taken outside of the actual descriptive benefit-risk framework, using the evidence in the framework to inform that assessment.

> * It should be clarified that PROTECT WP5 used the terminologies "favourable" and "unfavourable" effects in a different meaning than "benefit" and "risk": this is meant to express the difference between objective data about effects, using the agreed 'favourable effects' and 'unfavourable effects' that has been adopted by the EMA, as distinct from 'benefits' and 'risks', which include judgement about the clinical relevance of the effects.

On the other hand, *metric indices* do not properly frame the decision problems but do provide quantitative measurements for the effects and outcomes. Within *metric indices*, only the *trade-off metric indices* allow benefits and risks to be traded off, whilst the other sub-categories, *threshold indices* and *health indices,* may take benefits and risks into consideration but do not allow trade-offs. The *quantitative frameworks* are designed to do both in terms of framing the decision problems to some extent (but the decision context is not as explicitly structured as in the descriptive frameworks) and then progressing to provide integrated measures of the benefit-risk balance. It is clear that an integrated measure of a benefit-risk balance is not necessarily preferred or valued in every instance of a benefit-risk assessment.

Adopting a framework, therefore, plus or minus additional methodologies, allows to: a) enhance transparency, b) address prioritisation, c) look at weighing up benefits and risks, d) generate an integrated measure of benefit-risk, or e) further analyse uncertainty.  The choice of that framework should be determined by the complexity of the question at hand.  If a less technical approach is not answering the question adequately, a more complex method may be necessary, and additional resources and expertise may be needed as well.

The timeframe in which a decision is required may necessitate a staged approach in which the more straightforward aspects are answered first and additional analysis of the complex issues follow.

The more specific consideration for incorporating multiple methodology approaches in a benefit-risk assessment is the evidence side of decision problems.  For medical benefit-risk decisions, there are usually two forms of evidence to consider: *data* that aim to provide objective estimates of the effects of drugs (and may be drawn from, e.g., clinical trials, observational studies, registries, spontaneous reporting databases, literature, preclinical information, or clinical judgements) and the *preference values* of the key stakeholders.

Although simple benefit-risk assessment models may be appropriate in straightforward assessments, such simple models may impose implausible assumptions on the evidence and thus become inappropriate in more complex situations, such as when the benefit-risk balance is unclear, when there are many conflicting benefits and risks, or when the risk tolerance/benefit acceptance varies greatly between stakeholders. In these cases, complex evidence synthesis and simulation models (*estimation techniques*) may be more appropriate.

Preference values may require formal elicitations from relevant stakeholders; and this can be achieved through standard activities within certain *quantitative frameworks* by taking the specific stakeholder perspective for a

qualitative framework and by utilising the *utility survey techniques* investigated in PROTECT WP5. Preference values may also affect the results as was noted in the Wave 1 rimonabant case study where the subgroup demonstrated that the benefit-risk balance differs when the data were analysed using different stakeholder preferences (weights).

The application of the different approaches is not necessarily exclusive; several approaches can be used in combination or in parallel. The combination of key features of multiple approaches may facilitate a more robust benefit-risk assessment decision, in that the benefit-risk question is viewed from multiple methodological perspectives, and may result in aspects of the problem being considered that would otherwise have been overlooked. Additionally, many approaches are related to one another as many have been derived based on the idea of another. Utilising a combination of approaches does not necessarily mean that the different approaches are used in whole or in sequence. It could just be that certain features or proposals of an approach that may be lacking in another are used to come to a better decision model. A simple example of this is displaying uncertainty rather than using a more deterministic method (some examples are found in the Exploration section).

The choice and complexity of the approach (or approach combination) should be appropriate to the complexity of the benefit-risk assessment to be executed relative to the problem to be solved.

The authors recognise that not all readers may agree with the conclusions reached by the PROTECT WP5 Methodology Review (Mt-Isa et al., 2012) regarding the capacities of the various approaches.

Multiple combinations of methodologies were employed in the PROTECT WP5 case studies and are shown in Table 2.

**Table 2 PROTECT WP5 case studies: methodologies used**

|  | Case study | Methodologies used | Rationale for the combination of methods |
|---|---|---|---|
| WAVE 1 | Efalizumab | BRAT<br>PrOACT-URL<br>MCDA<br>BRR | Test the most available existing qualitative frameworks. Both were used to ensure comparison. In addition, BRAT was completed with an integrative simple Metric for trade-off between one major Benefit and one major Risk of the drug (BRR). PrOACT-URL was completed with a quantitative method integrating swing weighing of multiple criteria and using heterogenic data from RCTs and post-marketing (MCDA) |
|  | Telithromycin | BRAT<br>PrOACT-URL<br>MCDA<br>SMAA<br>SBRAM<br>BRR<br>PSM | PrOACT-URL formed a base for most of the comprehensive frameworks and was used to prepare the case study. Both BRAT and PrOACT-URL were evaluated to ensure transparency, to structure a benefit-risk decision problem, and to allow for comparisons between the two frameworks. MCDA provided a comprehensive approach to enable the benefit-risk balance to be represented numerically by incorporating the weighted value or utilities of favourable and unfavourable effects. SMAA was used to build a model by using a distribution of data rather than a |

| | Case study | Methodologies used | Rationale for the combination of methods |
|---|---|---|---|
| | | | single value as in MCDA. PSM (via Monte Carlo Simulation) was used to explore the statistical uncertainty in the benefit-risk balance obtained from the BRR metric. SBRAM is similar to BRAT, PrOACT-URL, and MCDA in its stepwise approach to structure the decision process, and was used to allow for comparison between the frameworks. |
| | Natalizumab | BRAT<br>PrOACT-URL<br>ITC<br>NNT/NNH<br>MCDA | Both qualitative frameworks are suitable and were used so they could be compared. ITC was used because active arms from studies cannot reasonably be compared because of differences in patient populations between studies. ITC "calibrated" treatments to a common placebo population. NNT and NNH is a popular method which is relatively simple to apply. It would be insightful to assess how it compares to more principled methods for assessing benefit-risk. MCDA is one of the most general methods. |
| | Rimonabant | BRAT<br>PrOACT-URL<br>NNT/NNH<br>PSM<br>Impact numbers<br>BRR<br>MCDA<br>SMAA | Both BRAT and PrOACT-URL were evaluated to ensure transparency, to structure a benefit-risk decision problem, and to allow for comparisons between the two frameworks. NNT/NNH, Impact Numbers, and BRR were tested mainly because of their simplicity, and PSM was also included because it allows a more complex benefit-risk model to be constructed, taking into account various uncertainties in input values. MCDA was tested because it provides a comprehensive approach to integrating and assessing benefit-risk balance. SMAA was regarded as an extension to MCDA with the added simulation. |
| WAVE 2 | Natalizumab | BRAT<br>ITC<br>PSM<br>MCDA | BRAT was used as in Wave 1 since we found little difference between the qualitative frameworks. ITC was used because active arms from studies cannot reasonably be compared because of differences in patient populations between studies. ITC "calibrated" treatments to a common placebo population. PSM was used because we could find the uncertainty in the clinical outcomes at the population level, and we wanted to assess how this affected the benefit-risk balance. MCDA is one of the most general methods. |
| | Rimonabant | PrOACT-URL<br>MCDA<br>SMAA<br>ITC/MTC<br>DCE | Results from Wave 1 demonstrated that BRAT and PrOACT-URL were observed to be very similar. PrOACT-URL was used because it is more complete. MCDA and SMAA were considered the most suitable benefit-risk assessment methods when compared against all those evaluated due to their ability to synthesise benefits and risks within the regulatory context. The main advantage of SMAA over MCDA was that SMAA allowed for |

| | Case study | Methodologies used | Rationale for the combination of methods |
|---|---|---|---|
| | | | flexibility in the uncertainty regarding the data range and criteria weight information. This was the reason why SMAA was used in Wave 2. ITC/MTC was used for the synthesis of data from different sources (studies), and DCE was used to elicit preference values from lay people. |
| | Rosiglitazone | PrOACT-URL<br>PSM<br>MCDA | Applying the PrOACT-URL framework in a facilitated group workshop enabled the construction of a deterministic MCDA model that clarified why this drug was so controversial. Another workshop in Wave 2, by incorporating probability distributions about the data of all the effects and using Monte Carlo simulation to explore many scenarios about the effects, provided even more certainty about the benefit-risk balance than could be seen in the deterministic model. |
| | Warfarin | BRAT<br>PSM<br>SMAA<br>Individualised benefit-harm method | Applying the BRAT framework helped to identify the key favourable and unfavourable effects for this older medicine and then to visualise those effects. With this as the basis, PSM and SMAA allowed the group to compare the impact of increased uncertainty on the benefit-risk assessment (e.g., due to the lower quality of older clinical trial data). In view of the availability of data in CPRD, an individualised benefit-risk approach was also taken to see whether incorporating information on the identified benefits and risks from the earlier models with the individual patient data affected the interpretation of the benefit-risk assessment. |

## Part 5: Criteria for appraisal of methodologies

During the various stages of PROTECT WP5, criteria were selected to identify the most effective benefit-risk assessment methodologies and visualisations.

In the Methodology Review (Mt-Isa et al., 2012), benefit-risk assessment methodologies were assessed in four dimensions:

(1) fundamental principle,
(2) its features,
(3) associated visual representations, and
(4) assessability and accessibility.

Further details on these dimensions and the specific criteria used are provided in Appendix 1.

The Discussion section of the case study reports were structured as follows:

(1) appropriate frame,
(2) meaningful reliable information,
(3) clear values and trade-offs,
(4) logically correct reasoning, and
(5) commitment to action.

Further details on these dimensions and the specific criteria used are provided in Appendix 2.

Visualisation techniques in the Stage 1 Visualisation Review (Mt-Isa et al., 2013a) were appraised using:

(1) the Wickens' principles of display design (Wickens et al., 2004) which had been adapted to benefit-risk assessment,
(2) the Lipkus' five elements of visual communication (Lipkus and Hollands, 1999), and
(3) Carswell's taxonomy (Carswell, 1992).

In the Stage 2 Visualisation Review (Mt-Isa et al., 2013b), which covers broader visualisation aspects, the visualisation techniques were appraised using several key criteria:

(1) the intended audience,
(2) the intended message,
(3) the knowledge required to understand visualisations,
(4) any unintentional message that may be associated with visual types,
(5) any missing information that may be needed to understand the visualisations, and
(6) potentially useful software to reproduce these visuals.

We summarise the most important findings in a format that can be easily digested. More detailed supporting information can be found in the case studies, as well as in the Methodology Review (Mt-Isa et al., 2012) and the Stage 1 and Stage 2 Visualisation Review Reports (Mt-Isa et al., 2013a; Mt-Isa et al., 2013b).


## Part 6: Document structure

This document discusses the various benefit-risk methodologies that have been investigated by PROTECT WP5. These are set out in the following categories, corresponding to key stages of the benefit-risk assessment process; in order to help the reader understand the application of benefit-risk methodologies and visualisations when progressing through a benefit-risk assessment, we have deliberately remained agnostic to the technical terminology used in benefit-risk frameworks described in the literature and detailed under each of these stages:

1) Planning
2) Evidence Gathering and Data Preparation
3) Analysis

4)
5)

Section 2 deals with each of the above benefit-risk assessment stages in turn.  The document thus acts as a practical guide for those interested in undertaking a benefit-risk assessment, proceeding in a broadly linear fashion from the initial stages (Planning) to the conclusion (Conclusion and Dissemination).

However, it is of vital importance that the reader appreciates the overlapping and iterative nature of the key benefit-risk assessment stages.  Those carrying out a benefit-risk assessment will inevitably find that tasks initially tackled at the earlier stages need to be revisited in light of what is revealed later on.  Equally, choices made in the early stages of the benefit-risk assessment process should be guided by knowledge of the data, resources, and methods that will be available further down the line.  In other words, it is important to take a *holistic* view and an iterative approach to the entire benefit-risk assessment process – and to accept that some steps may need to be revisited and refined in an iterative fashion - rather than concentrating on each stage of the process in isolation.

For example, the Wave 2 warfarin case study examined the available clinical data (part of the Evidence Gathering and Data Preparation stage) before deciding on the outcomes to be included in the final decision model (part of the Planning stage).  The Wave 2 natalizumab case study employed a probabilistic uncertainty analysis (part of the Exploration stage) that involved extracting data in a different format to the main Wave 1 analysis (revisiting the Evidence Gathering and Data Preparation stage). Because of the iterative nature of the process, readers who have turned to this document for guidance for a benefit-risk assessment of their own are encouraged to study all parts of Section 2 at the outset prior to embarking on a benefit-risk model.

As a medicinal product matures, so does the body of evidence regarding its effects.  Estimates of the benefit-risk balance are therefore dynamic, and it is natural to revisit benefit-risk assessments many times over the life cycle of a product.  Furthermore, the relationship between evidence and benefit-risk assessment cannot be seen as only one-way.  The design of interventional and observational studies should be guided by both the results of benefit-risk assessments that have already taken place and the needs of those studies that are planned or in progress.  A similar point can be made regarding the elicitation of preference values of patients and other stakeholders.  Some users may have the resources to integrate the collection of clinical data, preference values, and the assessment of benefit-risk into a single overarching process (e.g., pharmaceutical companies carrying out clinical trials).

Each benefit-risk assessment process stage is introduced by simple questions that the reader may raise at that stage, followed by a narrative regarding the findings experienced in the test case studies, concluded by some discrete summarised "recommendations."

## Introduction:  Summary of key points

- Out of an exhaustive review of 47 benefit-risk assessment approaches, PROTECT WP5 proposed a classification into 4 categories and selected 13 potentially relevant methods for application to benefit-risk balance of medicinal products.

- PROTECT WP5 also conducted, in two successive steps, an extensive review of Visualisation Methods intended to communicate benefit-risk assessment results to various audiences.

- Both benefit-risk assessment methods and Visualisation tools were tested in retrospective "real-life" case studies, consisting in the application of several qualitative and quantitative methods to historical examples originating from past years' regulatory decisions made in Europe.

- This Recommendations document proposes a sequential approach to the benefit-risk assessment process into five key stages: Planning; Evidence Gathering and Data Preparation; Analysis; Exploration; and Conclusion and Dissemination.

- These stages are the common thread structuring this report, with the acknowledgement that they overlap and are iterative in nature, requiring the user to take a holistic view and an iterative approach to the entire benefit-risk assessment process.

- All along this document, recommendations are made regarding the choice of methodologies, their possible combination, and the choice of visual tools, depending on the complexity of the problem to be resolved, the audience to be addressed, and the perspective of the decision maker.

- The public and/or patient's perspective is the subject of an ongoing dedicated subteam intended to explore how and where in the benefit-risk assessment process to involve this important stakeholder.

- This four-year research, conducted by a public-private partnership under the auspices of the EU and coordinated by the EMA and Industry representatives, confirms the added value of using more formal and structured approaches to benefit-risk assessment to improve the transparency and communicability of this process.

- None of the methodologies reviewed by PROTECT WP5 should be treated as a blind mechanistic process for making decisions; rather, they can present evidence to support the decision making process.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
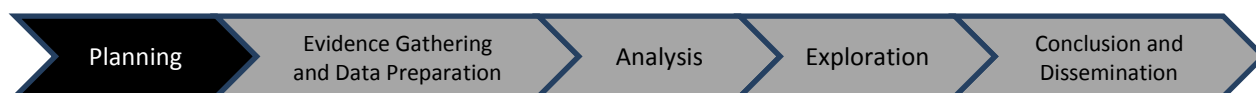
# Section 2   The benefit-risk assessment process

## Part 1:  Planning

In any decision analysis, let alone benefit-risk assessment in medicine, the actual question to be answered by the analysis needs to be clearly specified upfront. Together with good planning, this is required so that stakeholders can agree on the type and thoroughness of the assessment required for their decision maker needs. It also encourages value-focused thinking (Keeney, 1992), which is an approach designed to focus decision makers on the critical issues. This includes well clarified objectives, creative options, the key facts, and the most relevant uncertainties. It thereby helps optimise resource utilisation within an analysis (French et al., 2009).

The benefit-risk assessment process is initiated with planning so that:
- sufficient thinking and thorough discussions between stakeholders are undertaken to clearly define the purpose and context of the benefit-risk assessment;
- these discussions and their results are documented in such a manner that clearly summarises the details to allow evidence to be traced back to its source; and
- future analyses and updates can utilise the foundations of the analysis.

This section addresses the following key questions relating to the Planning stage of a benefit-risk assessment:

What key points should be documented at the Planning stage of a benefit-risk assessment?

What types of methodologies are available to help with the Planning stage?

Which descriptive frameworks were identified and reviewed by PROTECT WP5?

Which descriptive frameworks were evaluated in PROTECT WP5's case studies?

What are PROTECT WP5's recommendations regarding the use of these descriptive frameworks?

What are PROTECT WP5's recommendations regarding the use of visualisations at the Planning stage?

How were value trees constructed for PROTECT WP5's case studies and what lessons were learned?

### What key points should be documented at the Planning stage of a benefit-risk assessment?

The Planning stage is where the decision situation and its context are outlined to provide the starting point for the subsequent data extraction and analysis.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
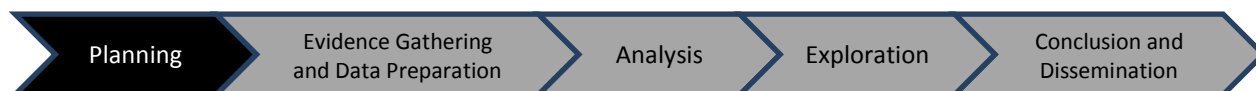
This is done by identifying and documenting the following details that are fundamental to the decision and the evidence supporting the analysis. A more comprehensive description is available within the PrOACT-URL framework (Mt-Isa et al., 2012).

- The **decision problem**
- The **comparators**
- The **benefits and risks** to include
- The **perspectives** that should be taken into account
- The **sources of evidence**
- The **resources** available to the decision maker
- **Time horizon** (short-term versus long-term benefits and risks)

### *The decision problem*

The most fundamental point to address is the nature of the decision for which a benefit-risk assessment is needed. In the case of a medicinal product, the decision usually relates to a specific point in the product's life cycle and may be triggered by a request from the regulatory authorities. For example, the aim may be to decide:

- whether a drug in early-phase trials should undergo further development
- whether a drug in development should be transitioned to the next phase
- whether a drug in development should be continued when a significant safety signal is detected
- whether additional studies or endpoints should be used to evaluate the benefit-risk balance
- whether a request for marketing authorisation should be submitted to regulatory authorities
- whether an optimal Risk Management Plan (RMP) would maintain a positive benefit-risk assessment
- whether an emerging safety risk in the post-marketing period has shifted the benefit-risk balance and whether a healthcare provider should continue using it (this was the question underlying several of the PROTECT WP5 case studies)
- whether another treatment alternative could replace an established treatment based on its benefit-risk assessment (this was considered in the Wave 2 warfarin case study)
- planned safety reports such as Periodic Safety Reports (e.g., Periodic Benefit-Risk Evaluation Reports (PBRERs) and Development Safety Update Reports (DSURs))

Being clear about the purpose of a benefit-risk assessment helps to ensure that appropriate **comparators**, **benefits and risks**, **perspectives**, and **sources of evidence** are adopted and used.

It is good practice to thoroughly consider and document the following factors in the decision context during planning: the treatment under investigation, its indication(s) and therapeutic action, unmet medical needs, the

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
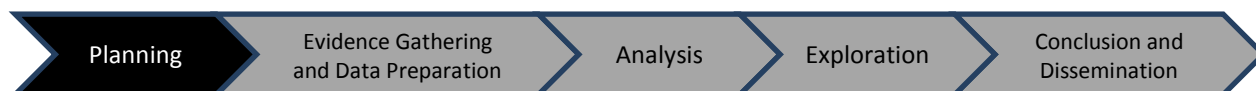
target population(s), and the assumed time horizon for the impact of the benefits and risks. The identity of the decision maker (i.e., the entity ultimately responsible for the decision) should also be mentioned.

## *Comparators*

Benefit-risk assessment is comparative in nature and usually involves weighing up the benefit-risk balance of a medicinal product compared to one or more alternatives. The assessment may compare a medicine's effects at one dose compared to a different dose, against a placebo, or against another active pharmacological or non-pharmacological treatment. Consideration is often given to standard of care and the impact of no treatment.

The choice of comparator(s) depends on the purpose of the assessment. For example, a drug may pass early phase trials if it appears to have a favourable benefit-risk profile compared to placebo or no treatment. However, later in a product's life cycle, it has been suggested that it may be more appropriate to compare the drug with other available treatments (or treatments that are anticipated to be available in the near future) for the same indication (Wave 2 warfarin case study).

## *Benefits and risks*

The benefit and risk measures to be included in (or excluded from) the analysis must be identified and justified. The choices of measures should reflect the strength of evidence, clinical relevance, generalisability, duration of effect, reversibility of risk, preventability, and public health impact (ICH E2C (R2) Guideline). This is a crucial and often difficult step. The basic consideration is to include only those measures that could affect the benefit-risk balance. For example, minor infections that do not require hospitalisation might not be included, while serious infections that would require hospitalisation could be included. If in doubt, include an effect; it can be deleted later if its inclusion was shown to have no effect on the benefit-risk balance.

Depending on a medicine's life cycle, information on its potential benefits and risks can be found in different sources. Some examples are shown in Table 3 below, and the list is not intended to be exhaustive. It is important for effective quantitative modelling to only include measures of benefit and risk that are relevant to the decision to be made (Phillips, 1986).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
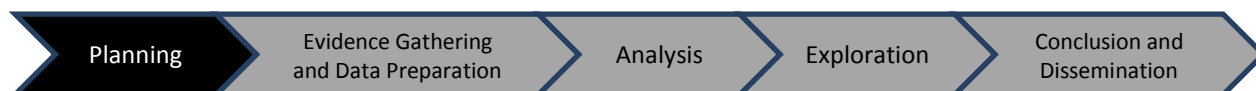
**Table 3 Information sources to identify the types of benefits and risks to include**
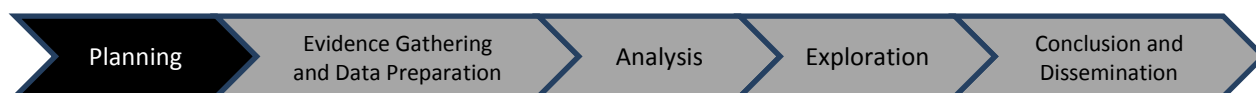
| Timepoint of analysis | Information source |
|---|---|
| Phase II or Phase III trials | • Results of earlier trials<br>• Clinical and pharmacological data |
| Post-marketing | • Post-marketing safety and efficacy studies, including extension studies<br>• Non-interventional (observational) studies<br>• Ad-hoc case reports<br>• Registration documents<br>• Periodic safety reports<br>• Literature searches |

Treatment benefits are normally given precise definitions within clinical studies, whereas risks are often defined using Medical Dictionary for Regulatory Activities (MedDRA) terms, some of which are very broad (e.g., arrhythmia), others relate to more specific conditions (e.g., Wolff Parkinson White syndrome), or by highly defined events (e.g., fatal myocardial infarction).   An analysis team needs to judge which is the level of precision that is required, and to ensure that there is minimal overlap between the definitions and that the data chosen (during the Evidence Gathering and Data Preparation stage) can appropriately reflect these benefit and risk definitions.

As an example for overlap, suppose a decision maker is interested in an individual's risk of ST elevated myocardial infarction (STEMI) in connection with a particular treatment.  A clinical trial has been carried out, recording the number of acute coronary syndrome (ACS) events and STEMIs in a group of patients given the drug.  To use both measures in a decision model would mean *double counting* the STEMIs, giving these events undue weight in the analysis (as they are included in both statistics) (refer to the Wave 1 efalizumab case study for another example of overlap relative to the PASI data).  Also, to use only the number of patients experiencing ACS events may involve *double counting* some patients and thereby overstating the risk, as each patient may have suffered more than one event.  In this example, the analysis team may decide to use the STEMI statistic, but perhaps this figure is not shown in the published report of the trial, so the team may instead choose to use the risk of cardiovascular events as a surrogate to represent STEMIs - but only if there are no other cardiovascular risks to consider, as this may also lead to double counting.

In addition to the above example, some measures (e.g., decrease in bleeding) could appear either as a risk or as a benefit, depending on the state of the patient experiencing them, and therefore need to be evaluated for the potential of double-counting.

As a conclusion, to avoid both double counting and the omission of key measures, it is helpful to take a holistic view across all benefits and risks rather than defining measures in isolation.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

PROTECT WP5's practical experience of identifying benefits and risks is summarised in the section on **constructing value trees**.

## *Perspectives*

A patient considering whether to take a medication and a regulator deciding whether to make a medicine available to a certain patient population may arrive at different preferred choices, even if the options available and the evidence considered are identical.

For example, in the assessment of vaccination decisions, the protection conferred by a sufficient degree of protection in a population (the so-called "herd-immunity" effect) will always be of importance to a public health decision maker, but may carry less weight for an individual patient when assessing whether or not to receive a vaccine.
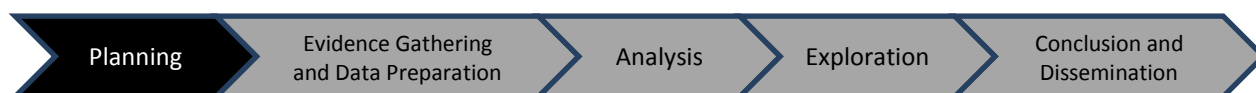
The perspectives should be made explicit at the beginning of the benefit-risk assessment process, and the following questions should be answered:

- who the decision maker is (the first party),
- who the decision is to be made for (the first or the second party), and
- any other stakeholders involved (the third party) in the decision making.

Care should be taken to ensure that second or third party perspectives are not adopted in a way that would interfere with the framing of the decision problem. This is not to say, however, that other perspectives cannot be taken on board. For example, in the **Wave 1 natalizumab case study**, the decision was framed from a regulatory perspective, but clinicians' advice was sought when selecting the benefit and risk outcomes, and patient representatives were involved in assigning preference weights.

Different stakeholders may prefer specific ways of formulating a set of choices and the context around these choices, and assign different weight of importance to the benefits and risks considered.

Different stakeholders may have different views on the benefits and risks criteria to be included in an analysis, and their relative importance. For example, clinicians may consider a particular prognostic biomarker to be a worthwhile endpoint, but from a patient's perspective, it may be impossible to appreciate the meaning and clinical usefulness of this endpoint to their individual situation. Therefore, we recommend to highlight which benefits and risks may be of relevance to different stakeholders and how this impacts the benefit-risk balance. In addition, the analysis should explicitly describe how the benefit-risk balance can differ between subgroups (e.g., infants and children; adults and elderly; male and female; pregnant or not pregnant).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|----------|------------------------------------------|----------|-------------|------------------------------|

Although patient involvement in regulatory decision making is increasingly seen as a priority, regulatory authorities will need to identify resources in order to evaluate patient preferences, and require expertise in the interpretation and incorporation of patient preference values into their regulatory assessment of a medicine.  In an ideal case, regulatory agencies would take into account patient preferences at the same time as considering public health perspectives that reach beyond the interest of an individual patient.

### Sources of evidence

At this stage, it is also helpful to look further ahead and begin considering the type and quantity, as well as summary level or patient level, of data that are available for the treatment of interest and its **comparators**, together with the metrics and quantitative methods that will be used for analysis.  The choice of analysis (which is often guided by the **resources** available) may to some extent determine the data that are required; whether or not these data are available can, in practice, influence many aspects of planning, such as identifying the **benefits and risks**, **comparators**, and **time horizon**. Allowing the use of several sources within one analysis usually carries the risk of bias and errors, and requires modelling assumptions that should be transparently documented, i.e., if pooling results from studies with varying designs.

### Resources

The time and expertise resources available for a benefit-risk analysis should be considered at the Planning stage. Data extraction and analysis may require varying amounts of statistical expertise depending on how they are performed.  This can influence both the methods that are deployed and the practical aspects such as timelines and budgets.

Particularly, extensive and forward planning of resources is needed if data are to be collected directly from patients. For example, the Wave 2 natalizumab case study team found that more time than anticipated was needed to recruit a sample of patients for a planned preference elicitation exercise.

### Time horizon

Another question to consider is how to determine the appropriate time horizon for measuring the occurrence of benefits and risks, and the often far longer (sometimes lifetime) horizon for the impact of these.  For acute effects or short-term treatments without long-term consequences, only a short time horizon may be necessary, but in other cases, it may be desirable to extend the time horizon as far as possible.  In practice, this is often challenging due to the data that are available. In particular, long term follow-up for safety, versus after a blinded efficacy and safety study, creates imbalances in that regard. Where data have been collected over a relatively short study-period for benefits and risks expected to impact the future, modelling may be helpful in understanding the longer-term impact of a medication.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
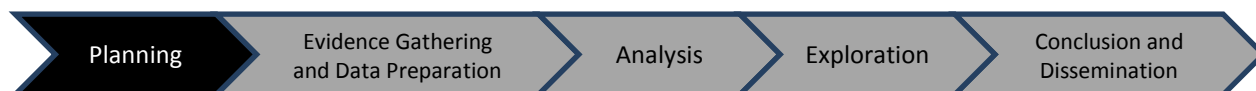
The time horizon should be included as appropriate in the documented definitions of benefits and risks. Different time horizons can be established for different criteria depending on what is considered to be the clinically relevant timeframe for each benefit and risk.

## What types of methodologies are available to help with the Planning stage?

**Descriptive Frameworks** provide a structured, consistent approach to decision making by facilitating the selection, organisation, summarisation, and interpretation of data and preferences relevant to the decision. Frameworks also serve as an aid to decision documentation and communication.
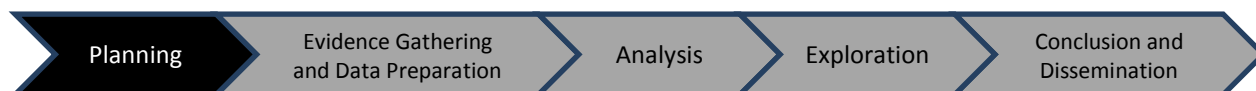
## Which descriptive frameworks were identified and reviewed by PROTECT WP5?

Eight **Descriptive Frameworks** were identified as:

- **ASF**
- **CMR CASS**
- **COBRA**
- **FDA BRF**
- **PhRMA BRAT**
- **PrOACT-URL**
- **SABRE**
- **UMBRA**

As noted in the Methodology Review (Mt-Isa et al., 2012) and in Appendix 3, we are aware of four descriptive frameworks under development. One framework is being developed by the FDA and is known as the FDA Benefit Risk Framework (BRF). FDA BRF aims at giving stakeholders, in this case the regulators, the "big picture" of the issues relevant to regulatory decision making, as well as being compatible with formal quantitative benefit-risk approaches (Jenkins, 2010).

Another framework is being developed by the CMR International Institute for Regulatory Science CASS group (CMR CASS) – Health Canada, Australia's Therapeutic Goods Administration, SwissMedic, and Singapore Health Science Authority. The initial CMR framework on benefit-risk assessment consists of a six-step process (Walker et al., 2009). The CMR CASS group tested the application of a similar framework (Liberti et al., 2010) by omitting the assessment of numerical scores and weights (Phillips et al., 2010). The CMR CASS has further evolved into the Consortium on Benefit Risk Assessment (COBRA) initiative and pursues a more qualitative approach to benefit-risk assessment (CIRS, 2012).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

The Southeast Asia Benefit-Risk Evaluation Initiative (SABRE) is also set up to further share the knowledge and to establish common working grounds between drug regulators in the Southeast Asian region, but there are no details yet available.

COBRA and CASS (as well as PhRMA BRAT) also joined forces with the Unified Methodologies for Benefit-Risk Assessment (UMBRA) Initiative led by the Centre for Innovation in Regulatory Science (CIRS) "to provide a platform for the coordinated development of benefit-risk assessment methodologies that can be used internationally during the drug development and regulatory review and post-approval periods" (http://213.120.141.158/UMBRA). UMBRA aims to increase transparency, predictability, and consistency in the benefit-risk assessment process globally by establishing a consensus on a scientifically acceptable framework for decision making (CIRS, 2012).

It would be too premature at this stage to formally appraise or consider these frameworks for applications in their current form, but they should be considered in the future when more details are available.

Further details are provided on each of these frameworks in Appendices A.5.3 and A.5.4 of the Methodology Review (Mt-Isa et al., 2012).

## Which descriptive frameworks were evaluated in PROTECT WP5's case studies?

Based on the results of the Methodology Review (Mt-Isa et al., 2012) (as noted in Appendix 4), two Descriptive Frameworks, **PrOACT-URL** and **BRAT,** were recommended to be taken forward for benefit-risk assessment methodologies.   The EMA Benefit-Risk Methodology Project Work Package 2 report Table 4 (Phillips et al., 2010) provides a detailed review of both frameworks and their usefulness for benefit-risk assessments.

Created by representatives within the pharmaceutical industry, the stated intention of the **BRAT** framework was to serve as a general platform for benefit-risk assessment, adaptable for use by all stakeholders.  **BRAT** proposed displaying the results as tabular output and graphical summaries, and that benefit-risk evidence should not be integrated but presented separately and individually, allowing the conclusions regarding the benefit-risk balance to be drawn outside of the framework or by the use of additional tools. This was consciously proposed to avoid the creation of single, summary statistics to characterise the overall benefit-risk profile.

**PrOACT-URL** (Hammond et al., 1999) is one of the earliest frameworks for decision making and has a long history in other fields such as operations research and ecological management. However, its application to medical benefit-risk decision making is relatively recent. The EMA Benefit-Risk Methodology Review project (European Medicines Agency, 2013a) has recommended the use of the **PrOACT-URL** framework for benefit-risk assessments in medicine and provided an adjustment of this framework to this context.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
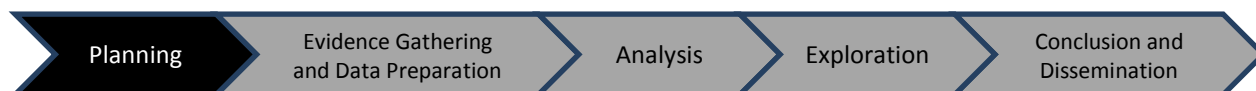
**PrOACT-URL** and **BRAT** both break down the assessment into a stepwise procedure. Such a breakdown is inevitably approximate, with fuzzy boundaries between some of the steps. Also, the process is iterative, and information that is uncovered during an investigation may lead to earlier stages of the process being revisited. Therefore, it would be inefficient to only follow each framework in a strictly linear fashion.

For similar reasons, the steps of the **BRAT** and **PrOACT-URL** frameworks do not map precisely onto one another. They are, however, set out individually below.

### BRAT framework
1) Define the decision context
2) Identify benefit and risk outcomes
3) Identify and extract source data
4) Customise the framework
5) Assess outcome importance
6) Display and interpret key benefit-risk metrics

### PrOACT-URL framework
1) Problem
2) Objectives
3) Alternatives
4) Consequences
5) Trade-offs
6) Uncertainty
7) Risk tolerance
8) Linked decisions

The **PrOACT-URL** and **BRAT** frameworks were used extensively in the PROTECT WP5 case studies, as shown in Table 4.

**Table 4 Descriptive frameworks tested in the PROTECT WP5 case studies**

| | Case study | BRAT | PrOACT-URL |
|---|---|---|---|
| WAVE 1 | **Efalizumab** | ✓ | ✓ |
| | **Telithromycin** | ✓ | ✓ |
| | **Natalizumab** | ✓ | ✓ |
| | **Rimonabant** | ✓ | ✓ |
| WAVE 2 | **Rosiglitazone** | | ✓ |
| | **Warfarin** | ✓ | |

| | Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|---|

| | Case study | BRAT | PrOACT-URL |
|---|---|---|---|
| | **Natalizumab** | | ✓ |
| | **Rimonabant** | | ✓ |
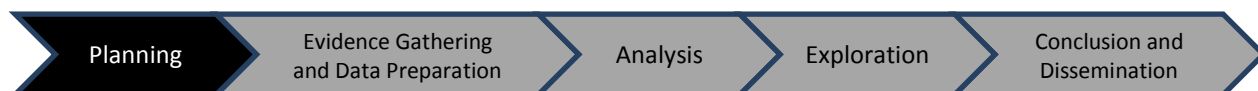
## What are PROTECT WP5's recommendations regarding the use of these descriptive frameworks?

Both the PrOACT-URL and BRAT frameworks emerged as being useful and favourable through the review and application in the case studies, with all the case study teams remarking that the structure provided by these descriptive frameworks served as a useful guide for the benefit-risk analysis process and are considered suitable for use at any stage of a medicinal product's life cycle.

There are two main operational differences between BRAT and PrOACT-URL that were noted in several case studies:

- If trade-offs (i.e., weighing up outcomes against one another) are critical to the benefit-risk balance, then the standard BRAT approach may need to be supplemented by additional methods. For example, in the Wave 1 efalizumab case study, the trade-off between the risk of PML and the benefits of treatment turned out to be pivotal to the results but was not addressed by the BRAT framework. An additional metric index, the benefit-risk ratio (BRR), was used to assess the trade-off. A similar trade-off was crucial in the Wave 1 natalizumab case study, where a weighted net clinical benefit (NCB) approach was used in combination with the BRAT framework.

- Application of the BRAT framework was facilitated by a Microsoft Excel©-based software tool. If populated with data, this could generate the benefit-risk summary table and interval plot visualisation. This software tool has limitations in the amount and type of data and metrics it can process, preferring benefits and risks to be expressed in terms of the relative frequencies of binary events, e.g., the proportions of patients achieving a given level of a favourable effect or of experiencing a defined side effect. However, it is often possible to customise the tool in order to address these limitations. For example, the Wave 2 natalizumab case study expressed a benefit (reduction in relapses) as a rate rather than as a proportion, but a simple mathematical model allowed the transformation of the data into the required form for the generation of the BRAT framework-associated visualisations.

The Wave 1 rimonabant case study team found that the implementation of the BRAT framework was overall easier than PrOACT-URL from the point of view of an inexperienced user, with clearer instructions and a straightforward software tool. However, none of the case study teams found PrOACT-URL particularly difficult to implement.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
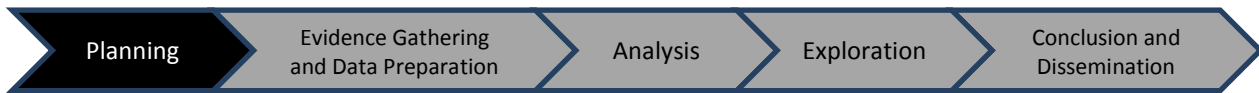
The use of a descriptive framework is not essential but is highly recommended, even for users with little prior experience of benefit-risk assessment. Benefit-risk decisions can be extremely complex, and the use of a framework can help to ensure that key aspects of the process are neither overlooked nor handled inappropriately.

Furthermore, and of particular relevance to the goals of PROTECT, frameworks enhance the transparency of the process and facilitate communication of the results. A structured stepwise analysis of the problem provides a clear audit trail to enable others to understand the reasoning behind a set of findings.  By documenting the interpretation of the available evidence, it allows traceability for reviews and related future assessments of the same medicinal product.

## What are PROTECT WP5's recommendations regarding the use of visualisations at the Planning stage?

When structuring a decision problem at the Planning stage, it is recommended to visually map all benefits and risks that are being considered for the analysis in a hierarchical diagram, called a value tree or attribute tree. This diagram lists and clusters the key benefits and risks that were initially identified. Those measures with particular relevance to the decision maker are focused on in the analysis and should be highlighted within the value tree in order to enhance transparency and the communication of the process, assumptions, and its results.

Various software tools are available to produce a tree diagram, including Microsoft Excel$^©$, Microsoft Word SmartArt Graphics, the BRAT Tool, and FreeMind Software, and are also typically included in MCDA software.

An example of a value tree from the Wave 2 warfarin case study is shown below (Figure 2), and additional samples are available in the other case study reports and in Appendix 6.  First, the benefit and risk categories (e.g., clusters of benefits and risks) are identified, and then specific outcomes within these categories can be determined.

**Figure 2 Value tree for the Wave 2 warfarin case study**

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
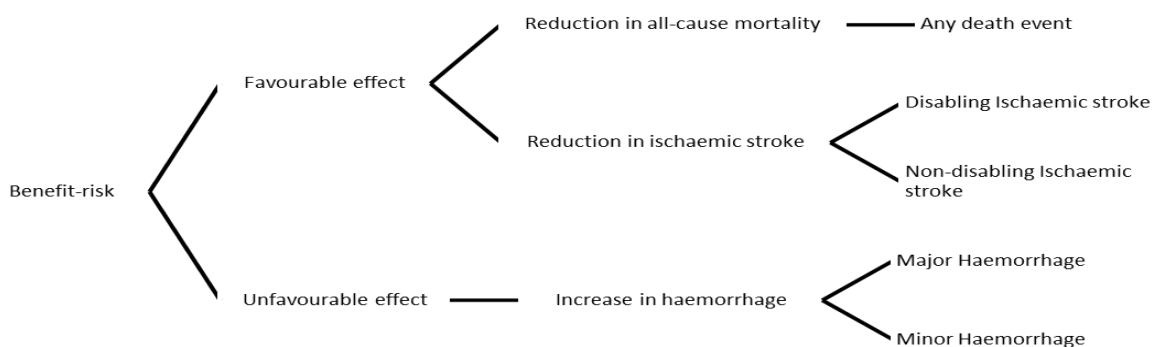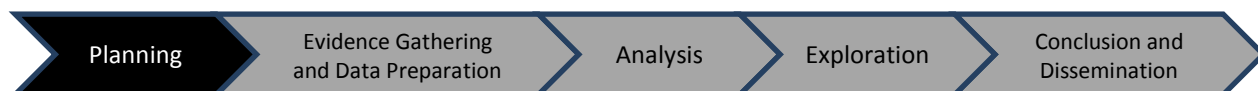
PROTECT WP5 also recommends that a table template (either 'effects table' or 'source data table') is prepared in order to represent the data that are required to be collected.  An example from the Wave 2 rosiglitazone case study is shown in Table 5.  At the Planning Stage, only the structure of the table is created, which constitutes the names and descriptions of all effects, the units characterising the measures of the effects, and the names of the criteria. The data will be filled in later. Again, the effects have been first named loosely (shown here as "Name") and then more precisely ("Description").  This table goes a step further by defining the precise measures and their units.

**Table 5 Structure for an Effects Table based on the Wave 2 rosiglitazone case study**

| | | Name | Description | Units | Rosi + adjunct | Adjunct only |
|---|---|---|---|---|---|---|
| **Favourable Effects** | | Glycaemic efficacy | (A surrogate marker of the quality of glucose regulation.) Mean change from baseline in the proportion of Hb in which A1c is greater than 48 mmol/ml. | % | | |
| | | Micro-vascular events | Incidence of new cases of microvascular events compared to baseline (Retinopathy requiring photocoagulation, vitreous haemorrhage, and fatal or non-fatal renal failure.) | % | | |
| **Unfavourable Effects** | | CHF | Proportion of patients experiencing congestive heart failure during the study period. | % | | |
| | MACE | CV death | The proportion of patients who died from any cardiovascular event including stroke. | % | | |
| | | Non-CV death | The proportion of patients who died from any non-cardiovascular event including stroke. | % | | |
| | | MI | Proportion of patients who experience a non-fatal heart attack. | % | | |
| | | Stroke | Proportion of patients who experience a non-fatal ischemia stroke. | % | | |
| | Other | Weight gain | Mean change from baseline in weight gain at 1 year. | Kg | | |
| | | Macular oedema | Proportion of patients who experience macular oedema. [Are data available?] | % | | |
| | | Bone fractures | Proportion of patients experiencing bone fractures. | % | | |
| | | Bladder cancer | Proportion of patients contracting bladder cancer. | % | | |

The effects table and value tree are two different ways of visualising the set of benefit and risk criteria that will be included in the benefit-risk analysis model.  The value tree in particular has been used frequently within benefit-risk

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

analyses, and the term "value tree" is often used to refer to the criteria themselves when taken as a whole, not just the visualisation.  We adopt this usage for the next section, which discusses the practical aspects of identifying the important benefit and risk criteria.

## How were value trees constructed for PROTECT WP5's case studies and what lessons were learned?

All case studies used a **sequential** approach for building a value tree, starting by creating an initial value tree that displayed all the benefit and risk outcomes of potential relevance to the decision problem. Building on the initial value tree, teams created an additional version of this visual to form the basis of the quantitative analysis.  This process is often called "pruning" and involves excluding the less relevant benefit and risk outcomes to achieve requisite size for a more efficient and insightful analysis. Documenting why benefits or risks were excluded from or included in the equation is key at this later "pruning" stage.

Different **sources of information** were chosen in the case studies on how to identify those benefits and risks that form the basis of the initial value tree, including examination of regulatory documents (the Wave 1 efalizumab case study), searching the clinical literature for relevant trials (the Wave 1 natalizumab case study), or a face-to-face brainstorming session (the Wave 2 warfarin case study).

The **time horizon for measuring the occurrence** of benefits and risks was usually determined by the length of the clinical studies from which data were drawn, or by the length of follow-up time within which the main risks were identified. In the Wave 1 efalizumab case study, benefit was measured over a 12-week study period whilst the risk of PML occurrence within a 3-year time horizon was taken into account.

Importantly, the **time horizon for the expected impact** of benefits and risks will often extend far beyond the time period within which the occurrence of the initial events was identified. In summary, a mixture of timeframes can be used in any analysis, provided that there is no double counting. However, there is no requirement that all criteria should work to the same time horizon provided that the time horizon is specified for each effect.  Thus, short-term benefits could be balanced against long-term side effects.

It was also noted that some attributes (e.g., blood pressure reduction) could in one context be **defined as a risk and in another more appropriately as a benefit**, depending on the health state of the patients receiving the medication. In most value trees, this can be addressed using clear labelling and should not impact the analysis results.

Identification of key benefits and risks within the value tree is an example of how findings from the Evidence Gathering and Data Preparation stage (extraction of clinical data) often feed back to the Planning stage of the benefit-risk assessment.  Hereby, documenting data that could not be found not only provides critical transparency on the process but is also considered an important way of identifying the need for further studies.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

Depending on the type of analysis that is considered possible, it may not be feasible to include particular outcomes within a specific methodology. For example, the Wave 1 rimonabant case study subteam investigating population impact numbers were unable to include the continuous endpoints that were used elsewhere in the case study as impact numbers can only be defined for binary endpoints.

On this issue of feasibility, it is critical to document the appropriateness of a chosen analysis methodology and to ensure the best methodology is chosen in light of all available evidence.

As a general principle, any effects for which a reasonable amount of evidence is available and which could also impact the benefit-risk balance should be included in the analysis. If there is substantial uncertainty on an effect's size, a sensitivity analysis should be performed to examine the potential level of impact on the overall benefit-risk profile of a medicine. Directly assessed preference values may be useful to assess the potential weight of an effect.

Once all relevant benefits and risks have been listed, a critical step is, for each of these, to identify and document **which measurements** from which data source will be used to form the basis for the quantitative analysis. (Refer to the Analysis stage for additional details and to the Wave 2 warfarin case study report Appendix 3 for a description of the iterative process.)

The validity of the analysis results depends on successfully applying to the value tree, and specifically to the measurements, the following criteria[1]:

1. **Completeness**: All criteria that could affect the overall result are included.

2. **Avoidance of double-counting**: If two outcomes are overlapping, or if one acts as a surrogate for another, only one should be included. Although it may not always be possible to fully eliminate double counting (for an example, see Section 5.3 of the Wave 2 warfarin case study), it should always be minimised as far as possible.
   - The Wave 1 natalizumab case study acknowledged that "double counting must be avoided" but also noted that "there was a general concern that the value tree will provide visual imbalance that may transfer into numerical differences later if the group is too selective on either benefits or risks."
   - Criteria that are very similar can be redefined as one criterion or investigated singly at the Exploration stage (e.g., the physician's global assessment of psoriasis and the patient's global assessment may be based on identical criteria; in which case, each should be considered singly in two separate analyses). A patient's response should be counted only once for a given effect; this is frequently violated in the benefit-risk assessment of drugs (e.g., the percentage of patients

---

[1] More detail is given in Section 5.4.4 of Dodgson, et al. (2000).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
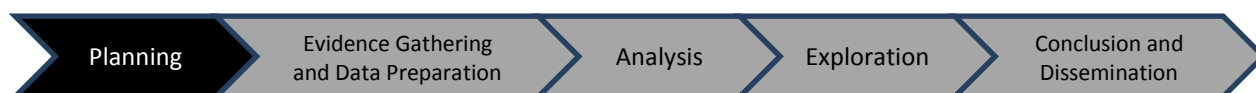
surviving to 24 months always also includes those surviving to 18 months and 12 months; the criteria should be redefined, e.g., as patients surviving up to 12 months, 12 months to 23 months, and 24 months or beyond).

3. **Preference Independence**: The preference order of options on any one effect is unaffected by the preference order on any other effect. This condition is weaker than statistical independence; scores on criteria can be statistically correlated but preference independent. For example, assessors might judge that their preferences for options on a primary endpoint would be the same however the options score on a secondary endpoint, thus establishing that the primary endpoint preferences are preference independent of preferences on the secondary endpoint. However, they might also argue that if performance of an option on a primary endpoint is poor, then they might assign a higher preference for an option on a secondary endpoint; thus, the preferences on the secondary endpoint would not be preference independent of primary endpoint preferences, thus violating the condition of preference independence. This is a requirement if weighted scores are to be interpreted unambiguously for any model approach in which weighted averages are taken.
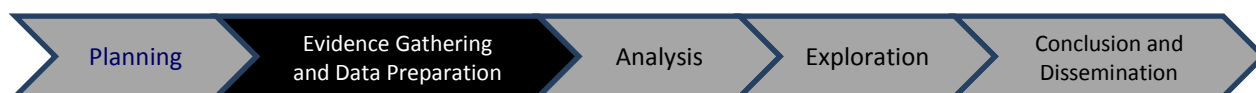
A key learning from the case studies is to ensure that sufficient time is allocated for constructing a value tree (and for the Planning stage in general). Several case study teams found this part of the process took longer than expected. In the case of the Wave 1 rimonabant case study, the team found it difficult in the time allocated to reach a consensus regarding which outcomes to include and which to exclude. This is particularly problematic because, if the problem is framed inappropriately at the Planning stage, it may lead to the need for additional revisions of the analysis later in the process and potentially invalidate the entire analysis.

A final recommendation emerged from looking back over the eight case studies listed in Table 1. All groups initially assumed that the benefits and risks listed in the important source documents were to be taken as given and could be used in the same benefit-risk analysis. However, each task group examined the available literature and found it helpful to **engage several people to advise on the structure of the value tree and the data table**.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

## Planning Stage:  Summary of key points

- It is recommended to agree with stakeholders on the purpose and type of assessment required so that the analysis is efficient and the result valid.

- Appropriate comparators can be a different dose, another active pharmacological or other non-pharmacological treatment, or no treatment.

- Depending on the analysis perspective, different risks and benefits may be included, or different weights of importance assigned. When describing and interpreting the results, any impact of the perspective taken on the benefit-risk balance should be acknowledged.

- Two descriptive frameworks, BRAT and PrOACT-URL, have been found useful to facilitate structured and consistent approaches as they help select, organise, summarise, and interpret data and preferences.

- In any analysis, a hierarchical visualisation, often called "value-tree," should be built to help list and organise all benefits and risks considered. Whilst initially all benefits and risks of potential relevance should be taken into account, an iterative process is recommended to highlight those with limited evidence or impact, and to then focus on those benefits and risks with potential impact on the benefit-risk balance.  A clear record of which effects were included and excluded needs to be kept for transparency.

- A critical recommendation at the Planning stage is to examine possible sources of double-counting, e.g., counting the percentage of patients experiencing an event within 12 months, as well as those experiencing an event within 24 months. Eliminating double-counting can be a challenge but is critical to the validity of an analysis.

- PROTECT WP5 also recommends that a table template ('effects table' or 'source data table') is prepared in order to represent the data that are required to be collected.

- Different time horizons can be established for different benefits and risks depending on what is considered to be the expected duration of each effect.

- Including patient preferences is seen as an important objective by many stakeholders. As the Wave 2 case study on natalizumab demonstrated, extensive and forward planning is required to elicit preferences directly from patients.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Part 2: Evidence Gathering and Data Preparation

This stage of the process concerns the identification and extraction of evidence relevant to the benefit-risk assessment, and data are needed on the performance of each alternative treatment in relation to each of the benefit and risk criteria. When preparing data, it is important to anticipate the methods that will be employed at the Analysis stage of the benefit-risk assessment process, as this may determine the form of the data that are required. PROTECT WP5 recommends that statistical and epidemiological expertise is engaged to ensure that data are handled appropriately. Clinical expertise is required to ensure that appropriate judgements are made regarding the relevance of benefits and risks criteria, and so that risk management options can be identified upfront.

The source data should be clearly documented, together with details of any manipulations that are applied to make the data suitable for the planned decision analysis model. The documentation should be sufficiently clear to enable the data preparation process to be replicated by others.

This section addresses the following key questions relating to evidence gathering and data preparation:

What steps are involved in evidence gathering and data preparation?

What types of methodologies are available to help with evidence gathering and data preparation?

Which estimation techniques were identified and reviewed by PROTECT WP5?

Which estimation techniques were evaluated in PROTECT WP5's case studies?
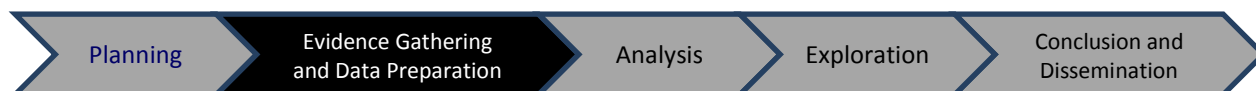
What are PROTECT WP5's recommendations regarding the use of these estimation techniques when gathering evidence and preparing data?

What are PROTECT WP5's recommendations regarding the use of visualisations when gathering evidence and preparing data?

### What steps are involved in evidence gathering and data preparation?

This stage of the benefit-risk assessment process can be further broken down into the following important steps:

- Identifying sources of evidence
- Deciding which source(s) of evidence to use
- Extracting the data for analysis
- Data transformations

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
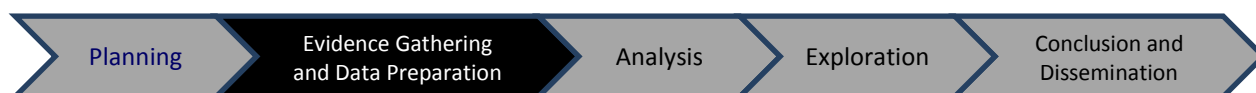
- Aggregating multiple sources of evidence
- Dealing with missing data

### *Identifying sources of evidence*

The availability of evidence regarding a medicine's effects may depend on who is carrying out the benefit-risk assessment, and hence the quantity, form, and quality of the available evidence relative to the point the medicine has reached in its life cycle. The need for evidence regarding the comparators should not be forgotten. PROTECT WP5 was limited to publicly available data for various reasons, including the fact that the case studies were all based on existing historical regulatory decisions and the possibility for subsequent research to be performed based on the same datasets. However, outside of the PROTECT context, decision makers (e.g., Industry or Regulatory Agencies) will have in practice to use both published and internal, or unpublished data. Therefore, adopting an extensive standpoint, typical sources of evidence include:

- Clinical/epidemiological studies: The body of literature on any given drug will accumulate over time, so the use of published data is well suited to benefit-risk assessments carried out during the later phases of the life cycle in addition to existing controlled data. However, Regulators and Industry may sometimes have access to and have to use data on file which are not yet publicly available. Double-blind clinical trials are the gold standard, but observational studies can also play an important role (e.g., the Wave 1 natalizumab case study used estimates of the incidence of PML based on observational data in the post-marketing setting; the Wave 1 efalizumab case study, for the same risk criteria of PML, used only spontaneously reported cases). This exemplifies the fact that, whilst clinical trials are the most efficient way for collecting efficacy data, some post-marketing situations may have to rely on less controlled data (see below Publicly-held Safety databases). The Wave 2 rimonabant case study, a post-marketing study in the UK, was used to estimate the incidence of psychiatric disorders. However, there are some aspects of published studies that can make their use frustrating for benefit-risk assessment purposes:

  o The published endpoints are usually few in number and may not correspond to the benefit and risk criteria that were selected for analysis. For example, in the Wave 2 rimonabant case study, the only endpoint available in the post-marketing study was psychiatric disorders. None of the benefits were analysed in the same publication.

  o The level of detail reported for each outcome may be insufficient for some analyses, particularly for the exploration of uncertainty. As noted by the Wave 2 rosiglitazone case study team: "Authors of publications often report only sample sizes, means, confidence intervals, and significance levels. Whilst this may be sufficient for making statistical inferences, it may not be adequate for the purposes of an MCDA, particularly for sensitivity analyses."

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

- o Data at the patient level (i.e., a record of all the relevant outcome measurements for each individual patient) is rarely made publicly available, but this does not prevent a decision maker (Regulator, Company) having access to non-publicly available individual data to include these in a benefit-risk assessment if this is relevant. This can facilitate exploration of aspects of the problem, such as variability, individualised benefit-risk assessment, and correlation of outcomes.

  - o The number of publications may be small for various reasons. For instance, the time window between approval and withdrawal may be too short to allow many studies to be performed and published. For example, in the Wave 2 rimonabant case study, there was only one published study for each drug being compared (orlistat, rimonabant, and sibutramine).

  - o Some uncontrolled data from post-marketing experience (e.g., spontaneous reports) may be difficult to incorporate statistically, despite providing information on key risks. This was exemplified in some case studies where the major risk was a very rare but very serious condition (e.g., PML in the Wave 1 efalizumab case study). Issues relating to post-marketing data are also discussed in the Conclusion and Dissemination section of this report.

- Public registration documents (e.g., EPARs, PSURs, PADERs, DSURs) are a convenient summary of the data from pivotal studies and were used extensively in PROTECT WP5's case studies.
  - o It may be important to note that EPARs in their current form are insufficient and unfit as a sole data source for benefit-risk assessment. They may also tend to be over-inclusive and generally discursive to be transparent as to which are the important data that inform the benefit-risk assessment and consequently the decisions.
  - o PROTECT WP5's case studies extracted data for analysis from the originally published pivotal studies identified through EPARs where possible.

- Publicly- and privately-held databases can provide extensive data at the patient level, but access may be limited. If a benefit-risk assessor wishes to use a database that is not administered in-house, it is important to consider:
  - o  how access will be obtained
  - o any training or software requirements for the extraction of data
  - o the length of time it would take to obtain access to these databases

- Spontaneous reports of adverse events can provide data relevant to the ongoing benefit-risk assessment of a marketed drug. Despite many limitations such as under-reporting (Belton, 1997), and incomplete numerators and denominators (Clarkson and Choonara, 2002; Sacristan et al., 2001), spontaneous reports are often the only

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
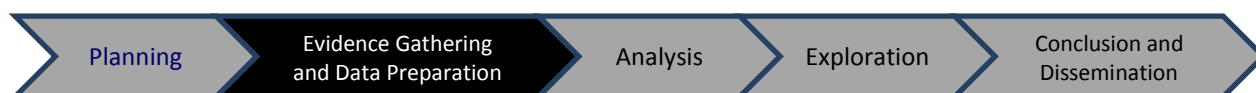
source of information of very rare but serious adverse events, allowing more detailed investigations (e.g., efalizumab PML cases).

*Deciding which source(s) of evidence to use*

In cases where the evidence is sparse, decision makers may have no choice but to use what little data are available. At the other end of the scale, for certain benefit-risk decisions, there may be an overwhelming amount of evidence. For example, in the Wave 2 warfarin case study, which considered a drug with a long history, a systematic literature review identified 37 relevant publications after a manual screening of the full text, but data were extracted from only a subset of these. This is also an example where an assessor may have to face the issue of low quality data, as much of the data on benefits and risks of warfarin pre-dates the GCP era.

Not all of the identified sources of evidence will necessarily be used to inform a decision. Decision makers should aim to ensure that the sources of evidence are appropriate to the decision problem:

- The outcome definitions and time horizon should match as closely as possible the value tree that was prepared at the Planning stage (e.g., in the Wave 1 rimonabant case study, a few trials were not used either because the outcome (smoking cessation) was not close enough to the value tree agreed by team members or because the length of the follow-up was too short).
- The population of patients from which data are drawn should match as closely as possible the population that will be affected by the benefit-risk decision.
- Any evidence whose reliability is doubted on either clinical or statistical grounds based on GCP standards should be excluded or given less weight relative to other more reliable evidence (e.g., in the Wave 1 natalizumab case study, a clinical trial was excluded due to concerns regarding its methodology and small sample size).
- Spontaneous reporting data may need to be used to address emerging risks which may not be available elsewhere; and the reliability of data and uncertainty from this source is to be addressed and documented. There are frequent situations (e.g., this is routine for the PRAC at EMA) where a benefit-risk assessment would have to be reviewed/updated based only on signal detection activity from spontaneous reports.

This is not to say, however, that each parameter in a benefit-risk model should be based on a single source alone. Data from multiple studies can be aggregated on either an impressionistic or statistical basis. This may be done in a way that gives different weights to different sources of evidence according to their perceived reliability (e.g., clinical trials may be given more weight than observational studies as they may be less prone to bias) (GAO-PEMD, 1992). Alternatively, the problem could be assessed using more than one dataset to investigate the effect of including or excluding particular sources of evidence.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

By choosing between multiple sources of evidence, assessors may unwittingly introduce bias into a benefit-risk assessment. Furthermore, published sources of evidence may be unrepresentative of the true clinical effects of a treatment owing to the well-known phenomenon of publication bias (Dickersin et al., 1987; Easterbrook et al., 1991). For these reasons, and to increase the overall transparency of the Evidence Gathering stage, assessors relying on published studies should consider carrying out a systematic review. This would involve specifying *a priori* a literature search protocol that sets out the inclusion and exclusion criteria that will be used together with details of how multiple evidence sources will be aggregated and whether any allowance will be made for publication bias.

The key to a good benefit-risk assessment lies in the representativeness of the data used in the benefit-risk framework / decision model and in the transparency with which the assumptions / decisions for inclusion / exclusion of data and their uncertainties are documented.

Whilst experimental data from clinical trials provide the "efficacy" (effects in perfect conditions) of a treatment as the benefit endpoint, the evidence from clinical databases can provide the evidence of "effectiveness" (effects in real life). Efficacy from a clinical trial may not be observed in real life use of drugs due to many external factors that cannot be controlled (Eichler et al., 2011).

Furthermore, the true benefit-risk balance emerges over time, not as a static "snapshot" as at the time of marketing authorisation applications. Similarly, some risks or adverse events may occur immediately and others may take longer to surface. This brings in the complications of the discordant timing of benefits and risks where it may take longer to observe the benefits or risks of a treatment, and these may not be observed within the trial period. These issues have been known to be the limitations of clinical trials. This is further discussed in the Exploration section.

There are many public clinical databases such as the Clinical Practice Research Datalink (http://www.cprd.com), The QResearch (http://www.qresearch.org), The Health Improvement Network database (http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database), and the Health Informatics Centre (http://www.dundee.ac.uk/hic/) in the United Kingdom alone. Vast amount of data from primary and secondary care are collected routinely in these databases, and can be and have been used for benefit-risk assessment of medicines. Observational data from these databases may be a good alternative but are not free from limitations:

- the larger uncertainty related to the accuracy, bias, and confounding
- the consistency of records varies and is also often questionable
- benefit endpoints that are recorded are limited; more serious events like myocardial infarctions and deaths are likely to be well-recorded, but less serious ones such as recovery from a symptomatic disease may not be recorded at all
- risks or adverse events may be captured through recorded medical complaints
- the severity of adverse events is not generally recorded

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
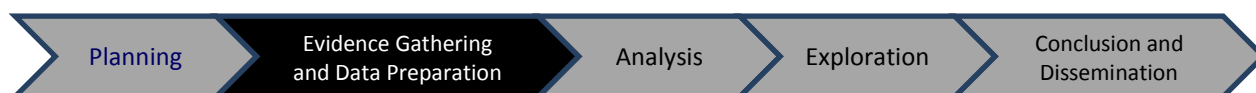
- the timing of the clinical events recorded may not be accurate due to the timing patients actually started experiencing adverse events to the time seeing the physicians

Therefore, when using these types of data sources, close collaboration with experts familiar with the strengths and limitations of the database information is recommended.

Bearing the above points in mind, input should be sought from stakeholders in order to reach agreement as to which evidence is most appropriate as they may not be able to rely on the conclusion of a benefit-risk assessment if they do not trust the evidence on which it is based.

In real life benefit-risk assessments, the circumstances are more complex than just identifying and using suitable evidence for the benefit-risk outcome measures. The process of identifying evidence itself is also important and should be made transparent. Justifications for the strategy utilised to find sources of evidence and evidence selection should also be documented to minimise biases. Future uncertainties and likely scenarios, e.g., as required by the EU Risk Management Plan (RMP) (CHMP/EMA, 2005), should also be considered in the preliminary benefit-risk assessments to accommodate accumulating future data. Addressing hypothetical scenarios, and past and future linked decisions would allow more informed decisions to be made, and consequently, could lead to better minimisation of risks in drugs use.

In the interest of transparency, any data potentially relevant for benefit-risk assessment (according to the final Value Tree) but not actually used in the final benefit-risk assessment for any reason should be precisely described and the reason given for not using it.

### *Extracting the data for analysis*

Once the sources of evidence have been identified, the data must be extracted and expressed in the appropriate metric(s) for the anticipated method of analysis.

If a single evidence source provides all the measures that are needed and in the correct form for analysis, then data extraction is extremely straightforward. In practice, however, the situation is complicated due to the need for:

- data transformations
- aggregating multiple sources of evidence
- dealing with missing data

### *Data transformations*

A data transformation is a mathematical manipulation of data numerical values that changes the scale on which the data are expressed. Transformations are needed when there is a mismatch between the way outcomes are defined or measured in the source data and the definitions and metrics that will be used for analysis. Stakeholders should

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
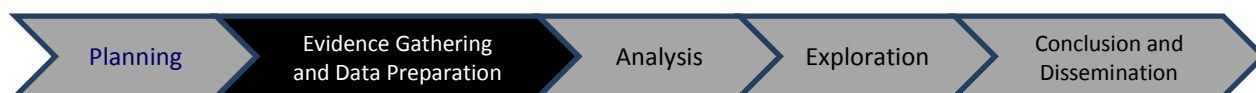
endorse a method of analysis before data preparation can be completed. The need for data transformations may not be obvious to those without statistical expertise.

The Wave 2 warfarin case study provides an example of a data transformation due to a mismatch between the metric in the source data and the metric intended for analysis. The source data (from a published meta-analysis) reported treatment effects as odds ratios. However, as noted in the Wave 2 warfarin case study report: "For the purpose of benefit-risk assessment, odds ratio metric can be quite difficult to interpret, especially when balancing across different criteria with different underlying frequencies." Consequently, odds ratio of the treatment effect was transformed to risk ratio and risk difference.

The Wave 2 natalizumab case study provides an example of a data transformation due to a mismatch in outcome definitions. One of the key benefits was a reduction in the rate of disability progression, where disability progression was defined as deterioration in a standard quality-of-life score, sustained for 6 months. Some of the source data, however, used a different definition, with the disability progression needing to be sustained for only 3 months. As noted in the Wave 2 natalizumab case study report: "The proportion of patients undergoing disability progression was adjusted by a factor of 0.71 in the treatment group and 0.79 in the placebo group to allow for this, based on the results of a trial for fingolimod (another treatment for relapsing-remitting multiple sclerosis), which monitored both definitions of the outcome."

The second example above demonstrates how statistical assumptions often play a role in data transformations – in this case, the assumption was that the ratio between the two outcomes observed in the fingolimod trial would also apply to the other study populations. Statisticians should carefully consider the implications of any assumptions adopted, as inappropriate assumptions may introduce bias to the analysis, particularly if the comparators are not all subject to the same data transformations. The impact of varying the assumptions should later be investigated as part of the Exploration stage.

For the sake of transparency, it is good practice to document any data transformations used. For example, the Wave 1 efalizumab case study report included a table detailing the formulae of the data transformations to be used in the analysis.

## *Aggregating multiple sources of evidence*

There may be no single source of evidence that can provide data on the performance of each comparator in relation to all of the benefits and risks. In such cases, it will be necessary to combine evidence from multiple sources.

Even for outcomes where data are already available from one source, there may be value in supplementing this with data from another source in order to arrive at a combined estimate. If performed appropriately, this makes efficient use of all the available evidence, and results in more accurate estimates and less uncertainty in the data. It also reduces the danger of placing too much reliance on a single source of evidence which may turn out to be flawed.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
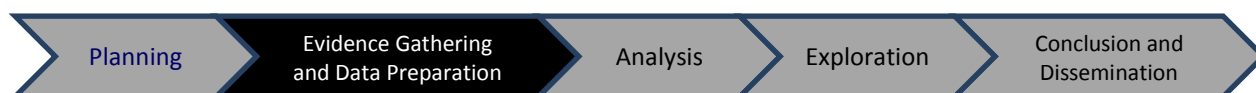
Aggregation of evidence can be done on an impressionistic basis, e.g., by a panel of clinicians considering the evidence and forming an opinion as to the appropriate parameter values (and documenting the rationale), or on a statistical basis using meta-analytical techniques. A technical overview of meta-analysis is beyond the scope of this report, but we later comment on the application of a specific technique, i.e., indirect/mixed treatment comparison. Good examples of data aggregation can be found in the Wave 2 rosiglitazone case study and the Wave 2 warfarin case study.

Aggregating data may not always be appropriate; however, different sources of evidence relating to the same treatment may be heterogeneous with regard to any aspect of study design including (but not limited to) the nature of the study (observational or RCT), target population, indication, disease severity, follow-up period, and dosage. When confronted with heterogeneous data, it may be more appropriate to assess the benefit-risk balance based on each evidence source separately and consider the relevance of any differences in the results.

### *Dealing with missing data*

Usually only known benefits and risks will be included in a decision model. However, if a benefit or risk is deemed important, e.g., a risk that has been observed in other drugs in the same class, but little data have been collected regarding the treatment of interest or no reliable evidence can be accessed, the decision maker:

- should consider using a different outcome definition or a surrogate outcome for which data are more readily available; this means revisiting the Planning stage and ensuring that the benefits and risks form a coherent, non-overlapping set
- may decide it is acceptable to omit the outcome from the analysis if it is not thought to be essential to the benefit-risk decision
- always has the option of leaving the outcome in the model and exploring the sensitivity of the results to different assumed values
- may consider using a different method of analysis that is more compatible with the available data

We would generally recommend the third approach over the second, on the grounds that it increases transparency and avoids second-guessing the outcome of the analysis. Moreover, outcomes that are irrelevant to the benefit-risk decision should not have been included in the first place. Conversely, outcomes which are relevant to a benefit-risk assessment should not be left out only on the excuse that data are unavailable, unreliable, or in a format not supported by the chosen method of analysis.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## What types of methodologies are available to help with the Evidence Gathering and Data Preparation stage?

**Estimation Techniques** may be required in order to transform source data into a suitable form for benefit-risk assessment. Estimation techniques range from the very simple, through to cutting-edge statistical methods for synthesis of complex data from multiple sources.

Well-established estimation techniques such as simple statistical transformations and basic meta-analysis (Sutton et al., 2000) have not been reviewed by PROTECT WP5 owing to their familiarity. However, they are extensively applied in benefit-risk analysis, including in PROTECT WP5's case studies.

## Which estimation techniques were identified and reviewed by PROTECT WP5?

Five **Estimation Techniques** were identified as:

- Directed Acyclic Graphs (DAGs)
- Probabilistic Simulation Method (PSM)
- Confidence Profile Method (CPM)
- Indirect/Mixed Treatment Comparison (ITC/MTC)
- Cross Design Synthesis (CDS)

Further details are provided on each of these estimation techniques in Appendix 7.

## Which estimation techniques were evaluated in PROTECT WP5's case studies?

Based on the appraisal execution results (as noted in Appendix 7), two estimation techniques were recommended to be applied in the case studies:

- Probabilistic Simulation Method (PSM)
- Indirect/Mixed Treatment Comparison (ITC/MTC)

PSM is a general framework for probabilistic estimation, and can be used to characterise the variability or uncertainty in the results of a benefit-risk model. It includes the well-established Monte Carlo method of sampling from statistical distributions a large number of times. PSM can characterise and quantify the uncertainty in the results of an analysis based on known probability distributions for the evidence data and assumptions.

Indirect treatment comparison (ITC) provides a statistical basis for drawing comparisons between treatments that have not been directly compared in the same study population. Mixed treatment comparison (MTC) is a

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

generalisation/extension of ITC that allows integration of direct and indirect evidence. ITC/MTC is concerned with appropriately combining different pieces of evidence in order to warrant the results of an estimation or simulation model.

PSM and ITC/MTC were used in several case studies, as shown in Table 6.

**Table 6 Estimation techniques tested in the PROTECT WP5 case studies**

|  | Case study | PSM | ITC/MTC |
|---|---|---|---|
| WAVE 1 | **Efalizumab** |  |  |
|  | **Telithromycin** | ✓ |  |
|  | **Natalizumab** |  | ✓ |
|  | **Rimonabant** | ✓ |  |
| WAVE 2 | **Rosiglitazone** | ✓ |  |
|  | **Warfarin** | ✓ |  |
|  | **Natalizumab** | ✓ | ✓ |
|  | **Rimonabant** |  | ✓ |

### What are PROTECT WP5's recommendations regarding the use of these estimation techniques when gathering evidence and preparing data?

#### *Probabilistic Simulation Method (PSM)*

Interpreting the results of PSM will be discussed in the Exploration section, whilst this section deals with the extraction of source data for use in such a probabilistic model.

PSM is flexible enough to be used with any model and any combination of benefit and risk outcomes provided the distribution of each uncertain outcome can be characterised when preparing data. Some users may find this is more demanding in terms of resources and statistical expertise than simply extracting average data values. Selecting a statistical model requires careful consideration of the sources of uncertainty that are of interest and how they can be characterised given the available data.

For example, the Wave 2 natalizumab case study used a model that allowed for sampling error in the average outcome values but did not characterise the full variability of responses that would be seen at the patient level. The Wave 1 rimonabant case study, by contrast, did allow for variability of binary outcomes at the population level. These two examples were both based on summary data from published studies.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

Patient-level data, if it is available, is likely to be very useful when attempting to characterise variability at the patient level, particularly if correlations between outcomes are to be taken into account. However, it is frequently unavailable to the public in its entirety, largely as a result of privacy laws that restrict the sharing of data points that could compromise the identity of the individual patient or violate the terms of informed consent. Anonymised patient-level data often can be shared. Out of the case studies, only the Wave 2 warfarin case study had access to patient level data from a routine healthcare database (CPRD) for this purpose. Other case study teams expressed regret at this limitation. For example, the Wave 1 telithromycin case study team noted that in their application of PSM, "criteria are assumed to be independent of each other. It is not clear how much the correlations affect the results." None of the case studies had access to patient-level data from clinical trials, but where such data are available, it may be helpful in revealing and adjusting for any correlations that could affect the results.

Several case studies (telithromycin, natalizumab, rosiglitazone) noted that the Bayesian statistical framework is particularly well suited to deriving probability distributions for benefit and risk outcomes.

### *Indirect/Mixed Treatment Comparison (ITC/MTC)*

ITC/MTC was found to be a valuable tool for bringing together evidence on different comparators. For example, the Wave 2 rimonabant case study report states that ITC/MTC was "extremely useful in estimating the relative response between comparators when direct comparisons are not available."

The Wave 2 rimonabant case study team also noted that the biostatistical assumptions underlying ITC/MTC may not always hold in practice, saying "results were generated under an assumption that the common [pivot] option between different comparison pairs is identical. One would argue that is not always clinically plausible."

In the Wave 2 natalizumab case study, the use of ITC/MTC for the evidence base, and the resulting complex data network, led to complications in the application of PSM. Although distributions for the source parameters were easily defined, the parameters were combined in the ITC/MTC framework under the assumption that they were independent, which may have affected the estimated variability of the decision model parameters.

ITC/MTC is a useful tool when direct comparative data are insufficient or unavailable to support a benefit-risk decision. There are strong arguments to suggest that evidence from direct comparison studies are most credible, but the ITC/MTC technique could still provide a benefit-risk assessor with a more comprehensive picture by amalgamating direct and indirect evidence.

### **What are PROTECT WP5's recommendations regarding the use of visualisations when gathering evidence and preparing data?**

Three visualisation techniques were identified as being useful at this stage:

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
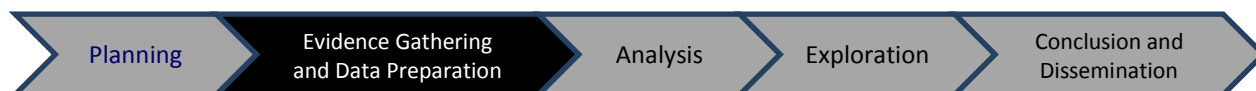
- Structured and colour-coded **tables** of evidence data and sources
- **Network graphs** of the relationship of evidence data and sources
- **Forest plots** of the evidence summary

Data tables can be used to show an overview of numerical values of several measures for different treatments. When setting up a benefit-risk assessment problem, specialised data tables like the 'Effects table' from the PrOACT-URL framework (Figure 3) and the 'Source data table' from the BRAT framework (Figure 4) were generally found to be useful in the case studies.  Such structured tables provide increased consistency and clarity to the decision problem.  Good tables ease cognitive burdens of users and decrease the time required to extract the information.

The amount of information that can be included in a table increases as the benefit-risk analysis process progresses. Initially, tables may simply list the criteria (and perhaps the units of measure) without including any data values (e.g., Table 5).  The table can then be filled in and extended as required – typically, this process would begin with central estimates of the benefit and risk outcomes for each comparator, and perhaps go on to include measures of uncertainty, data sources, transformations, preference weights, or other information that may be required.  When presenting a table for a specific purpose, however, it should be limited to only the relevant rows and columns (i.e., the options and criteria) to avoid adding cognitive burden when reading tables. Readability can be enhanced through the use of colour-coding to represent grouping and relationships.

**Figure 3 PrOACT-URL 'effects table' listing the criteria for the benefit-risk assessment model in the Wave 1 efalizumab case study**

| | Name | Description | Fixed Upper | Fixed Lower | Units | Weight | Efalizumab | Placebo |
|---|---|---|---|---|---|---|---|---|
| Favourable Effects | PASI75 | Percentage of patients achieving 75% reduction in baseline PASI[1] at week 12. | 60.0 | 0.0 | % | 1.0 | 29.5 | 2.7 |
| | PGA | Percentage of patients achieving Physician's Global Assessment[2] clear/almost clear at week12. | 40.0 | 0.0 | % | 0.8 | 295 | 5.1 |
| | OLS | Percentage of patients with Overall Lesion Severity rating of minimal or clear at FT (day 84). | 40.0 | 0.0 | % | 0.25 | 32.1 | 2.9 |
| | DLQI | Dermatology Life Quality Index[3]. Mean percentage of patients showing an improvement. | 10.0 | 0.0 | Change score | 0.8 | 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 | 0.2 | 41.0 | 24.0 |
| | Severe infections | Proportion of patients experiencing infections serious enough to require hospitalisation. | 3.00 | 0.00 | %/100ptyrs | 1.0 | 2.83 | 1.4 |
| | Severe Thrombocytopenia | Number of cases exhibiting severe (grade 3 and above) thrombocytopenia[4]. | 10 | 0 | number | 0.8 | 9 | 0 |
| | Psoriasis Severe Forms | Percentage of patients developing severe forms of psoriasis (erythrodermic, pustular). | 4.0 | 0.0 | % | 0.05 | 3.2 | 1.4 |
| | Hypersensitivity Reactions | Percentage of patients exhibiting hypersensitivity reactions, arthralgia, psoriatic arthritis, flares, back pain asthenia, ALT and Ph. Alk increase. | 10.0 | 0.0 | % | 0.05 | 5.0 | 0 |
| | Intersticial Lung Disease | Number of cases of intersticial lung disease. | 20 | 0 | number | 0.1 | 18 | 0 |
| | Inflammatory Polyradiculopathy | Number of cases of inflammatory polyradiculopathy. | 5 | 0 | Data | 0.02 | 4 | 0 |
| | SAEs | Number of cases of haemolytic anemia. | 25 | 0 | number | 0.12 | 24 | 0 |
| | PML | Number of cases of progressive multifocal leukoencephalopathy. | 5 | 0 | number | 1.0 | 3 | 0 |
| | Aseptic Meningitis | Number of cases of aseptic meningitis. | 30 | 0 | number | 0.1 | 29 | 0 |

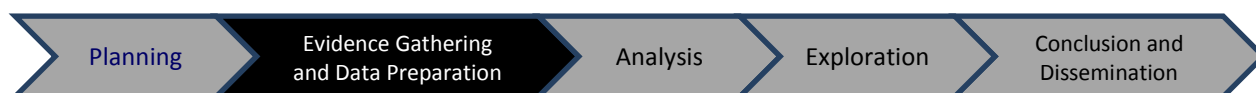| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

**Figure 4 BRAT 'source data table' listing the data for benefit-risk assessment model and their source in the Wave 1 natalizumab case study**

| Study ID | Value Tree Category | Outcome | Measure | Study Drug | Study Drug Estimate | Ref Group | Ref Group Estimate | Study Estimate |
|---|---|---|---|---|---|---|---|---|
| Polman 2006/ EPAR | Disease Activity | Relapse | Annualized Relapse rate[95%CI] | Drug A | 0.23 [0.19-0.28] | Placebo | 0.73 [0.62 –0.87] | 0.32 [0.26 – 0.40] |
| Jacobs 1996 | Disease Activity | Relapse | Annualized Relapse rate[95%CI] | Drug B | 0.67 [n.a.] | Placebo | 0.82 [n.a.] | 0.82 [0.56 – 1.20] |
| Johnson 1998 | Disease Activity | Relapse | Annualized Relapse rate[95%CI] | Drug C | 0.65 [n.a.] | Placebo | 0.91 [n.a.] | 0.71 [0.47 – 1.08] |
| … | | | | | | | | |
| Polman 2006 | Liver Tox | ALT>5x ULN | n/N (%) | Drug A | 31/627 (5%) | Placebo | 12/312 (4%) | RR = 1.25 |
| Jacobs 1996 | Liver Tox | ALT>5x ULN | n/N (%) | Drug B | Not reported | Placebo | Not Reported | RR = 1 |
| Johnson 1998 | Liver Tox | ALT>5x ULN | n/N (%) | Drug C | Not reported | Placebo | Not Reported | RR = 1 |
| … | | | | | | | | |

When gathering data, it is recommended that the table be annotated or colour-coded to highlight any problems with the data, e.g., using a specific colour to indicate cells with missing data. The tree diagram and table produced initially may need to be revised in the light of data quality and availability, and this will serve as a useful visual summary of the issues.

It is also recommended that a version of the data table is prepared with full details of evidence sources, data transformations, and estimation techniques, enabling the data extraction process to be reproduced.

Another visual that may be useful at this stage is the network graph. The network graph can be used to give an overview of the sources of direct and indirect evidence on different treatment effects, which may drive the choice of benefit-risk assessment approach in the Analysis stage. The visual was not applied in any of the case studies because of trivial relationships between data sources. In particular, a network graph may be a useful visualisation in the application of ITC/MTC.  For illustrative purpose, Figure 5 shows a network graph representing the direct and indirect evidence in the Wave 1 natalizumab case study (for more information, see Methodology Review (Mt-Isa et al., 2012): A.9.1 and A.9.4 and visual review: A.9 Network graph).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
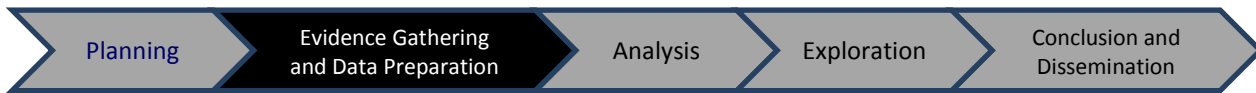
**Figure 5 Example of a network graph, based on the Wave 1 natalizumab case study. The solid lines represent direct evidence, and the broken lines represent indirect evidence. The texts next to each line refer to the literature reference.**
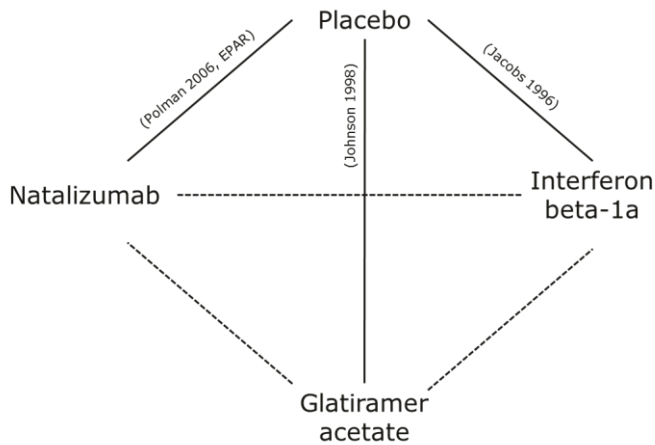


The forest plot or interval plot (Figure 6 and Figure 7, from the Wave 2 warfarin case study and Wave 1 rimonabant case study, respectively) can be used to communicate summary measures such as mean risk difference and risk ratios as well as their associated uncertainty (via confidence intervals). Forest/interval plots can be used as means of communicating benefit-risk data to specialist audiences such as physicians, regulators, and other experts. They may, however, be less intuitive for lay audiences to interpret.

**Figure 6 Forest/interval plot for risk differences between warfarin and control (The number in the box is a point estimate and the width of the box is the confidence interval. Benefits are coloured green, and risks are coloured red.)**

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
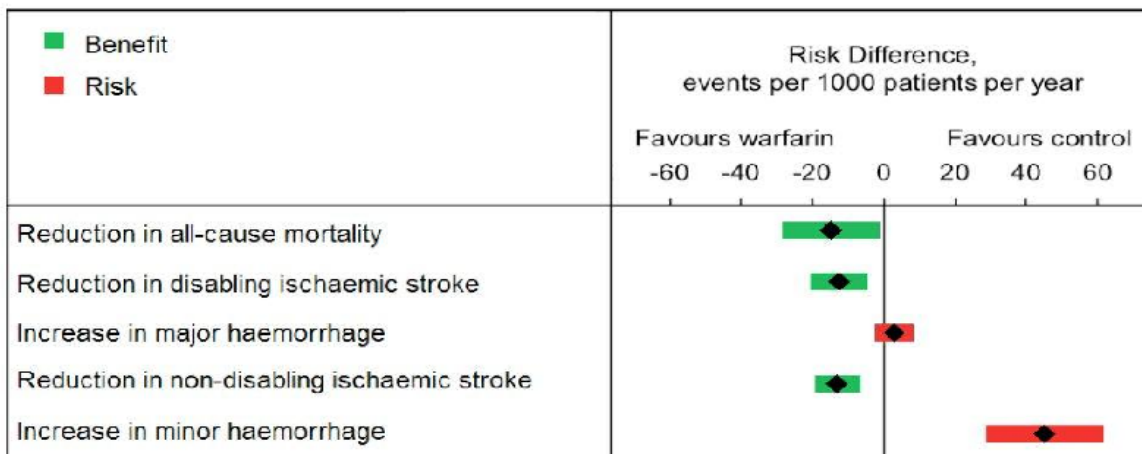
**Figure 7 Forest/interval plot for risk differences between rimonabant (Treatment A) and placebo (Treatment B) (The number in the box is a point estimate and the width of the box is the confidence interval. Benefits are coloured yellow, and risks are coloured blue.)**
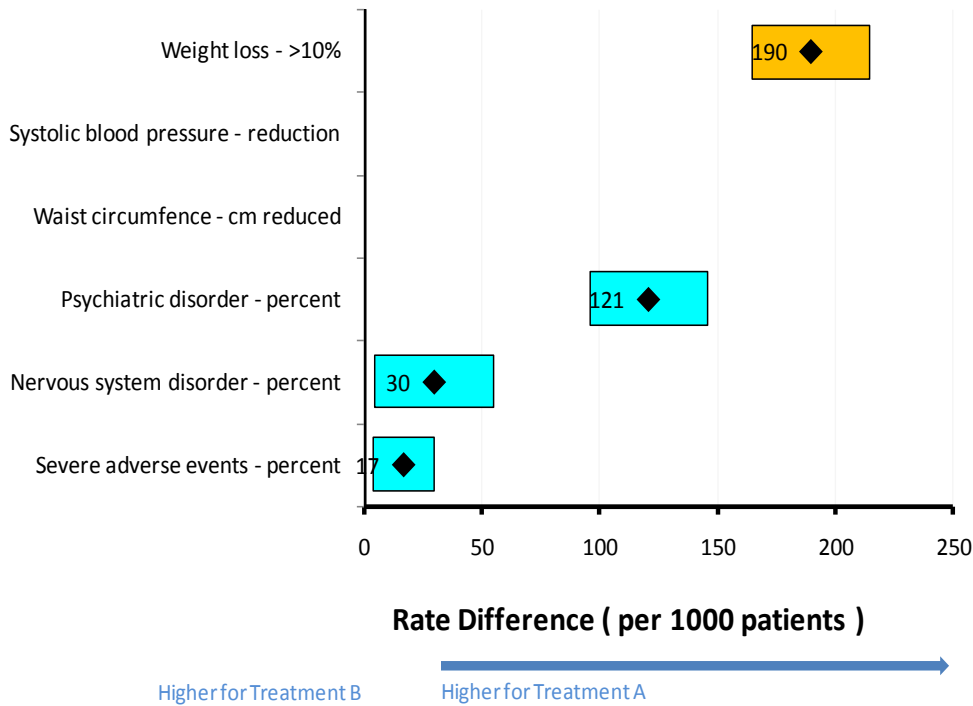


Figure 7 demonstrates a major limitation of forest/interval plots that was noted in the Wave 1 rimonabant case study: "The forest plot is most suitable when all outcomes can be measured in the same metric, e.g., percent." This is the reason why there is no data shown in Figure 7 for the two outcomes that were not expressed on the same scale as the others (i.e., Systolic blood pressure – reduction and Waist circumference – cm reduced).

Tables are easily produced in a variety of popular software packages. Forest plots can be produced easily in statistical software packages such as Stata, R, and SAS, and may be produced with a little more work in software such Microsoft Excel© and Tableau. The BRAT tool produces both these visualisations.

## Remarks

This section provides a systematic approach to gathering evidence and preparing data for a benefit-risk assessment. At the end of this stage, decision makers should have identified sufficient data for a first run of a benefit-risk model, aggregated them into suitable formats for analysis, and formed a clear idea of the types of analysis that are feasible given the data. Decision makers should also have finalised the criteria on the value tree to be used in the model. The initial model can be used to carry out 'what-if' analyses, which will help to direct the search only to information that

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
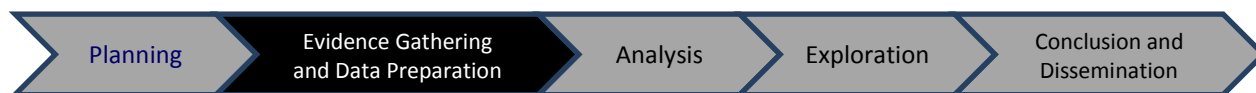
could affect the benefit-risk balance. The next section further discusses some of the benefit-risk assessment approaches available and provides some guidance on selecting appropriate ones that can deal with the data prepared.

## Evidence Gathering and Data Preparation: Summary of key points

- The Evidence Gathering and Data Preparation stage of the process concerns the identification and extraction of evidence relevant to the benefit-risk assessment.

- PROTECT WP5 recommends that clinical, statistical, epidemiological, and database expertise is engaged at this stage.

- The source of data should be clearly documented and justified for inclusion or exclusion, and be sufficiently clear to enable transparency in the data preparation process.

- Relative to deciding which sources of evidence to use, it is recommended that decision makers are aware that the sources of evidence:

  o must be appropriate and sufficient to the decision problem
  o can be aggregated if multiple studies are involved
  o may be aggregated using different weights according to their perceived reliability
  o should match criteria definitions and time horizon as closely as possible
  o should be based on a population of patients that resembles as closely as possible the population that will be affected by the medicinal product

- It is recommended that the decision maker be prepared to discuss the available sources of evidence with other stakeholders in order to reach agreement as to which sources of evidence are most appropriate to be used in a benefit-risk assessment.

- It is recommended that future uncertainties and likely scenarios should also be considered in the preliminary benefit-risk assessments.

- Data from the sources of evidence must be extracted and expressed in the appropriate, well-defined metric(s) for the anticipated method of analysis.

- Data extraction may be straightforward if a single evidence source provides all the measures that are needed and in the correct form for analysis; however, prior to analysis, some data may:

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
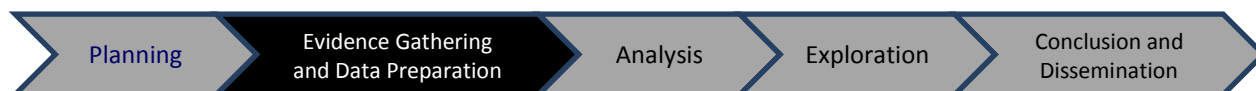
- o need a mathematical transformation which should be documented

- o need to be combined if there is no single source of evidence that can provide data on the performance of each comparator in relation to all of the benefits and risks

- o be combined to make efficient use of all the available evidence, resulting in more accurate estimates and less uncertainty in the data

- o be aggregated based on expert opinion (e.g., by a panel of clinicians considering the evidence and forming an opinion as to the appropriate parameter value and documenting the rationale), or on a statistical basis using meta-analytical techniques

- If there is no reliable evidence for a particular benefit or risk criterion, the recommended options should be to:

  - o consider using a different criterion definition or a surrogate criterion for which data are more readily available
  - o leave the criterion in the model and explore the sensitivity of the results to different assumed values [PROTECT WP5 recommended option]
  - o consider using a different method of analysis that is more compatible with the available data
  - o decide it is acceptable to omit the criterion from the analysis if it is not thought to be essential to the benefit-risk decision

- Estimation Techniques may be required when dealing with complex data or multiple sources of evidence in order to transform source data into a suitable form for benefit-risk assessment. Probabilistic Simulation Method (PSM) and Indirect/Mixed Treatment Comparison (ITC/MTC) were specifically evaluated in PROTECT WP5 case studies.

- Three visualisation techniques were identified as being useful at the Evidence Gathering and Data Preparation stage:
  - o Structured and colour-coded tables of evidence data and sources
  - o Network graphs of the relationship of evidence data and sources
  - o Forest plots of the evidence summary

- At the end of the Evidence Gathering and Data Preparation stage, decision makers should have identified all available data, aggregated them into suitable formats for analysis, and should have a clear idea of the types of analysis that are feasible given the data; as well as having finalised the criteria on the value tree to be used in the model, e.g., a key benefit-risk summary (KBRS) table or effects table.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Part 3: Analysis

At the Analysis stage, the data are evaluated to quantify the magnitudes of benefits and risks for the drugs of interest. Depending on the purpose and context of the benefit-risk assessment, the benefits and risks may be weighted and integrated to provide a quantitative measure of the benefit-risk balance.

This section addresses the following key questions relating to the Analysis stage of a benefit-risk assessment:

What type of analysis is required?

What types of methodologies are available to help with the Analysis stage?

Which metric indices were identified and reviewed by PROTECT WP5?

Which metric indices were evaluated in PROTECT WP5's case studies?

What are PROTECT WP5's recommendations regarding metric indices?

Which quantitative frameworks were identified and reviewed by PROTECT WP5?

Which quantitative frameworks were evaluated in PROTECT WP5's case studies?

What are PROTECT WP5's recommendations regarding quantitative frameworks?

Which utility survey techniques were identified and reviewed by PROTECT WP5?

Which utility survey techniques were evaluated in PROTECT WP5's case studies?

What are PROTECT WP5's recommendations regarding utility survey techniques?

What are PROTECT WP5's recommendations regarding the use of visualisations at the Analysis stage?

Are there any other comments or recommendations relevant to the Analysis stage?

### What type of analysis is required?

The starting point for analysis is a fully populated data table (or a forest plot if it can accommodate all of the benefits and risks).

"Fully populated" does not mean that every item of clinical evidence is known with certainty. As highlighted in the Evidence Gathering and Data Preparation section, there will always be uncertainty associated with the clinical evidence in the data table, and assumptions or provisional estimates may have been used in place of hard data. The

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
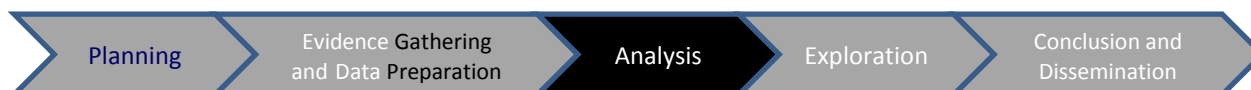
sensitivity of the analysis to these uncertainties will be examined at the Exploration stage; but for the purpose of this Analysis section, we assume that the figures in the data table can be taken at face value.

The level of analysis that is required will depend on the decision maker's judgement regarding the weight of evidence in the data table. Specifically, the decision maker should consider whether a *qualitative analysis* based on inspection of the data table is sufficient to enable a justifiable benefit-risk decision, or whether a *quantitative analysis* would strengthen the decision by providing a transparent, integrated measure of the benefit-risk balance.

To illustrate this point, three scenarios involving hypothetical anti-obesity drugs are shown below in order of increasing complexity. To keep things simple, we have not given precise definitions of the benefit and risk measures, which are not of particular importance to the discussion.

## Scenario A – a dominant drug

In the data table below, the percentages in the table represent the proportion of patients who experience each benefit or risk outcome.

| Benefits | Drug 1 | Drug 2 |
|---|---|---|
| Reduction in cholesterol | 50% | 40% |
| Weight loss | 20% | 18% |
| **Risks** | | |
| Transient nausea | 15% | 20% |
| Gastric ulcer | 5% | 6% |

Drug 1 is an example of a *dominant* alternative – it outperforms Drug 2 with respect to all of the benefit and risk criteria (i.e., it provides the greatest benefits and the smallest risks). On the basis of this evidence, any sensible decision maker would favour Drug 1. This is clear from a *qualitative analysis* of the data table, and further quantification to support the decision is not required.

## Scenario B – a simple weighting problem

Drug 3 is a new drug that has been developed for the same indication. Replacing Drug 2 with the new treatment results in the following data table:

| Benefits | Drug 1 | Drug 3 |
|---|---|---|
| Reduction in cholesterol | 50% | 60% |
| Weight loss | 20% | 20% |
| **Risks** | | |
| Transient nausea | 15% | 18% |
| Gastric ulcer | 5% | 5% |

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

In this example, neither treatment dominates. Drug 3 achieves a greater cholesterol reduction than Drug 1 but causes slightly more transient nausea. If the decision maker judges that the reduction in cholesterol benefit outweighs the nausea risk, Drug 3 would be favoured.

This is a simple example of *weighting*, whereby a judgement is made regarding the relative importance of the criteria in the data table. In this case, the weighting was *implicit*, i.e., the decision maker did not have to quantify the difference in importance between nausea and reduced cholesterol, but may be able to make a sound benefit-risk decision based on a *qualitative analysis* of the data table.

Implicit weighting is adequate for simple cases where the following conditions are satisfied:

- Only a small number of benefit and risk criteria (or sets of criteria) are to be weighed against each other;
- Only a small number of comparators are involved; and
- The difference in importance between the criteria being traded off is intuitively clear.

However, if any of these three conditions are not met, the cognitive strain involved in implicit weighting becomes overwhelming, as illustrated by the next scenario.

### Scenario C – a complex weighting problem

Sometime later, new studies have been carried out on Drugs 1, 2, and 3, providing updated estimates of the benefit and risk outcomes in the data table. The benefit-risk assessment is repeated with the resulting new data values:

| Benefits | Drug 1 | Drug 2 | Drug 3 |
|---|---|---|---|
| Reduction in cholesterol | 45% | 42% | 51% |
| Weight loss | 21% | 23% | 19% |
| **Risks** | | | |
| Transient nausea | 17% | 14% | 21% |
| Gastric ulcer | 3% | 5% | 7% |

In this scenario, the decision maker's task is considerably more difficult. Each of the three drugs has its own strengths and weaknesses amongst several key criteria. Attempting to weigh up the options implicitly and in a principled, transparent fashion becomes more challenging: such an approach would essentially constitute an attempt to mentally calculate and integrate the attractiveness of the three drugs' performance in relation to each of the four benefit and risk criteria, together with the relative importance of those four criteria. Even if the decision maker feels comfortable making the decision in this way, explaining the basis of the decision to a third party may not be as feasible. In other words, there is no transparency.

It is in such situations that *quantitative analysis* methods can help. Quantitative decision models are an aid to the cognitive process of decision making. They disaggregate a complex problem like Scenario C into simpler components

that are easier to understand and weigh up, and then may go on to use principled methods to integrate the components into a measure of the overall benefit-risk balance. This approach:
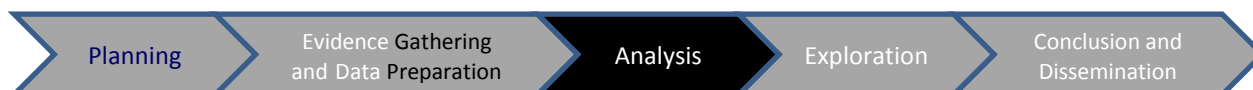
- facilitates clear thinking;
- provides a path to resolving disagreement regarding the benefit-risk balance (by pinpointing the aspects of the problem where those disagreements occur); and
- leaves a clear audit trail of the process.

Even if a qualitative analysis is sufficient, there are reasons why a decision maker might wish to incorporate preference weights and express the benefit-risk balance in quantitative terms. Explicit weighting can increase the transparency of the decision process, help to ensure consistency with other decisions, and establish priorities for the development of new treatments. Expressing the benefit-risk balance numerically can facilitate the sensitivity analysis, helping to ensure that decisions are robust.

On the other hand, quantitative methods require technical expertise and are generally more demanding in terms of resources than qualitative methods. Quantitative analysis of simple benefit-risk decisions may not be appropriate, and each case will usually need to be judged on its own merits.

There may be a tendency amongst some decision makers to distrust quantitative models on the grounds that they reduce multi-faceted problems to a single number representing the benefit-risk balance, thus giving a false impression of simplicity and discarding important nuances, as well as negating the impact of expert clinical judgement. However, this is a misunderstanding regarding the nature and interpretation of such models. Although the output may be a single number, it should never be interpreted as an estimate of a universal truth; rather, it must be interpreted in light of the preference values used or elicited by the model.

It is in the preference values that the complex nuances of the problem are represented. However, this may be a difficult point to communicate to audiences not familiar with the methods and the assumptions, and so there remains a danger that the output of quantitative models may be misinterpreted. Decision makers should therefore think carefully about when to use quantitative modelling and to whom the benefit-risk analysis will be communicated and, if it seems likely that quantitative measures will be misinterpreted, consider presenting only a qualitative analysis or additional clarification of the quantitative results. Given the ongoing efforts in data transparency, it must be assumed that benefit-risk modelling results will be published if consequences, such as approval or rejection of a marketing authorisation, are based on it.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## What types of methodologies are available to help with the Analysis stage?

PROTECT WP5 identified three types of methodologies as being useful during the Analysis stage:

- **Metric indices** provide numerical representations of benefits and risks, and for some metric indices in which benefits and risks are traded off, such as QALYs, the trade-offs are implicit.
- **Quantitative Frameworks** facilitate the creation of customisable models for trading off of benefits and risks, and providing an integrated measure of the benefit-risk balance.
- **Utility Survey Techniques** are methods for eliciting stakeholder preference information, which can be used to integrate benefits and risks in quantitative decision models.

## Which metric indices were identified and reviewed by PROTECT WP5?

The metric indices can be classified into three sub-categories: those that provide indices that are used as thresholds, those that characterise health outcomes and implicitly trade off benefits and risks, and those that explicitly trade off the quantified benefits and risks but may not necessarily be specific to health outcomes. There are other basic metric indices commonly used in epidemiology such as the incidence rates, relative risks, odds ratios, and attributable risks. These are also suitable to quantify benefits and risks for the purpose of decision making in medicine, but have not been reviewed by PROTECT WP5 as a strong body of literature already exists to guide and evaluate their application.

Table 7 shows the metric indices identified by PROTECT WP5's Methodology Review (Mt-Isa et al., 2012).

**Table 7 Metric indices**

| Threshold indices | NNT / NNH |
| | AE-NNT / NEAR |
| | RV-NNH |
| | Impact numbers |
| | MCE |
| | RV-MCE |
| | MAR |
| **Health utility indices** | QALY / DALY / HALE |
| | Q-TWiST |
| **Trade-off indices** | UT-NNT |
| | INHB |
| | BRR |
| | GBR |
| | Principle of three |
| | TURBO |
| | BM |

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Which metric indices were evaluated in PROTECT WP5's case studies?

Based on the appraisal execution results (as noted in Appendix 8), the following metric indices were recommended to be applied in the case studies:

- Number Needed to Treat (NNT) and Number Needed to Harm (NNH)
- Impact numbers
- Quality Adjusted Life Years (QALY)
- Quality adjusted Time Without Symptoms and Toxicity (Q-TWiST)
- Benefit-Risk Ratio (BRR)

NNT and NNH indicate the number of patients that would need to be given a treatment in order for a particular benefit (NNT) or risk (NNH) event to occur. Mathematically, they are calculated as the reciprocal of the difference between the event probabilities in the treated and untreated (or comparator) populations. The classical NNT/NNH approach to benefit-risk assessment allows only a single benefit and a single risk to be compared; and treatment is favoured if NNT>NNH. Implicitly, this gives equal weighting to the benefit and risk events. Extensions to NNT/NNH that allow weighting of multiple benefits and risks have also been developed.
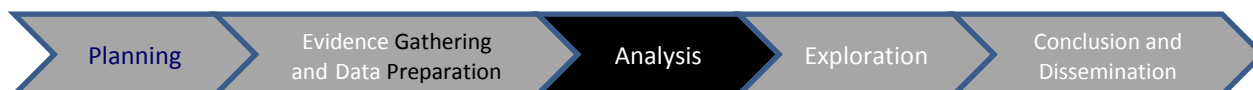
Impact numbers are an extension of the NNT/NNH concept that indicates the numbers of people that will be affected by medical conditions and/or treatments in specific populations. As with NNT/NNH, if multiple benefits and risks are to be weighed against each other, then additional techniques must be adopted.

Quality Adjusted Life Years (QALYs) are a measure of a patient's remaining lifespan adjusted for quality of life within each health state within the lifespan. This is achieved by measuring the duration of time periods where quality of life is expected to be impaired and multiplying the duration with a measure of quality of life on a scale from zero to one. The quality of life judgements effectively require implicit weighting of benefits and risks. QALYs (and other related measures) are well-established in the treatment of chronic diseases, where their ability to account for the time spent in specific disease states is particularly important.

Q-TWiST is an extension of QALY specifically developed for application in cancer treatments based on discrete health states experienced by the patients. It was first proposed in breast cancer trials (Goldhirsch et al., 1989). Q-TWiST is obtained by dividing survival time into discrete health states: TOX (time subject to toxicity effect), TWiST (time without symptoms and toxicity), and REL (time of relapse to death).

Benefit-Risk Ratio (BRR) is based on the ratio of benefits to risks.

All of these metric indices except QALY and Q-TWiST were evaluated in the Wave 1 case studies, as shown in Table 8 below, as no suitable PROTECT WP5 case studies could be found due to data availability. This exemplifies the need to select the appropriate methodology relative to the benefit-risk assessment. The Wave 2 case studies did not use

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

any of these specialised indices, but instead used more well-known epidemiological metrics (e.g., incidence) in combination with quantitative frameworks.

**Table 8 Metric indices tested in the PROTECT WP5 case studies**

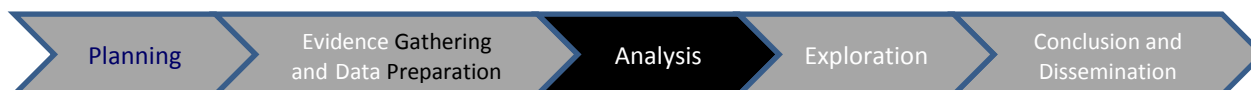| | Case study | NNT/NNH | Impact numbers | QALY | Q-TWiST | BRR |
|---|---|---|---|---|---|---|
| WAVE 1 | **Efalizumab** | | | | | ✓ |
| | **Telithromycin** | | | | | ✓ |
| | **Natalizumab** | ✓ | | | | |
| | **Rimonabant** | ✓ | ✓ | | | ✓ |
| WAVE 2 | **Rosiglitazone** | | | | | |
| | **Warfarin** | | | | | |
| | **Natalizumab** | | | | | |
| | **Rimonabant** | | | | | |

## What are PROTECT WP5's recommendations regarding metric indices?

### *Number Needed to Treat / Number Needed to Harm (NNT/NNH)*

NNT/NNH is characterised by its simplicity, which presents both advantages and disadvantages.

The concept has a simple interpretation and can be easily understood. Those who directly compare NNT and NNH for decision making purposes, however, should be aware of the implicit assumption they are making, i.e., that the benefit and risk events would be equally important. Furthermore, in its basic form, NNT/NNH has trouble dealing with more than one benefit and one risk, as noted in the Wave 1 rimonabant case study: "In analyses involving many criteria, the results from NNT are difficult to communicate and do not readily lend to a conclusion." We recognise that NNT or NNH is routinely used for communicating the likelihood of an individual event, particularly in physician-patient settings, and it may continue to be useful for this purpose. However, we cannot recommend NNT and NNH alone as useful measures for the purpose of weighing up multiple benefits and risks. Where benefits and risks of different relevance are included, "It does not speak to clinical relevance of any of the effects, does not deal with multiple benefits or multiple risks, or any of the trade-offs between the benefits and risks" (European Medicines Agency, 2013a).

The metric is restricted to binary endpoints, limiting the range of benefit and risk criteria that can be included. The advantage of this is that the statistics involved are simplified, particularly when exploring uncertainty. This has facilitated the development of probabilistic extensions to the basic NNT/NNH approach (Sutton et al., 2005).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Impact numbers

Impact numbers are useful descriptive tools when focusing on a single benefit or risk criterion (or very few criteria). As noted in the Wave 1 rimonabant case study, "the results from the analyses are directly applicable to the population of interest where the context can be placed immediately in terms of number of people who would be affected by the decisions." This makes impact numbers appealing as a tool for communicating the consequences of particular courses of action (e.g., to provide additional transparency and support for regulatory decisions).

As decision making tools, however, impact numbers appear less attractive, particularly for problems involving multiple benefit and risk criteria. However, the method may be useful for public health practitioners investigating the impact of interventions at the population level, such as vaccination programmes. There is no integration of benefits and risks, making complex problems difficult to digest. As with NNT/NNH, impact numbers use "the same unit for both benefit and risk criteria, but the scales may not be directly comparable" (the Wave 1 rimonabant case study), presenting a danger that inappropriate comparisons will be made. The potential for misapplication is increased by the fact that there are several impact numbers with different interpretations and there may be confusion as to which is the most relevant in any given situation.

Similar to NNT/NNH, impact numbers can only be defined for binary endpoints and ignore preference values.

## Benefit-Risk Ratio (BRR)

BRR requires a single measure for "benefits" and a single measure for "risks." It may, however, be applied to problems involving many criteria, either by focusing on the key benefit and key risk (as in the Wave 1 efalizumab case study and in the Wave 1 rimonabant case study), or by using an integrated measure for benefits and a separate one for risks. None of the case studies attempted the latter approach, although it might be appropriate in disease areas where integrated benefit and risk measures are routinely collected.

It is important to consider the interpretation of the benefit-risk ratio in terms of the relative value of the benefits and risks that have been included. For the purpose of transparency in decision making, it is good practice to make this value judgement explicit by defining an "acceptability threshold" or "acceptability curve." (An example from the Wave 1 telithromycin case study is shown below in Figure 8.)

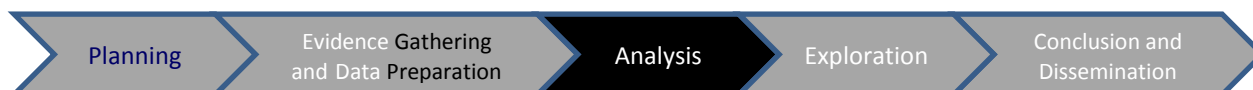| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

**Figure 8 Benefit-Risk acceptability curve for the probability that telithromycin is net-beneficial relative to comparator at any risk–benefit acceptability threshold. For example, if preferences were such that one is willing to accept 1 risk event to 1 benefit event, the probability that the drug provides a net benefit is 0.96**
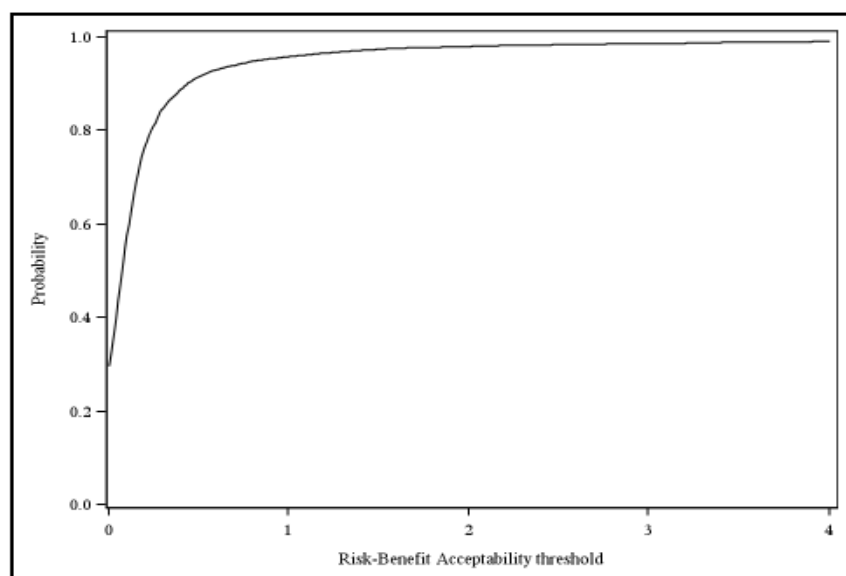


Unless the benefit-risk ratio is very large, adding information on the INHB can provide useful complementary information.

## Which quantitative frameworks were identified and reviewed by PROTECT WP5?

Nine (9) **Quantitative Frameworks** were identified as:

- **BLRA**
- **CUI**
- **Decision Tree**
- **DI**
- **MCDA**
- **MDP**
- **NCB**
- **SBRAM**
- **SMAA**

Further details are provided on each of these frameworks in Appendix 9.

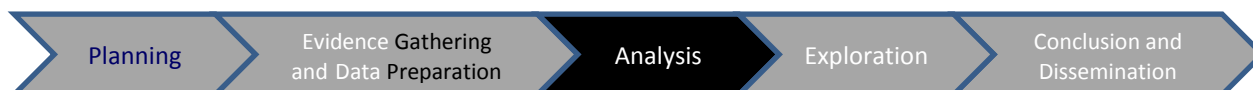| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Which quantitative frameworks were evaluated in PROTECT WP5's case studies?

MCDA and SMAA were recommended to be taken forward for benefit-risk assessment execution methodologies.

MCDA is a sound and flexible framework for integrating multiple benefit and risk criteria based on their perceived value, and using this as a basis for comparisons between alternative treatments.

SMAA is an extension of MCDA that incorporates probabilistic modelling of the treatments' performance data and does not require the weighting of benefits and risks to be specified a priori – instead, it explores all possible combinations of weights.

Two quantitative frameworks that were not recommended by the Methodology Review (Mt-Isa et al., 2012) were also applied in at least one case study: relative-value adjusted Number Needed to Treat (RV-NNT) and Sarac's Benefit-Risk Assessment Method (SBRAM).
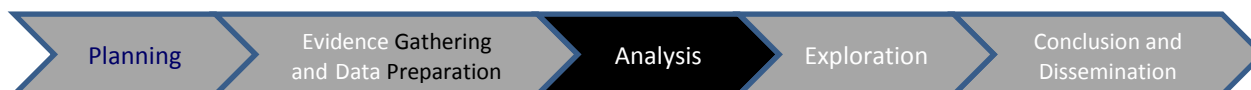
RV-NNT is an extension of the NNT metric that allows weighting of multiple benefits and risks (Holden, 2003). The Wave 1 natalizumab case study team demonstrated that RV-NNT is equivalent to a preference-weighted implementation of another quantitative framework, Net Clinical Benefit (NCB). RV-NNT is a quantitative framework that compares the overall difference in favourable and unfavourable effects, and also corresponds to a special case of the more general MCDA framework.

Sarac's Benefit-Risk Assessment Method (SBRAM) is conceptually similar to MCDA but uses a simplified system for scoring and weighting. Each benefit or risk criterion is weighted as either high, medium, or low importance; and the difference between a treatment and its comparator with respect to each criterion is reduced to a trichotomous measure (treatment is inferior, treatment is superior, or no difference).

Table 9 shows the quantitative frameworks that were evaluated in each PROTECT WP5 case study.

**Table 9 Quantitative frameworks tested in the PROTECT WP5 case studies**

|  | Case study | MCDA | SMAA | RV-NNT / NCB | SBRAM |
|---|---|---|---|---|---|
| WAVE 1 | Efalizumab | ✓ |  |  |  |
|  | Telithromycin | ✓ | ✓ |  | ✓ |
|  | Natalizumab | ✓ |  | ✓ |  |
|  | Rimonabant | ✓ | ✓ |  |  |
| WAVE 2 | Rosiglitazone | ✓ |  |  |  |
|  | Warfarin | ✓ |  |  |  |
|  | Natalizumab | ✓ |  |  |  |
|  | Rimonabant |  | ✓ |  |  |

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

### What are PROTECT WP5's recommendations regarding quantitative frameworks?

#### *Multi-Criteria Decision Analysis (MCDA)*

MCDA was found to be a highly effective methodology in all the case studies where it was used. Its systematic approach to breaking down complex problems facilitates "a transparent judgement of value between risk and benefit" (the Wave 1 rimonabant case study). The method "divides a complex problem into smaller criteria for assessment…this approach leads the decision makers to develop a deeper insight into the problem to be addressed as well as the alternatives to be considered" (the Wave 1 telithromycin case study).
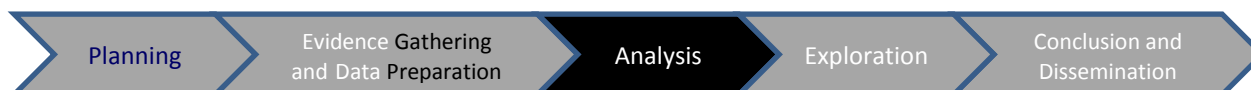
MCDA works particularly well in conjunction with the PrOACT-URL descriptive framework, as it provides the technical means for completing many of the framework steps. The Wave 1 rimonabant case study team described MCDA as "a natural progression of PrOACT."

It is worth pointing out that MCDA is an umbrella term for a broad range of related (but distinctly formulated) approaches to decision making involving several criteria, and not all formulations have been evaluated by PROTECT WP5 (Figueira et al., 2005). The Wave 1 telithromycin case study team observed that it is important to be "clear in our mind that there are many adaptions to this framework, and we have only tested one adaption." The differences between the forms of MCDA generally relate to the technical details of the underpinning theory, however. The MCDA framework that the PROTECT WP5 case study teams used is based on decision theory, and there are various ways to realise and apply that theory. There are other methods and theories that claim to be MCDA but that are not actually based on decision theory; and if these are to be applied, it is important to understand their underlying constructs (Belton and Stuart, 2002).

Most of the case study teams used the Hiview software to implement MCDA. This software was generally found to be efficient and easy to use, and its ability to produce key visualisations was appreciated, as noted by the Wave 1 efalizumab case study team: "MCDA is applied in the Hiview3 software with several graphical representations (Effects Tree, various coloured bar graphs) which provide easily understandable visualisation of results. This is easily provided by the software itself."

The Wave 1 telithromycin case study identified a significant limitation of Hiview3: The software only allows a single value for each benefit or risk, so it cannot directly "account for the uncertainties and random error with the statistical estimates…this is crucial in making medical judgements." However, it is possible to implement MCDA in a variety of software packages, including spreadsheets, and many limitations of standard software packages can therefore be overcome, given sufficient resources. This may allow for direct handling of uncertainty within the MCDA model; if such an implementation is not possible, then we recommend that the sensitivity of MCDA results to uncertainty is thoroughly investigated at the Exploration stage.

A major strength of the MCDA framework is that each benefit or risk can be expressed using any measure, as long as it is possible to convert the measure into a preference value or utility. This makes it particularly useful in benefit-risk

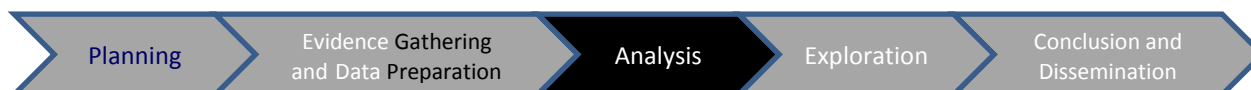| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

assessments later in a product's life cycle, where the data for benefits and risks may be drawn from different sources and therefore compatibility of measures cannot be guaranteed. This was observed by the Wave 1 efalizumab case study team: "Extreme heterogeneity of measures (absolute numbers, proportions with various denominators) is manageable in MCDA, which is useful in a post-marketing evaluation where measures are very heterogenic in nature and in units." The Wave 1 telithromycin case study team noted that, in fact, "the MCDA model is also applicable where there are few or no objective measures but only preferences."

On the other hand, MCDA usually relies upon direct data which is then transformed to common units of preference values so it must be possible for stakeholders to attach values to the benefit and risk measures that are included. In practice, this means using absolute measures of the outcomes of treatments in the target population instead of the relative measures of treatment effect (ratios or differences between groups) that are frequently reported in the medical literature. For example, if it is known that 50% of patients taking a drug experience a particular adverse event, we can associate a utility score with this outcome. If, on the other hand, we only know that the relative risk of the adverse event is 5 (compared to patients who take placebo), we cannot translate this into a utility because we do not know the magnitude of the risk.

The Wave 1 rimonabant case study team suggested that preference information may be "difficult to obtain in real life, particularly when the number of criteria is large." There are practical considerations when trying to elicit preferences, such as the cognitive burden on an individual or group of individuals asked to make distinct comparisons within a large number of scenarios, perhaps within a short period of time. These problems may be partly or wholly due to unfamiliarity with what are relatively new methods, and can be mitigated by the use of an experienced facilitator. There is also the question of whether the preferences elicited by a particular group on any given occasion are representative of the views of all relevant stakeholders.

The Wave 1 telithromycin case study and Wave 1 rimonabant case study teams found that it was not possible to obtain precision in weights and suggested that this can be explored using SMAA, which shows the effects on the benefit-risk balance by exploring all potential combinations of preference weights. Of course, there is also uncertainty in the data, and this too can be explored in SMAA as well as in probabilistic simulation method (PSM) in general.

Another potential difficulty was noted by the Wave 1 natalizumab case study team: "Outcomes should be expressed in non-comparative terms (i.e., using measures observed in a particular population, rather than comparative measures such as risk ratios that relate to the difference between populations) as this is the scale [on which] preference values are expressed." By contrast, it is outcomes on a comparative scale (i.e., relative to a comparator or no treatment) that are fundamental to the results of a benefit-risk assessment and most frequently reported in the medical literature. The requirement under MCDA to convert outcomes to an absolute scale for the purpose of preference elicitation may increase the complexity of analysis, particularly if complex data networks or probabilistic models are employed. The multiple steps involved in transforming outcomes, converting to utility scores, and

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

weighting may increase the opportunities for bias to creep into the analysis. The above use of the term 'absolute' should not be confused with the term as it is used in measurement theory, where it simply refers to the measurement of numerousness, as in counting the number of patients who experienced a side effect. In general, when measures of effects are combined to give overall results, meaningful interpretation of the results expressed as ratios requires ratio-scale data (arbitrary unit) on all effect data. Results expressed as differences can be based on any mixture of interval-scale data (arbitrary unit and zero point) and ratio-scale data. Utilities are assessed as interval scales, so ratios of utilities can be misleading (Krantz et al., 1971).
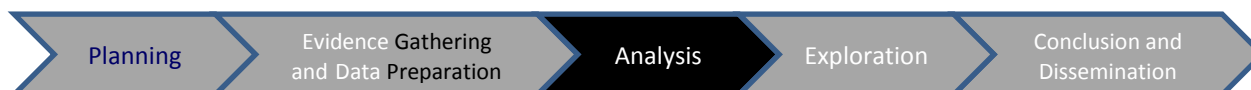
### *Stochastic Multi-criteria Acceptability Analysis (SMAA)*

SMAA attempts to overcome two of the most commonly cited shortcomings of standard MCDA models: It allows for flexibility in the handling of performance estimates and preference information. Performance estimates can be input as point estimates or using a distribution. Preferences can be input as missing, ordinal (i.e., outcomes are ranked), or cardinal (i.e., the exact value or the interval of the preference data can be specified). Distributions can be used to describe the performance estimates and preference information in an SMAA model, which allow for: (a) the incorporation of uncertainty in the performance of a treatment, e.g., the effect of confidence intervals on the decision can be investigated, and/or (b) the incorporation of uncertainty in preferences, e.g., the effect of varying preferences can be accommodated if stakeholders do not come to consensus. Instead of calculating the best option with a given set of data and preferences, SMAA estimates the probability that each comparator is the best option available conditional on the probability of different data and preference information.

The Wave 1 rimonabant case study used specialised J-SMAA software, which is available free online (a version still in development was used by the PROTECT WP5 case study teams, who discovered limitations in the current version). It is also possible to implement an SMAA-style approach in other packages: The Wave 2 rimonabant case study developed STATA code and a Microsoft Excel$^{©}$ spreadsheet for this purpose, and the Wave 2 natalizumab case study used WinBUGS to allow for uncertainty of clinical data (but not preference weights). However, a custom-built implementation such as these will require statistical expertise, whereas J-SMAA may be more suitable for naive users – though this can lead to errors in interpretation if the user does not understand what is going on "under the hood." There may be implicit assumptions that are important to consider when presenting the results. For example, it is typical to assume that the different benefit and risk criteria, although variable, are not correlated in any way; in other words, the expected score for each benefit or risk criterion is not affected by the scores on the other criteria. In reality, one might expect some of the benefits or risks to be correlated, as noted in the Wave 1 telithromycin case study report: "The performances of an alternative on different criteria are likely to be correlated. Currently, they are taken as independent in SMAA simulations. … It is not clear how much the correlation affects the results."

The use of SMAA is recommended in the following situations:

- To investigate the impact of uncertainty on an existing MCDA model;
- To model the distribution of the benefit-risk balance based on patient-level data; or

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

- When clear preference information is missing or when a consensus cannot be reached.

Documenting preference information which has been elicited from stakeholders is desirable for transparent benefit-risk assessments. Therefore, we believe it is appropriate to elicit weights where possible and to supplement this approach with SMAA if necessary to demonstrate robustness to variability (as suggested by the Wave 1 rimonabant case study team: "Although precise weighting information is not needed … A decision conference would be useful to elicit the stakeholder preference between criteria to examine the sensitivity of the model.").

## Relative-Value Adjusted Number Needed to Treat (RV-NNT) and Net Clinical Benefit (NCB)

RV-NNT has been developed as an extension of the NNT/NNH family of indices that allows multiple benefits and risks to be considered simultaneously and weighted according to elicited preference values (Holden, 2003).
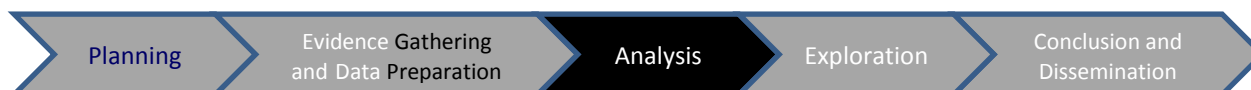
Net Clinical Benefit (NCB), as described by Sutton et al. (2005), is a measure of the difference in favourable and unfavourable effects between two treatments. NCB is defined as the sum of the difference in benefits minus the sum of the difference in risks; however, the framework is flexible with regard to how the benefits and risks are measured and summed.

The Wave 1 natalizumab case study team observed that the RV-NNT framework can be seen as an implementation of NCB with explicit preference weighting. Furthermore, it is equivalent to a special case of MCDA in which every benefit or risk outcome is expressed as a binary variable with a particular form of linear value function. The method is recommended as long as these restrictions are considered appropriate.

## Sarac's Benefit-Risk Assessment Method (SBRAM)

Sarac's Benefit-Risk Assessment Method (SBRAM) was designed for use by pharmaceutical companies during the drug development process (Sarac et al., 2012). SBRAM is an extension of MCDA that uses 3-point scoring and weighting systems. A drug is scored relative to a comparator on each criterion as superior (+1), non-inferior or equivalent (0), or inferior (-1). Scoring rules are established for both discrete and continuous data. A simple approach to uncertainty provides for scores to be expressed as intervals. Criterion weights appear to represent the relative importance of the effects, with "a weight/importance of 1 (low), 2 (medium), or 3 (high)." Scores are multiplied by weights to give weighted scores that can range from +3 to -3, or as an interval. These weighted scores are displayed as a tornado-like diagram, with sections coloured to indicate inferiority, non-inferiority, or superiority of the drug for each effect.

As this particular application is not within PROTECT WP5's remit, we have not carried out extensive testing of SBRAM. However, it was evaluated in the Wave 1 telithromycin case study. This case study team found the method to be quite demanding in terms of the statistical work involved ("the process of scoring criteria is not straightforward for layman and there exists no finished software for the methods") and also noted restrictions on the form and

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

source of evidence (the data for each benefit or risk must be drawn from a single trial and "the method cannot (in this development stage) accommodate input from Meta-analysis").

On a more fundamental level, simplified scoring systems such as that used in SBRAM have been criticised elsewhere, as they discriminate poorly between alternatives and may increase the potential for bias (Nutt et al., 2010).

Overall, PROTECT WP5 does not recommend SBRAM as a tool for benefit-risk assessment during later-phase trials or after marketing registration.

## Which utility survey techniques were identified and reviewed by PROTECT WP5?

Four Utility Survey Techniques were identified as:

- **Stated Preference Method**
- **Contingent Valuation Method**
- **Conjoint Analysis**
- **Discrete Choice Experiments (DCE)**
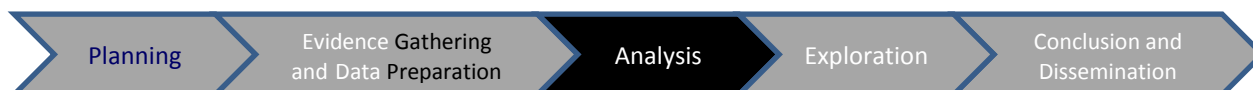
Further details are provided in Appendix 10.

## Which utility survey techniques were evaluated in PROTECT WP5's case studies?

Based on the results of the Methodology Review (Mt-Isa et al., 2012) (as noted in Appendix 10), Conjoint Analysis and Discrete Choice Experiments (DCE) were recommended to be applied in the case studies.

### *Conjoint Analysis / Discrete Choice Experiments*

Both conjoint analysis and discrete choice experiments are versions of stated preference methods. To apply either approach, participants in elicitation sessions are shown two different items to be compared. Each item is defined by a specific level of achievement on each criterion that is relevant to the item, such that the combination of levels on item A is different from the combination on item B. As applied to drugs in PROTECT WP5, the criteria are the favourable and unfavourable effects, and the levels of achievement are possible realised performance of a drug for the effects. Many possible drugs are then generated, and assessors are asked to compare two different drugs at a time, stating which of the two they prefer. From many preference statements about many pairs of hypothetical drugs, criterion weights and utilities or preference values can be calculated.

**Swing weighting** is perhaps the most direct Conjoint Analysis method. The first step is to define the range of values for the benefit and risk criteria by choosing realistic "worst" and "best" outcome for each criterion. Participants are

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

then asked to imagine that all benefits and risks are at the worst score and to choose the outcome they would most like to move to the best score. They are then asked, "How big is the worst-best difference on this criterion, and how much do you care about it?" on this criterion compared to the one with the biggest swing. This is a "thought stepping stone" for putting preference weights on these outcomes. For this, the top ranked criterion is given a weight of 100, and the participants assign weights to the other criteria to reflect their relative importance.
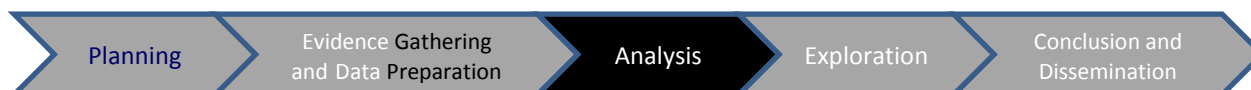
**Analytic Hierarchy Process (AHP)** breaks down the problem into a set of pairwise comparisons between clusters of benefit and risk criteria. Participants are asked to judge the relative importance of each pair of criteria on a numerical scale from 1 (equally important) to 9 (extreme preference for one criteria). AHP assesses the consistency of these pairwise judgements and translates them into unitless priority numbers (weights, preferences, or likelihoods, depending on what questions are asked of the assessors) that range from 0 to 1.0.

**MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique)** does not require participants to assign numerical values. Instead, the difference in attractiveness between each pair of benefit and risk criteria is expressed on the following categorical scale: neutral, very weak, weak, moderate, strong, very strong, and extreme. MACBETH assesses the consistency of these pairwise judgements and translates them into numerical preference values. If the question posed to the assessors is to compare the difference in worst-to-best attractiveness on one criterion with another using the same seven categorical difference descriptions, then the result is a set of relative importance weights for the criteria. In a similar fashion, value functions can be obtained with the categorical scoring technique.

Table 10 shows the utility survey techniques that were tested in PROTECT WP5's case studies (and by the PPI workstream, whose work developed out of an extension to the Wave 1 natalizumab case study).

**Table 10 Utility survey techniques tested in the PROTECT WP5 case studies**

|        | Case study | Swing weighting | AHP | MACBETH | DCE |
|--------|------------|:---------------:|:---:|:-------:|:---:|
| WAVE 1 | **Efalizumab** | ✓ | | | |
|        | **Telithromycin** | | | ✓ | |
|        | **Natalizumab** | ✓ | | | |
|        | **Rimonabant** | ✓ | | | |
| WAVE 2 | **Rosiglitazone** | ✓ | | | |
|        | **Warfarin** | | | | |
|        | **Natalizumab** | | | | |
|        | **Rimonabant** | | | | ✓ |
| PPI    | **Natalizumab** | ✓ | ✓ | ✓ | ✓ |

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

## What are PROTECT WP5's recommendations regarding utility survey techniques?

It is recommended to hold a decision conference to organise the process of eliciting preference information for a benefit-risk analysis. This was the approach used in the efalizumab, telithromycin, natalizumab, and rosiglitazone case studies. A decision conference is a meeting between individuals representing key stakeholder views, where the importance of the benefits and risks is discussed and elicited using techniques like those in Table 10.

As noted in the Wave 1 rimonabant case study and the Wave 1 telithromycin case study, it is recommended that: "(1) Stakeholders selected for the decision conference need to be wide enough to accommodate views from different parties – regulators, physicians, and patients. (2) Information on criteria would need to be available for review prior to meeting."

Disagreements often arise between the stakeholders involved in a decision. The decision conference format facilitates debate and the sharing of relevant experience, which may help the team arrive at a consensus. If substantial disagreement persists, the effect of using different preferences for analysis can later be explored. Some utility survey techniques – including AHP, MACBETH, and DCEs – can be designed to incorporate more than one set of preferences and provide an average result and, in some cases, the distribution of results.
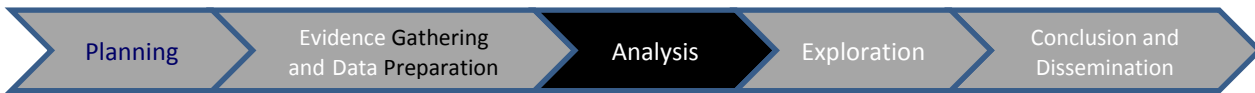
It is important that the patients' perspectives are considered in the benefit-risk assessment process. For example, patients can contribute their views about what benefits and risks should be included, and how relatively important each one of these is to them.

As a matter of being pragmatic, a team may choose to rely on internal expertise as a surrogate for specific external perspectives as this may be more efficient than finding external participants to represent stakeholder views. For example, the Wave 1 rimonabant case study used case study team members to represent the perspectives of physicians, non-physicians, and regulators based on their respective areas of expertise.

As noted in the Wave 1 efalizumab case study, it is recommended that: "The weight given to some outcomes would deserve thorough discussion on their medical relevance (e.g., reversibility of serious risks, long term continuation of short term benefit)."

As noted in the Wave 1 efalizumab case study, it is recommended that "structured and validated questionnaires should be developed and used for these methods." Questionnaires may need to include supporting material such as glossaries to explain any unfamiliar concepts, and these should be pitched at the correct level for the participants. Sufficient time should be budgeted for preparing and validating the supporting material.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## What are PROTECT WP5's recommendations regarding the use of visualisations at the Analysis stage?

Our recommendations regarding visualisations at the Analysis stage are set out below, grouped according to their specific purpose.

A relevant point throughout is that graphics must be well-designed for the intended audience. Higher level of education and perceived numeracy skills are associated with higher clarity and understanding of information on a visual display. Therefore, more work should be done when benefit-risk information needs to be communicated to audience with low perceived numeracy skills (Dolan et al., 2012).
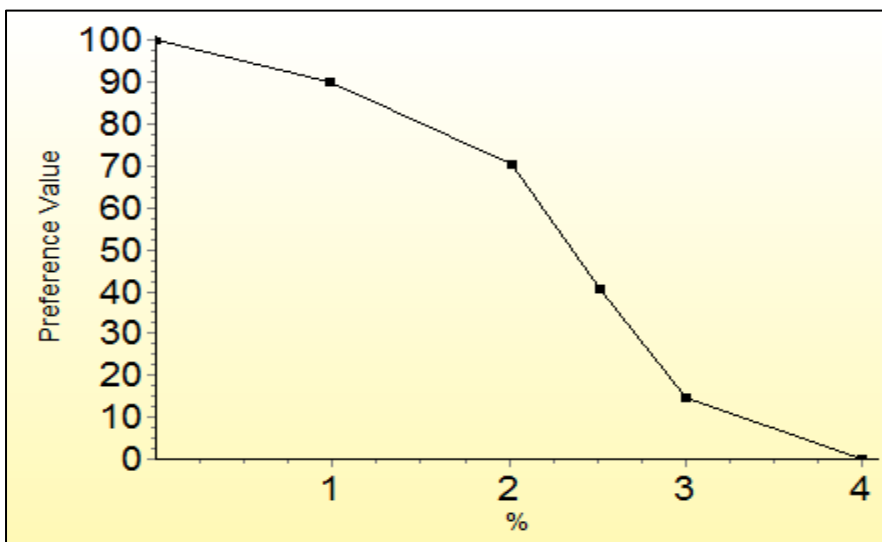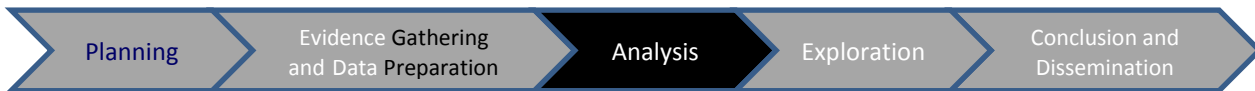
### *Eliciting preferences*

The use of visualisation methods for preference elicitation can reduce cognitive burden on the stakeholders, ensure proper understanding of the message, and engage stakeholders in the elicitation process.

Using a value-tree diagram to communicate the structure of the decision problem to the stakeholders is recommended for all elicitation methodologies. This should be supplemented where necessary by a glossary of the relevant medical terminology.

There are also specific visualisation methods that work well with particular elicitation methods. For example, Hiview3 provides an interactive graph whose data points can be dragged to the desired position in order to build a preference value function (Figure 9).

**Figure 9 A non–linear value function for the percentage of patients who experienced congestive heart failure in a clinical trial.**

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

The standard method of weight elicitation in MCDA is swing-weighting, assessing the swing in preference from the worst to the best positions on a criterion scale. Thermometer-like displays make it easier to visualise the difference between worst and best so the largest clinically relevant swing can be assigned 100 and the other swings judged as ratios relative to 100 (Figure 10).

**Figure 10 Swing weights as displayed in Hiview3 for four unfavourable effects criteria considered in the rosiglitazone case study.**
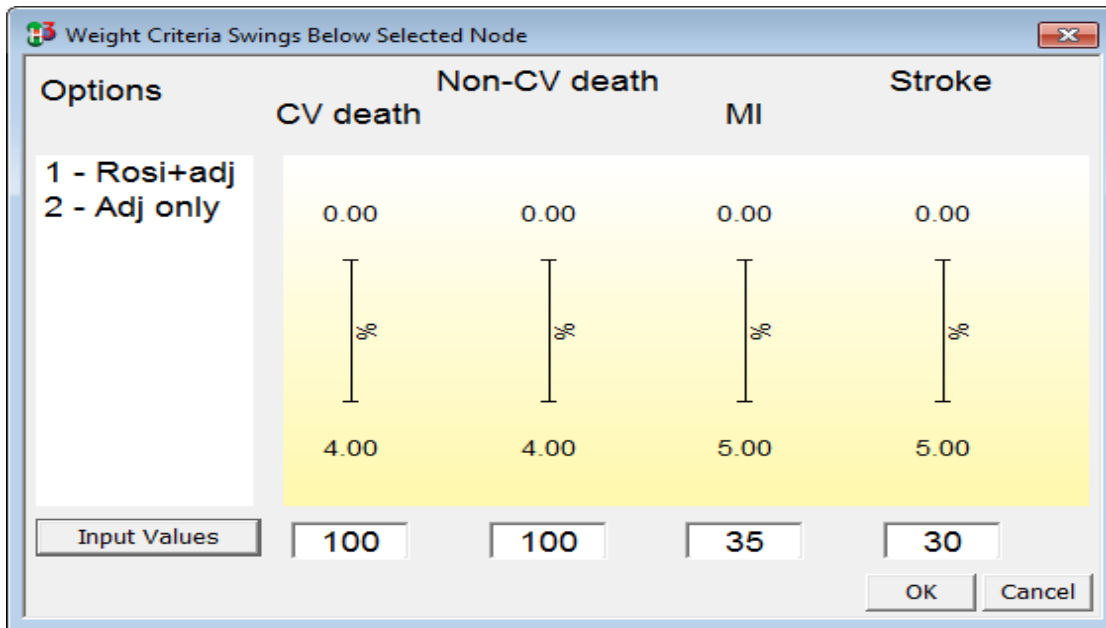


MACBETH for Hiview3 makes use of an interactive table display to enable assessors to make judgements about differences in criterion-weights according to MACBETH's categorical scale (Figure 11).
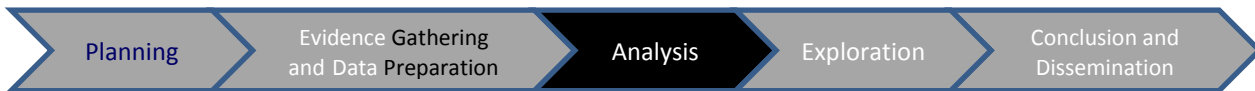
| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

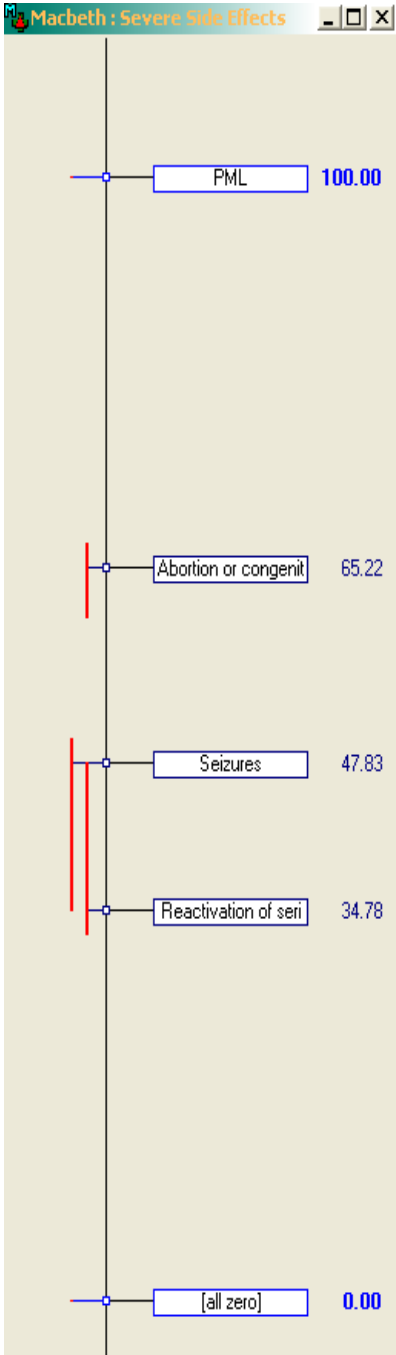**Figure 11 The MACBETH interactive table in Hiview3 for eliciting qualitative differences in criterion weights.**



Consistency checks made on the categorical judgements as they are inputted help the assessors to provide reliable assessments and thus the integrity of the resulting MCDA model. An interactive thermometer scale is also used to visualise the resulting criterion weights, allowing users to fine tune their preferences about weights within their consistency bounds (Figure 12).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|----------|------------------------------------------|----------|-------------|------------------------------|

**Figure 12 Thermometer scale of relative weights on node "severe side effects" from Macbeth**
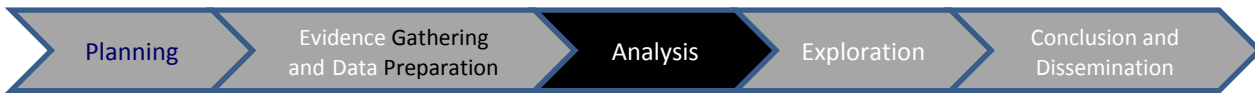
**Figure 13 A typical table used for eliciting preference weights in AHP**



The analytical hierarchy process (AHP) has the least established visualisation method. One common technique is to present the criteria within a group in a matrix or table (Figure 13) to be compared directly, but users need to specify the direction of preference and assign quantitative preference values manually (typed in). We have developed an example of how this approach can be adapted by creating a web-based survey in SurveyMonkey (http://www.surveymonkey.com).   This provides a user-friendly interface to replace the typical table. The quantitative values have also been replaced with categorical statements (right-most column in Figure 14) to help support judgement, but in this case, this comes at the expense of fewer categories for "importance" ratings. The advantage is that the interactive drop-down lists are likely to ease cognitive burden to the stakeholders. The categorical statements are later converted to the predetermined set of ordinal numeric values for analysis but are unknown to the responders at the time of answering the survey.
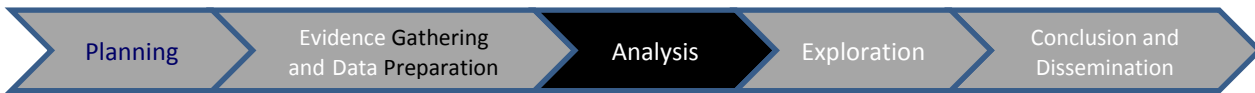
| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

**Figure 14 AHP drop-down list of response choices**



*Presenting results*

The use of appropriate visualisations can speed up stakeholders' response times when reading outputs from a benefit-risk analysis.
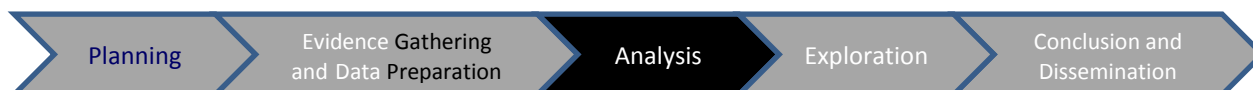
## Qualitative or partially quantitative analysis

To present the results of a qualitative or partially quantitative analysis, a table or forest/interval plot is recommended.

Tables provide fast and efficient readability across issues displayed in rows and columns. They can serve as a common means for benefit-risk communications because of their simple structure, flexibility, and the ease with which they can be adapted. Although some individuals may not intuitively think of tables as a form of visual representation, tables can be very powerful as a communication tool whilst also conveying a substantial amount of information. They can be used when communicating benefits and risks to all audiences including the general public, mass media, patients, doctors, regulators, and other experts such as analysts.

The ability to comprehend tables is highly dependent on the verbal and numerical format of the display. For tables representing summary statistics and specialist benefit-risk metrics, a statistical background may be required. Likewise, tables loaded with medical terms require some medical knowledge to be understood.

Good tables ease cognitive burdens of users and decrease the time required to extract the information. Tables should be limited to the requisite number of rows and columns to avoid adding cognitive burden when reading

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

tables. Readability can be enhanced through the use of colour-coding to represent grouping and relationships, as done in the BRAT framework (Figure 18).

Numerical presentation in tables (Figure 15) can influence how an individual may perceive the benefits or risks of a treatment. Any misunderstanding of the numerical presentation of a benefit-risk metric could lead to an incorrect interpretation and the potential for erroneous treatment decisions.

Tables sometimes are thought of as containing a list, which could give a false impression on benefit-risk balance because people tend to perceive a drug with a long list of risks as having an unfavourable benefit-risk balance without taking into account the actual quantitative data. Hierarchies may be perceived when reading a table since the information appears by lines and inevitably would be read as such. There may also be some issues of overlapping information presented in a table; e.g., when presenting events which are not mutually exclusive such as measuring "all deaths" and "death from cancer" in an analysis leading to double-counting the (latter) events. The existence or non-existence of hierarchies and overlapping information should be clarified when presenting information in tables, such as by accompanying tables with a tree diagram to visualise hierarchy or a Venn diagram to visualise inclusivity.

**Figure 15 PrOACT-URL 'effects table' listing the criteria for the benefit-risk assessment model in the Wave 1 efalizumab case study**

| | Name | Description | Fixed Upper | Fixed Lower | Units | Weight | Efalizumab | Placebo |
|---|---|---|---|---|---|---|---|---|
| Favourable Effects | PASI75 | Percentage of patients achieving 75% reduction in baseline PASI[1] at week 12. | 60.0 | 0.0 | % | 1.0 | 29.5 | 2.7 |
| | PGA | Percentage of patients achieving Physician's Global Assessment[2] clear/almost clear at week12. | 40.0 | 0.0 | % | 0.8 | 295 | 5.1 |
| | OLS | Percentage of patients with Overall Lesion Severity rating of minimal or clear at FT (day 84). | 40.0 | 0.0 | % | 0.25 | 32.1 | 2.9 |
| | DLQI | Dermatology Life Quality Index[3]. Mean percentage of patients showing an improvement. | 10.0 | 0.0 | Change score | 0.8 | 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 | 0.2 | 41.0 | 24.0 |
| | Severe infections | Proportion of patients experiencing infections serious enough to require hospitalisation. | 3.00 | 0.00 | %/100ptyrs | 1.0 | 2.83 | 1.4 |
| | Severe Thrombocytopenia | Number of cases exhibiting severe (grade 3 and above) thrombocytopenia[4]. | 10 | 0 | number | 0.8 | 9 | 0 |
| | Psoriasis Severe Forms | Percentage of patients developing severe forms of psoriasis (erythrodermic, pustular). | 4.0 | 0.0 | % | 0.05 | 3.2 | 1.4 |
| | Hypersensitivity Reactions | Percentage of patients exhibiting hypersensitivity reactions, arthralgia, psoriatic arthritis, flares, back pain asthenia, ALT and Ph. Alk increase. | 10.0 | 0.0 | % | 0.05 | 5.0 | 0 |
| | Intersticial Lung Disease | Number of cases of intersticial lung disease. | 20 | 0 | number | 0.1 | 18 | 0 |
| | Inflammatory Polyradiculopathy | Number of cases of inflammatory polyradiculopathy. | 5 | 0 | Data | 0.02 | 4 | 0 |
| | SAEs | Number of cases of haemolytic anemia. | 25 | 0 | number | 0.12 | 24 | 0 |
| | PML | Number of cases of progressive multifocal leukoencephalopathy. | 5 | 0 | number | 1.0 | 3 | 0 |
| | Aseptic Meningitis | Number of cases of aseptic meningitis. | 30 | 0 | number | 0.1 | 29 | 0 |

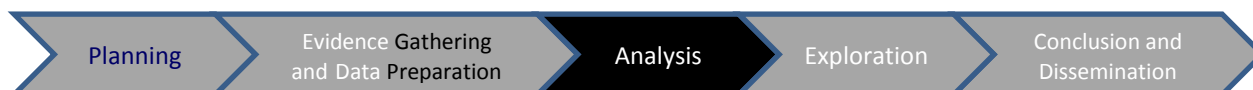| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

**Figure 16 BRAT 'source data table' listing the data for benefit-risk assessment model and their source in the Wave 1 natalizumab case study**

| Study ID | Value Tree Category | Outcome | Measure | Study Drug | Study Drug Estimate | Ref Group | Ref Group Estimate | Study Estimate |
|---|---|---|---|---|---|---|---|---|
| Polman 2006/ EPAR | Disease Activity | Relapse | Annualized Relapse rate[95%CI] | Drug A | 0.23 [0.19-0.28] | Placebo | 0.73 [0.62 –0.87] | 0.32 [0.26 – 0.40] |
| Jacobs 1996 | Disease Activity | Relapse | Annualized Relapse rate[95%CI] | Drug B | 0.67 [n.a.] | Placebo | 0.82 [n.a.] | 0.82 [0.56 – 1.20] |
| Johnson 1998 | Disease Activity | Relapse | Annualized Relapse rate[95%CI] | Drug C | 0.65 [n.a.] | Placebo | 0.91 [n.a.] | 0.71 [0.47 – 1.08] |
| ... | | | | | | | | |
| Polman 2006 | Liver Tox | ALT>5x ULN | n/N (%) | Drug A | 31/627 (5%) | Placebo | 12/312 (4%) | RR = 1.25 |
| Jacobs 1996 | Liver Tox | ALT>5x ULN | n/N (%) | Drug B | Not reported | Placebo | Not Reported | RR = 1 |
| Johnson 1998 | Liver Tox | ALT>5x ULN | n/N (%) | Drug C | Not reported | Placebo | Not Reported | RR = 1 |
| ... | | | | | | | | |

**Figure 17 Example of a table showing increased risk of headaches and nausea caused by taking pills (reproduced from Hawley et al., 2008)**

| | No pill | Pill A | Pill B |
|---|---|---|---|
| **Get mild headaches** | 10 out of 100 | 27 out of 100 | 27 out of 100 |
| **Get severe nausea** | 1 out of 100 | 9 out of 100 | 13 out of 100 |

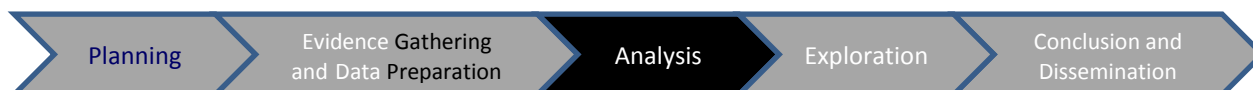| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

**Figure 18 A  colour-coded table within BRAT framework as applied in Wave 1 natalizumab case study**

| | | Outcome | Natalizumab Risk/1000 pts | Comparator Risk/1000 pts | Risk difference (95%CI)/1000 pts | |
|---|---|---|---|---|---|---|
| Benefits | Convenience benefits | Convenience (weight 0.6%) | - | - | - | (-,-) |
| | Medical benefits | Relapse (weight 3.9%) | 280 | 450 | -260 | (-326,195) |
| | | Disability progression (weight 5.6%) | 110 | 230 | -120 | (-,-) |
| Risks | Infection | Re-activation of serious herpes viral infection (weight 6.7%) | 80 | 70 | 10 | (-26,45) |
| | | PML (weight 55.9%) | 2 | 0 | 2 | (-,-) |
| | Liver toxicity | Transaminases elevation (weight 11.2%) | 50 | 40 | 10 | (-16,38) |
| | Reproductive toxicity | Congenital abnormalities (weight 5.6%) | - | - | - | (-,-) |
| | Neurological Disorders | Seizures (weight 5. %) | 0 | 0 | 0 | (-,-) |
| | Other | Infusion/injection reactions (weight 2.8%) | 236 | 180 | 56 | (6,114) |
| | | Hypersensitivity reaction (weight 1.1%) | 90 | 40 | 50 | (20,82) |
| | | Flu-like reactions (weight 1.1%) | 399 | 400 | -1 | (-114,114) |

Higher for natalizumab

Higher for comparator

## Quantitative analysis

To present the results of a quantitative benefit-risk assessment, a bar chart is recommended.

Bar charts come in many types. Several examples are shown in Appendix 11.  Bar charts can be used to communicate magnitude of any measure (e.g., benefit-risk score, probability of outperforming a comparator). Stacked bar charts can be used to depict proportions. Bar charts can also be used to display the benefit-risk trade-offs and to compare between options (stacked bars), as in the MCDA software Hiview3.  A specific application of a bar chart within Hiview3 is known as the 'difference display,' where the bars are indicative of the difference in benefit and risk scores between two treatments (Figure 19).  This is a particularly important visualisation because of the comparative nature of benefit-risk assessment.
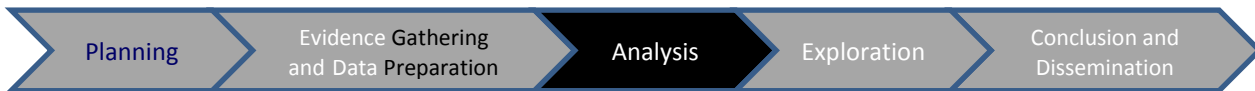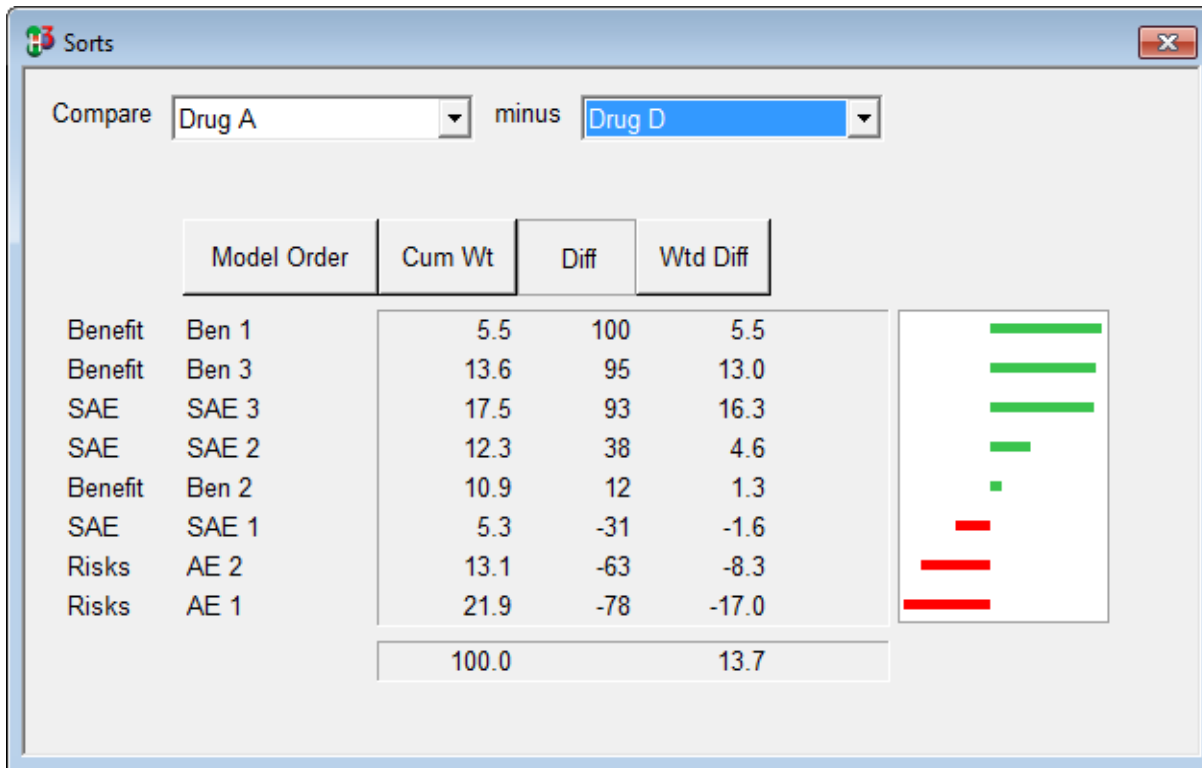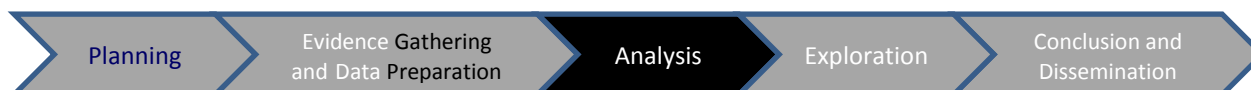
| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

**Figure 19 An example of a difference display from Hiview3**



Bar charts may be suitable to be used as a visual communication tool to a large variety of audiences such as the general public through the media, patients, physicians, regulators, and other experts. The 'difference display,' in particular, has been found to be a very useful visualisation of benefit-risk balance in regulatory decision making (European Medicines Agency, 2012). Any medical terminologies used obviously require some explanation if presented to an audience with no medical knowledge.
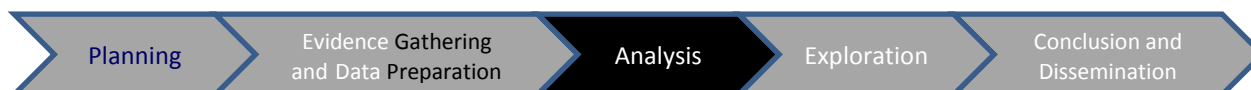
When a quantitative benefit-risk assessment approach is used, we recommend that stakeholders' value preferences (i.e., the weights of the benefit and risk criteria) and the magnitudes of the final benefit-risk metrics should be visualised using simple or stacked bar charts. The importance of presenting the preference weights should not be overlooked; as we have already mentioned, this information is key to interpreting an integrated benefit-risk metric.

To visualise the contributions of the different benefit and risk criteria in the benefit-risk analysis, the use of stacked bar charts, difference displays, or grouped bar charts is recommended.

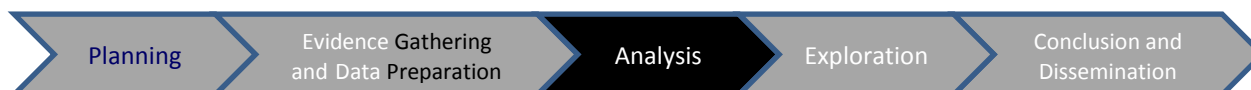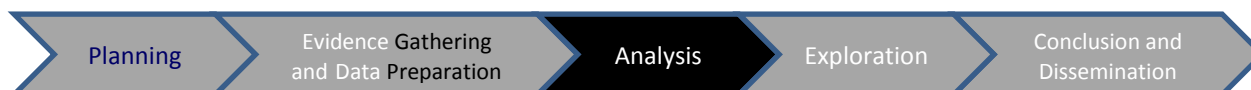| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Analysis: Summary of key points

- A full populated data table or a forest plot, if the plot can accommodate all of the benefits and risks, should be the starting point for a benefit-risk analysis.

- The type of analysis is determined by the scope and complexity of the data, as well as the decision maker's judgement regarding the weight of evidence in the data table. Specifically, the decision maker should consider whether a *qualitative analysis* based on inspection of the data table is sufficient to enable a justifiable benefit-risk decision, or whether *quantitative methodologies* would strengthen the decision by providing a transparent, integrated measure of the benefit-risk balance.

- Data are analysed to quantify the magnitudes of benefits and risks for the drugs of interest. Depending on the purpose and context of the benefit-risk assessment, the benefits and risks in addition may be weighted and combined to provide a single integrated measure of the benefit-risk balance.

- Weighting is an exercise in which a judgement is made regarding the relative importance of the criteria in the data table.

  o Implicit weighting (i.e., where the decision maker does not have to quantify the difference in importance of specific benefits and risks) is adequate for simple cases where the following conditions are satisfied: (1) Only a small number of benefit and risk criteria (or sets of criteria) are to be weighed against each other; (2) Only a small number of comparators are involved; and (3) The difference in importance between the criteria being traded off is intuitively clear. However, if any of these three conditions are not met, the cognitive strain involved in implicit weighting becomes overwhelming.

- When implicit weighting is inadequate, quantitative modelling preference between options can facilitate an integrative approach and assist decision making.

- Even if a qualitative analysis is sufficient, there are reasons why a decision maker might wish to incorporate preference weights and express the benefit-risk balance in quantitative terms. Explicit weighting can increase the transparency of the decision process, help to ensure consistency with other decisions, and establish priorities for the development of new treatments. Expressing the benefit-risk balance numerically can better facilitate the sensitivity analysis, helping to ensure that decisions are robust.

- Quantitative methods require technical expertise and are generally more demanding in terms of resources than qualitative methods.

- When deciding whether to use quantitative methods and models, decision makers must consider the intended audience for the assessment and to whom it will be communicated, and adjust the outputs from the assessment to meet the needs of the audience.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

- Three types of methodologies were evaluated for their usefulness during the Analysis stage:

    - Metric Indices
    - Quantitative Frameworks
    - Utility Survey Techniques

- Though commonly known and used for benefit-risk assessment decision making, especially by clinicians, the metric indices were generally found to be inadequate for the purposes of benefit-risk assessment.

- Some complex problems, e.g., involving weighted criteria to be appraised for several alternatives, were addressed by using decision making tools derived from so-called Multi-Criteria Analysis (MCA) methods. One of these, MCDA, extensively described by Keeney and Raiffa (1976), was tested in all Wave 1 and most of the Wave 2 case studies, and was found to be efficient for this purpose. SMAA was also tested in some case studies with positive results.

- Of the four Utility Survey Techniques identified by PROTECT WP5, Discrete Choice Experiments (DCE) was recommended to be applied in the PROTECT WP5 case studies.

- PROTECT WP5 recommends holding a decision conference (i.e., a meeting amongst the individuals representing key stakeholder views) to discuss and arrive at consensus judgements of the importance of the benefit and risk criteria, and to organise the process of eliciting preference values for a benefit-risk analysis.

    - It is recommended that the set of stakeholders selected for the decision conferences is wide enough to accommodate the views of the different parties (e.g., regulators, physicians, patients).

    - Because they can inform which criteria to include, as well as their relative importance, where practical, it is recommended that the perspectives of patients be included in the benefit-risk assessment.

- The use of appropriate visualisations can expedite stakeholders' response times when reading outputs from a benefit-risk analysis. Therefore, it is recommended that graphics be well designed for the intended audience (i.e., additional work may be required to design effective graphics for communicating benefit-risk information to an audience with perceived low numeracy skills (Dolan et al., 2012)).

- A table or forest/interval plot should be used to present the results of a qualitative or partially quantitative analysis. In particular, tables provide fast and efficient readability across issues displayed in rows and columns. Tables can:

    - serve as a common means for benefit-risk communications because of their simple structure, flexibility, and the ease with which they can be adapted

    - be very powerful as a communication tool whilst conveying a substantial amount of information

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

- o   be used when communicating benefits and risks to all audiences, including the general public, mass media, patients, physicians, regulators, and other experts such as analysts

- o   ease cognitive burden of the users and decrease the time required to extract the information

- o   be limited to the requisite number of rows and columns to avoid adding cognitive burden when reading tables

- o   have their readability enhanced through the use of colour-coding to represent grouping and relationships

- The use of a value tree is recommended to communicate the structure of the decision problem to the stakeholders for all elicitation methodologies.

  - o   For preference elicitation, the use of visualisation methods can reduce cognitive burden on the stakeholders, ensure proper understanding of the message, and engage stakeholders in the elicitation process.

- It is recommended to use a bar chart to present the results of an integrative benefit-risk assessment.

  - o   Bar charts (e.g., simple, stacked, grouped) can be used to communicate the magnitude of any measure (e.g., benefit-risk score, probability of outperforming a comparator), and stacked bar charts can be used to depict different parts or components (proportions).

  - o   Bar charts can be used to display the benefit-risk trade-offs and to compare between options.

  - o   Bar charts (e.g., the Hiview3 difference display) can be used to indicate the difference in benefit and risk scores between two treatments.

    - ▪   This is a particularly important visualisation because of the comparative nature of the benefit-risk assessment.

    - ▪   The difference display has been found to be a very useful visualisation of benefit-risk balance in regulatory decision making (European Medicines Agency, 2012).

  - o   Bar charts may be a useful visual communication tool for a large variety of audiences, such as the general public through the media, patients, physicians, regulators, and other experts because of the simplicity in their design concept.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

- It is recommended that simple or stacked bar charts are used to visualise stakeholders' value preferences (i.e., the weights of the benefit and risk criteria) and the magnitude of the final benefit-risk metrics when a quantitative benefit-risk assessment approach is used.

- The importance of presenting the preference weights should not be overlooked as this information is key to interpreting an integrated benefit-risk metric.

- It is recommended that stacked bar charts, difference displays, or grouped bar charts are used to visualise the contributions of the different benefit and risk criteria in the benefit-risk analysis.

- The use of an interactive table display is recommended in the application of MACBETH to ensure the categorical judgements from stakeholders are consistent across the different criteria.

  - The use of an interactive thermometer scale may be useful to visualise the elicited preference values, allowing users to fine-tune the preference values within their consistency bounds.

- For weight elicitation, thermometer-like vertical sliders may be used; and interactive sliders may allow direct comparison and weighting of benefit and risk difference criteria on a pre-determined scale.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

## Part 4: Exploration

Subsequent to the main analysis, the results need to be assessed for robustness and sensitivity to the various assumptions, divergent views, and sources of uncertainties. Decision makers have different attitudes to uncertainty, and the nature of the key uncertainties should be explained as clearly as possible to enable an informed decision to be made. Since many of the inputs in a benefit-risk assessment, such as clinical data values and preference weights, are subject to uncertainty, it is important to assess whether the strength of the conclusion is affected by these uncertainties.

It is also important to explore further the consequences of a decision, and consider whether the results of the benefit-risk assessment may inform related decisions on risk management plans (RMPs) or benefit-risk assessments of similar medicinal products.

Statistical and modelling expertise are key resources at the Exploration stage, though clinical knowledge is still important.

This section addresses the following key questions relating to the Exploration stage of a benefit-risk assessment:

How does uncertainty arise in a benefit-risk assessment?

How does uncertainty affect the analysis results?

How does uncertainty affect the decision that needs to be made?

What related decisions might also be affected by a benefit-risk assessment?

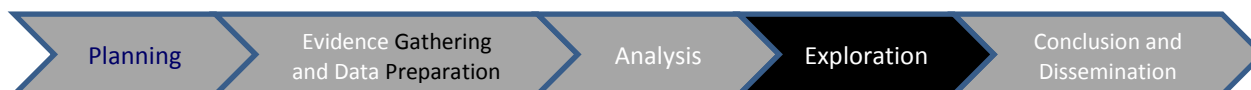Which of the methodologies tested by PROTECT WP5 can help with the Exploration stage?

What are PROTECT WP5's recommendations regarding the use of specific methodologies at the Exploration stage?

What are PROTECT WP5's recommendations regarding the use of visualisations at the Exploration stage?

### How does uncertainty arise in a benefit-risk assessment?

Published studies in the medical literature tend to report measures of uncertainty only using confidence intervals and p-values, which relate to sampling error only. There may be other sources of uncertainty that may affect the study results. Advanced statistical modelling may be able to account for some of these uncertainties but not for all. In any case, uncertainty should be addressed and not ignored.

Factors that can contribute to the uncertainty of results in a benefit-risk analysis include the following:

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

- Mismatched populations
  - between and within medicinal effects data, elicited preferences, and target population
- Sampling error
  - the population in which the medicinal effects data or elicited preferences came from does not reflect the population for which the decision is to be made
- Within-population variability
  - certain data, e.g., the average results in a study population, may not be generalisable to different subgroups of the target population
- Missing data (e.g., data were not collected or observed on a particular benefit or risk,  lack of comparator data in long term extension clinical studies, poorly recorded data that cannot be used for analysis, inconsistent or varying definitions of clinical outcomes or endpoints)
- Unresolved disagreements amongst assessors (e.g., choice of criteria, criterion weights,  appropriate comparators, data sources, preferences, and interpretation of results)
- Bias (systematic error in a model or process) can arise in many areas of a benefit-risk assessment, e.g.:
  - in medicinal effects data (e.g., due to dropouts or treatment crossover in clinical trials, or selection bias or confounding in observational studies)
  - in elicited preferences (e.g., due to poorly framed questions)
  - between comparators due to the structure of the benefit-risk decision model (e.g., due to inappropriate assumptions, or key benefits and risks having been omitted)

Some sources of bias that contribute to uncertainty in a benefit-risk assessment could be minimised by designing the benefit-risk assessment carefully and using data from highly relevant and validated sources.  However, one cannot guarantee that bias has been completely eliminated.
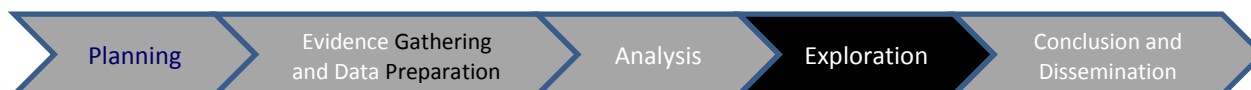
With access to sufficient patient-level data, more complex methods of analysis can be used to reduce the problems presented by mismatched populations and within-population variability.  For example, the Wave 2 warfarin case study constructed a model to predict the benefit-risk balance for patients based on their individual characteristics.


## How does uncertainty affect the analysis results?

*Sensitivity analysis* is a well-established technique for exploring the robustness of a model.  It is widely employed in a number of fields.

Sensitivity analysis aims to uncover the extent to which changes in the inputs of a model affect the results.  This is achieved by re-running the model using different input scenarios and observing the change in the results.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

Although full sensitivity analysis involving multiple outcome effects is possible and encouraged, in practice it is likely to be limited to a small number of scenarios.  This is because the amount of information to be presented would otherwise be overwhelming.  It is with quantitative benefit-risk models that uncertainty analysis really shows its value, as it is not necessary to present every simulated scenario but simply summary statistics that convey the variability of the benefit-risk balance. This allows systematic evaluation and presentation of a more comprehensive range of possibilities.   An example of this approach is the probabilistic simulation method (PSM), which is aimed at revealing the entire distribution of possible values of the benefit-risk balance.

Presenting an uncertainty analysis is central to ensuring proper interpretation of the results of decision models.  It serves as a reminder that the "single number" output of quantitative models is dependent on multiple uncertain inputs and encourages decision makers to consider whether that uncertainty is enough to cast doubt on the results and how it affects the decision.

## How does uncertainty affect the decision that needs to be made?

Attitudes to uncertainty may vary amongst stakeholders.  What this means is that the impact of uncertainty on a benefit-risk assessment depends not only on the extent of the uncertainty but also on the perspective that is adopted.

Attitudes to uncertainty vary between individuals for reasons that may not always be tangible. For example, it has been observed that assessors tend to become more risk-averse as they become more experienced; and that female assessors are more willing to take risks than males (Beyer et al., 2013).  Attitude to uncertainty is therefore one of the main reasons for disagreement within groups, and this can be a barrier to transparent decision making. Exploration of uncertainty cannot necessarily resolve all such problems, but it can help to ensure that the nature of any uncertainty is better understood so that the issues can be tackled in a logical way.

Attitudes to uncertainty in benefit-risk assessment are also subject to external influences.  For instance, major safety concerns about any individual drug may lead to a more cautious regulatory climate and thereby impact upon the assessment of seemingly unrelated treatments.  A stark example of this is the thalidomide scandal, which led to significant changes in regulatory requirements for medicines, such as the introduction of the Kefauver-Harris Drug Amendment Act in the USA.  The attitude of patients and the public may also have an effect, e.g., pressure from patient groups resulted in a licensing change for the use of natalizumab for multiple sclerosis in the EU.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
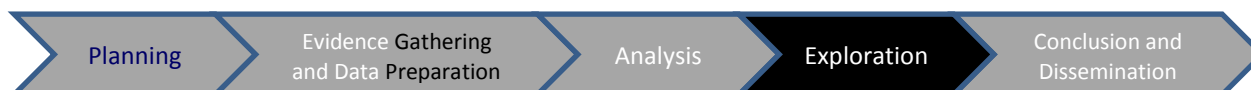
## What related decisions might also be affected by a benefit-risk assessment?

The benefit-risk methodologies reviewed by PROTECT WP5 are aimed at increasing the transparency of benefit-risk decision making.  One of many advantages of increased transparency is that it can help to ensure that related decisions are made on a consistent basis.
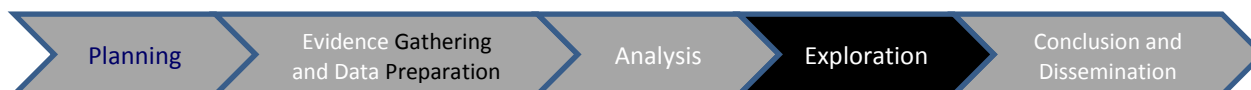
Benefit-risk assessments can be considered to be related if they share one or more elements; e.g., two assessments may have treatments, benefits and risks, target populations, data sources, or methodological approaches in common.   A decision maker who visits any particular concept in more than one benefit-risk assessment should consider the consistency of the approaches that were chosen in each instance and be prepared to justify any differences on the basis of the decision context.

For instance, both the Wave 1 efalizumab case study and the Wave 1 natalizumab case study shared the key risk criterion of progressive multifocal leucoencephalopathy (PML), and both case studies considered the regulatory question of whether the treatment should be given marketing authorisation. The weight given to PML in each of their MCDA decision models was significantly higher for natalizumab (55.9%) than efalizumab (12.8%), which might suggest inconsistency. However, caution should be exercised in comparing weights or weighted preference values between different methods or models.  The reason is that no modelling approach, with the exception of QALYs or DALYs, uses a unit of utility that stays constant across different models; the scales are defined locally for a given model, not globally across all models.  In addition, the number of criteria and scale ranges can affect the values of normalised weights.

What could be compared across two separate MCDA models are *ratios* of weights for criteria held in common between the models.  For efalizumab and natalizumab, only PML is held in common; there is not a second identical criterion.  However, potentially life-threatening outcomes, for different reasons, are common to both drugs: transaminases elevation (which could indicate liver damage) in natalizumab and serious infections (requiring hospitalisation) in efalizumab.  The ratio of PML to transaminases elevation is 55.9 ÷ 11.2 = 4.99, and the ratio of PML to serious infections is 12.8 ÷ 2.6 = 4.92.  The near-identical ratios could be a coincidence but might also suggest a degree of consistency between the judgements of the two teams that modelled these drugs.

Another example relates to the merging of new drugs/competitors to the market, as in the Wave 2 warfarin case study. Initially, the benefit-risk balance of warfarin was deemed positive against no treatment. However, it is not equally clear that the benefit-risk balance of warfarin would have been deemed positive if the assessment were to be made today against newer drugs.

Benefit-risk assessments also play a part in certain key regulatory requirements for pharmaceutical products.  For example, in the post-marketing setting, the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) has recently published guidance recommending that the

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

Marketing Authorisation Holder (MAH) should submit a Periodic Benefit-Risk Evaluation Report (PBRER) to the regulatory authorities.

Another key regulatory document is the Risk Management Plan (RMP), which identifies options to prevent, minimise, or mitigate the impact of risks related to the medicine. RMPs have a two-way relationship with the benefit-risk assessment: benefit-risk can help to identify and prioritise the key risks to be managed, and the RMP, particularly the assessment of the effectiveness of risk minimisation, can help to refine estimates of the impact of those risks in future assessments.

Benefit-risk assessments may also influence the information that is recorded on Company Core Data Sheets (CCDS) for pharmacovigilance purposes.

### Which of the methodologies tested by PROTECT WP5 can help with the Exploration stage?

It may not be necessary to employ additional methodologies for exploration purposes. The Exploration stage may simply involve revisiting earlier stages of the assessment and performing sensitivity testing using the same model with different scenarios. Alternatively, the effect of using different methodologies could be deliberately explored. The PROTECT WP5 cases were chosen because of their complexity which invoked the application of the more quantitative methodologies.

A few of the methodologies discussed in the previous sections have features that facilitate exploration in that they can help to quantify uncertainty:

- Indirect/Mixed Treatment Comparison (ITC/MTC)
- Utility survey techniques
- Probabilistic simulation method (PSM)
- Stochastic Multi-criteria Acceptability Analysis (SMAA)

### What are PROTECT WP5's recommendations regarding the use of specific methodologies at the Exploration stage?

#### *Indirect/Mixed Treatment Comparison (ITC/MTC)*

ITC/MTC provides a principled framework for propagating source data through Bayesian data networks that may be required for comparative benefit-risk assessments. The method offers increased transparency in terms of clarifying the sources of evidence used and the extent to which uncertainties in the source data translate to uncertainty in the

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

benefit-risk assessment metrics. The statistical uncertainty associated with using indirect evidence is greater than when direct evidence is used, and ITC/MTC makes allowance for this.

The method's ability to compare treatment options allows a wider range of comparators to be considered, which may be another aspect of the analysis to be explored at this stage.

The section on Evidence Gathering and Data Preparation includes further discussion of ITC/MTC.

## Utility survey techniques

The Wave 2 rimonabant case study used a Discrete Choice Experiment (DCE) to elicit preferences from the public. Analysis of the DCE provided estimates of the variability of the preference weights. This approach could be used in tandem with a quantitative benefit-risk assessment in order to explore the uncertainty related to preference weights and has the advantage that the range of weights explored is based on experimental evidence (compared to the standard SMAA approach whereby all possible combinations of weights are explored). It may also be possible to quantify the uncertainty of value preferences using other utility survey techniques, such as the Analytic Hierarchy Process (AHP).

Further discussion of the utility survey techniques can be found in the Analysis section.

## Probabilistic Simulation Method (PSM)

PSM can be used in conjunction with other quantitative benefit-risk models to explore the impact of uncertainty in input data on the final benefit-risk balance. It can be based either on theoretical distributions for the uncertain input parameters (Monte Carlo simulation) or a re-sampling from individual patient data (if such data are available).

PSM has two important advantages compared to a standard sensitivity analysis: The use of individual patient data or appropriate probability distributions can ensure that the scenarios explored are realistic; and it also provides a more complete picture of the variability of the results. In summary, it is a sound method from a statistical point of view. The danger is that it can lead to overconfidence in the results, in particular because it is unlikely that all sources of uncertainty can be accounted for in PSM. In other words, the fact that one source of variability has been thoroughly explored using PSM does not mean that any other source of uncertainty can be ignored. The probabilistic assumptions underlying applications of PSM should always be considered. For example, all applications in the PROTECT WP5 case studies assumed that the probabilities of benefits and risks were uncorrelated with one another. Such assumptions may not always be realistic.

## Stochastic Multi-criteria Acceptability Analysis (SMAA)

SMAA is an implementation of PSM within an MCDA framework, allowing for uncertainty of treatment effects data and preferences. The general comments above regarding PSM, therefore, also apply to SMAA. A particular strength

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
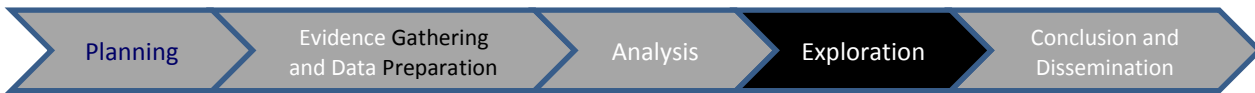
of SMAA is its ability to explore preference uncertainty with very little prior information, which may be a more straightforward approach than attempting to characterise preferences using a probability distribution.

## What are PROTECT WP5's recommendations regarding the use of visualisations at the Exploration stage?

The visualisations discussed below all have features that may be helpful when exploring and communicating the uncertainty of the benefit-risk balance.

### Box plot

Box plots (also known as box and whisker diagrams, or by further variants thereof) are used to convey statistical information about the range of values taken by a variable. Box plots come in many varieties (see Figure 20), but the basic principle in the same: the lengths of the different segments of the plot provide information about the spread of values and reflect the bias or skewness in the data. Outliers are sometimes presented as points away from the main box plot.  Due to the technical constructions of box plots, their use may be limited to experts or trained audiences who have some understanding of statistical summary measures (e.g., medians, means, quartiles, outliers); without this, there is a danger that the segments will be interpreted as having some other (perhaps clinical) meaning.

**Figure 20 The anatomy and variations of box plots (reproduced from Potter K. Methods for presenting statistical information: the box plot. 2006.)**
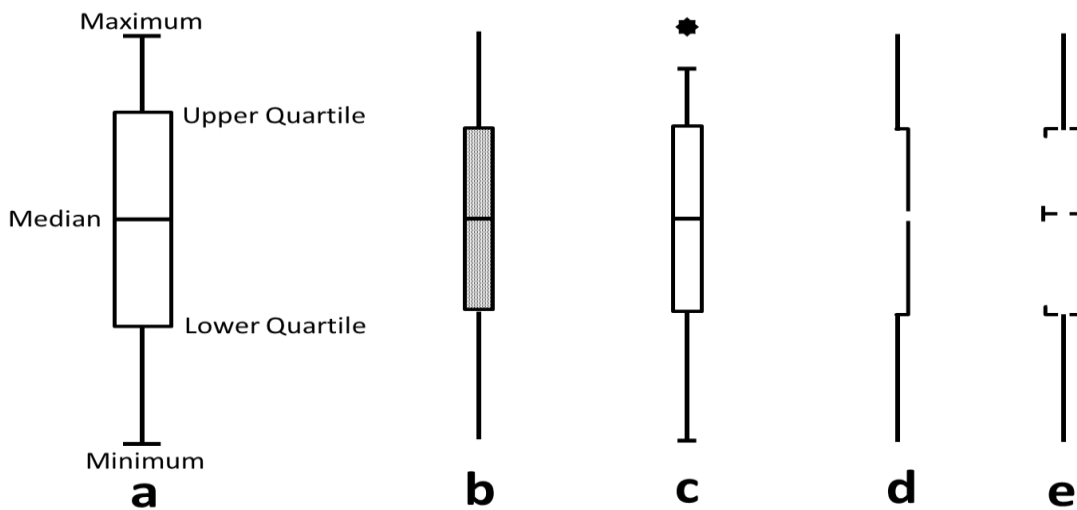


(a)   The anatomy of a box plot, (b) The range-bar chart, (c) The box plot, (d) the quartile plot, (e) the abbreviated box plot

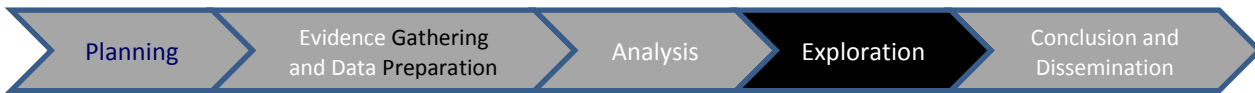| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|
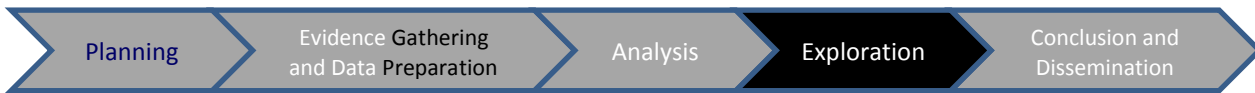
## *Distribution plot*

Distribution plots are the only widely used graphics that display the shape of an entire probability distribution. As such, they contain more information than simpler graphics like box plots or interval plots. For this reason, distribution plots must be based on large datasets; typical uses include showing the range of values of a measurement taken in a large group of patients or summarising the thousands of iterations produced by probabilistic simulation models.

Although these diagrams may be appealing to specialist audiences, they may be less helpful for those without the necessary technical background.

Distribution plots can be useful when making comparisons between treatments or groups as they convey a sense of statistical significance. However, box plots and interval plots also share this property to some extent and are arguably simpler to generate and communicate to a wider range of audiences.

**Figure 21 Distribution plot showing the distribution of the benefit-risk score for four treatments: from the Wave 2 natalizumab case study**

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
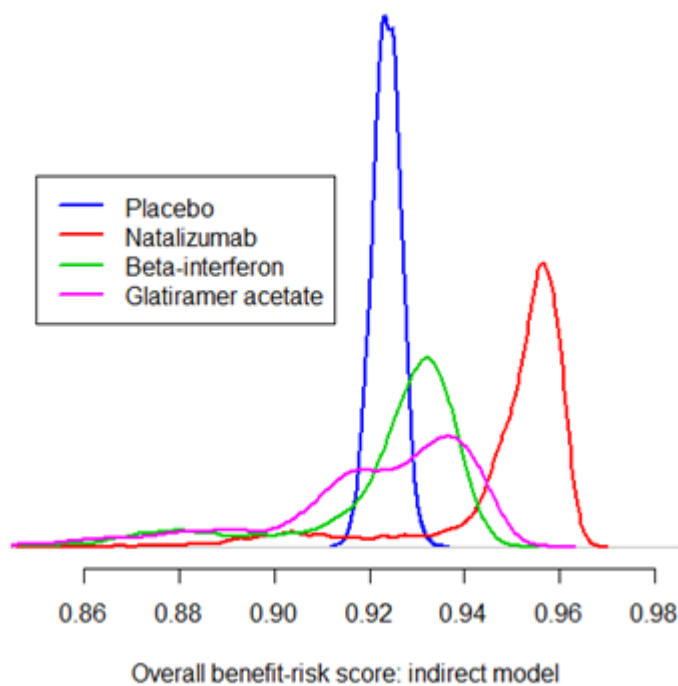
## Forest/interval plot

Forest/interval plots provide a sense of the distribution of a range of values without all the detail of distribution plots. By convention, forest/interval plots show means and 95% confidence limits, which are arguably more relevant to medical decision making than the measures of spread shown by a box plot. Other measures could be used if desired.

Forest/interval plots are recommended in situations where it is important to visualise the mean effect sizes and confidence intervals of two or more criteria alongside one another. One example is where benefits and risks are not integrated but simply presented side by side, as in Figure 22, which shows a confidence interval for each benefit and risk, aligned according to which treatment is favoured and with the neutral line of "no effect" clearly marked. However, this plot can only be produced if all the benefits and risks are expressed on the same scale (in this case, the risk difference between two treatments).

**Figure 22 Forest plot showing the difference in risk per 1000 patients in using a constructed triptan vs. another constructed triptan for treating acute migraine (reproduced from Levitan et al., 2011)**
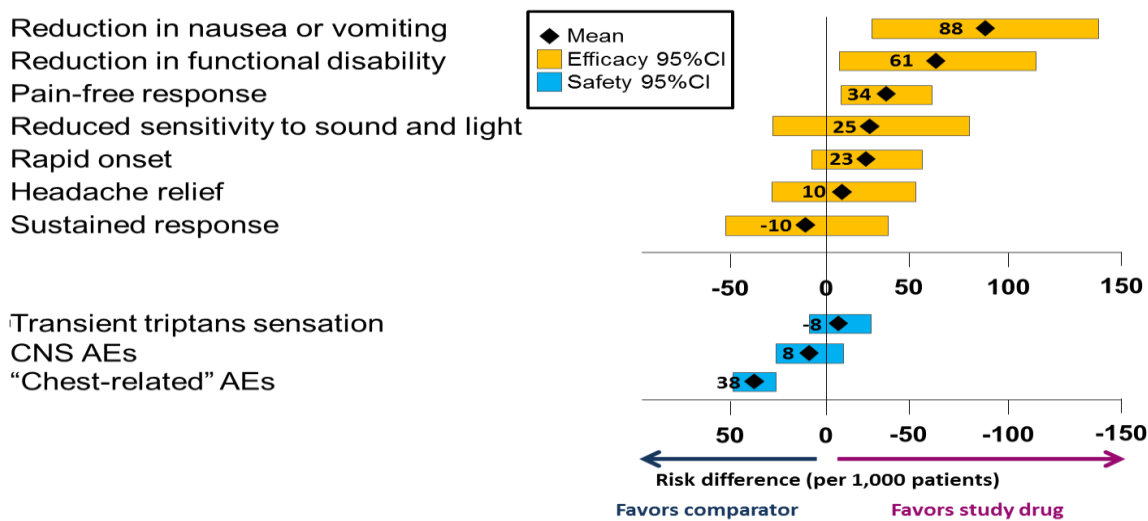


One possible drawback of this kind of forest plot was noted by the Wave 1 telithromycin case study team: "Wide variability, i.e., a long bar in the forest plot, can tend to overemphasise a less important variable if the data are not weighted." Furthermore, presenting the benefits and risks alongside each other on the same scale could be taken to imply equal weighting. The Wave 2 warfarin case study addressed this issue to some extent by ordering the benefits

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
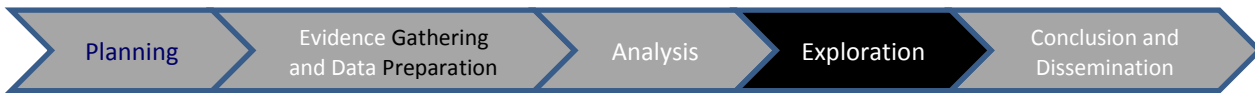
and risks according to elicited preference weights (Figure 23); the Wave 1 natalizumab case study team suggested taking this a step further by incorporating the weights in the labels.

**Figure 23 Forest plot illustration of the difference in consequence using risk difference per 1000 patients per year for the Wave 2 warfarin case study. The criteria are listed in order of importance (highest rank at the top)**
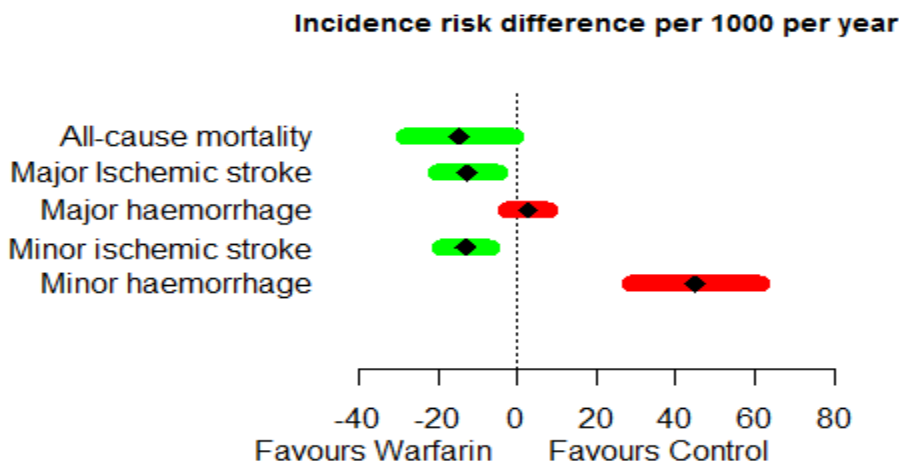


*Tornado diagram*

A tornado diagram is a specialised bar chart that is designed to show the results of a sensitivity analysis. An example from the Wave 1 natalizumab case study is shown in Figure 24. For each benefit or risk, the width of the red and green bars shows the effect on the overall benefit-risk score if the treatment effect is changed from the central estimate (shown under the bars) to a higher or lower value (shown beside the bars).

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
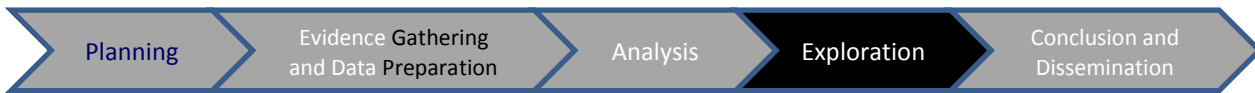
**Figure 24 Tornado diagram illustrating sensitivity of the benefit-risk score to change in treatment effects data in the Wave 1 natalizumab case study**
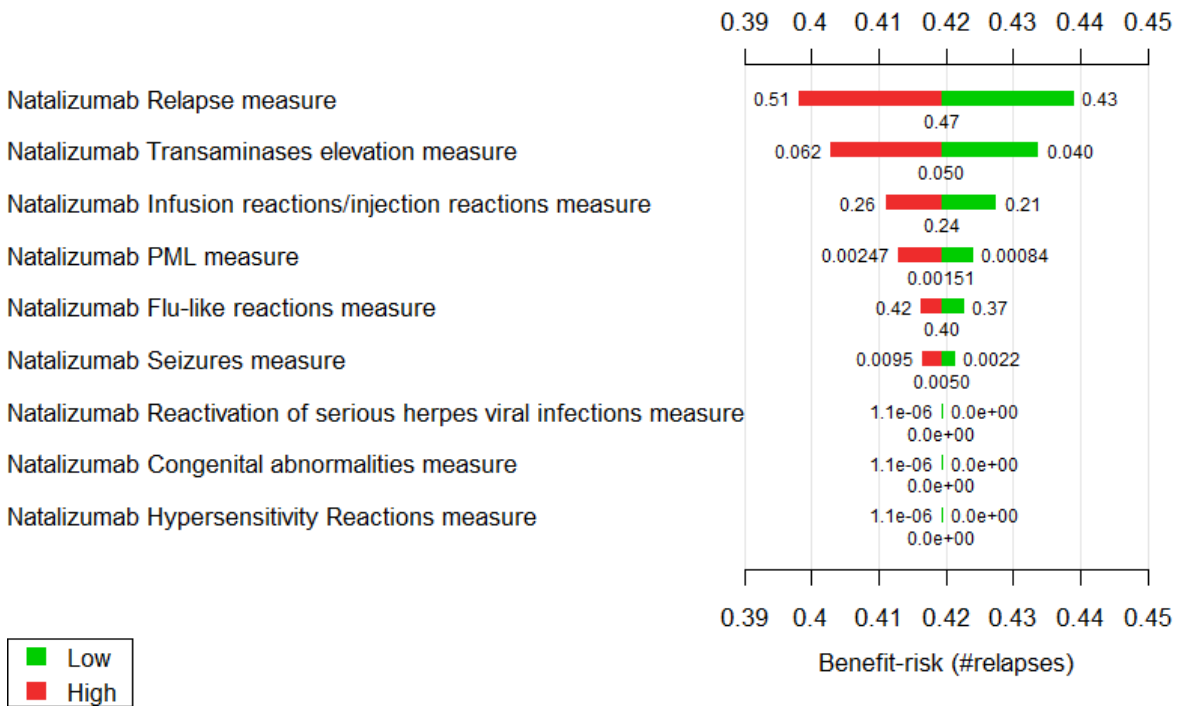


Interpretation of tornado diagrams can be confusing. Audiences who are unfamiliar with the technical details can easily misunderstand their purpose and intended message. Even with the right technical knowledge, it is difficult to gain a sense of the overall uncertainty of the benefit-risk balance based on the information provided in Figure 24. However, the diagram does effectively convey which criteria are most fundamental to the benefit-risk balance.

### Scatter graph

Scatter graphs are simple depictions of multiple points in two dimensions. Like distribution plots, they are generated from large datasets, but each point encodes two variables instead of just one. This makes scatter graphs useful for examining the relationship between variables. For example, Figure 25 compares two outputs of multiple simulation runs in the Wave 1 telithromycin case study.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
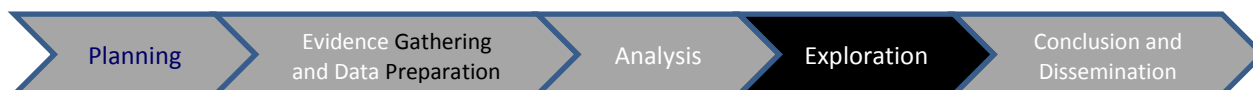
**Figure 25 Scatter plot of incremental harm versus incremental benefit from PROTECT WP5 Wave 1 telithromycin case study**
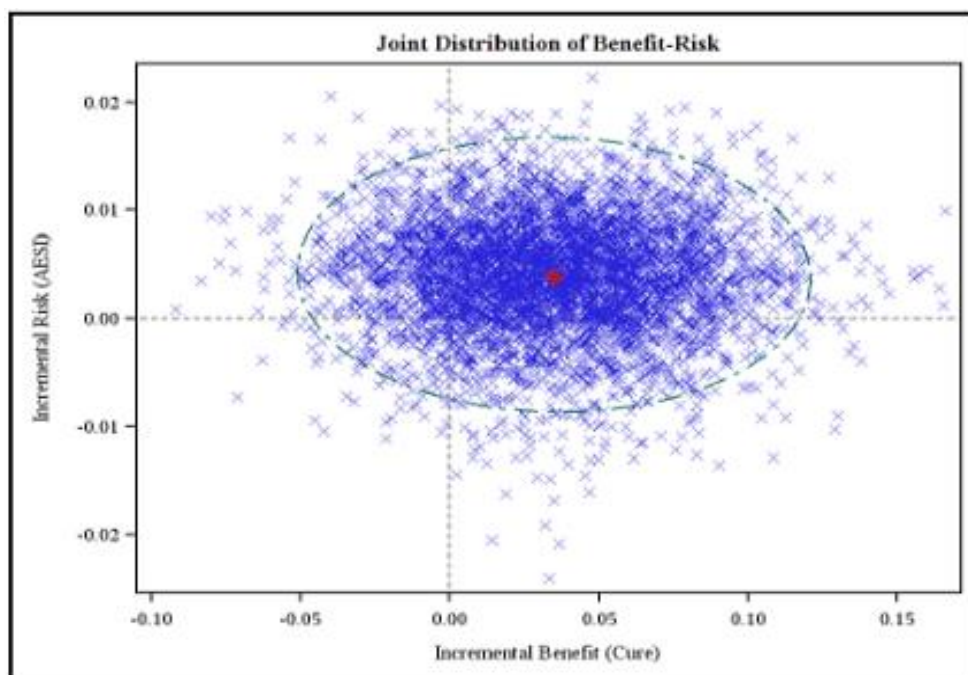


## Interactive visualisations

Interactive visualisations on a computer screen provide a user with the ability to change key parameters and observe the effect on the output of a benefit-risk model. These visualisations, although technically demanding and therefore not widely used, have great potential as part of the Exploration process. Interactive visuals enable active participation of the audience, which can increase attention and perception. Through interactive visualisations, it is possible to personalise the information communicated, by allowing the audience to investigate various aspects of a problem that they consider important (to personal decision making) or to explore areas which are still unclear in the primary visuals.

It is recommended that related interactive visualisations required to make a decision should be presented on a dashboard. A dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance (Few, 2004).

The Wave 2 rimonabant case study team developed several interactive dashboards: a screenshot from one is shown below (Figure 26), and the full dashboard can be found online.
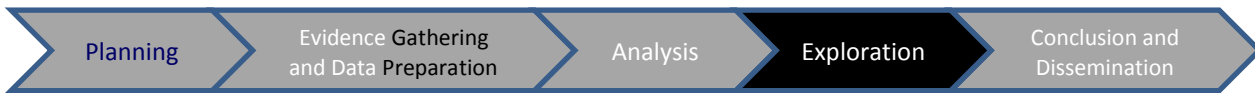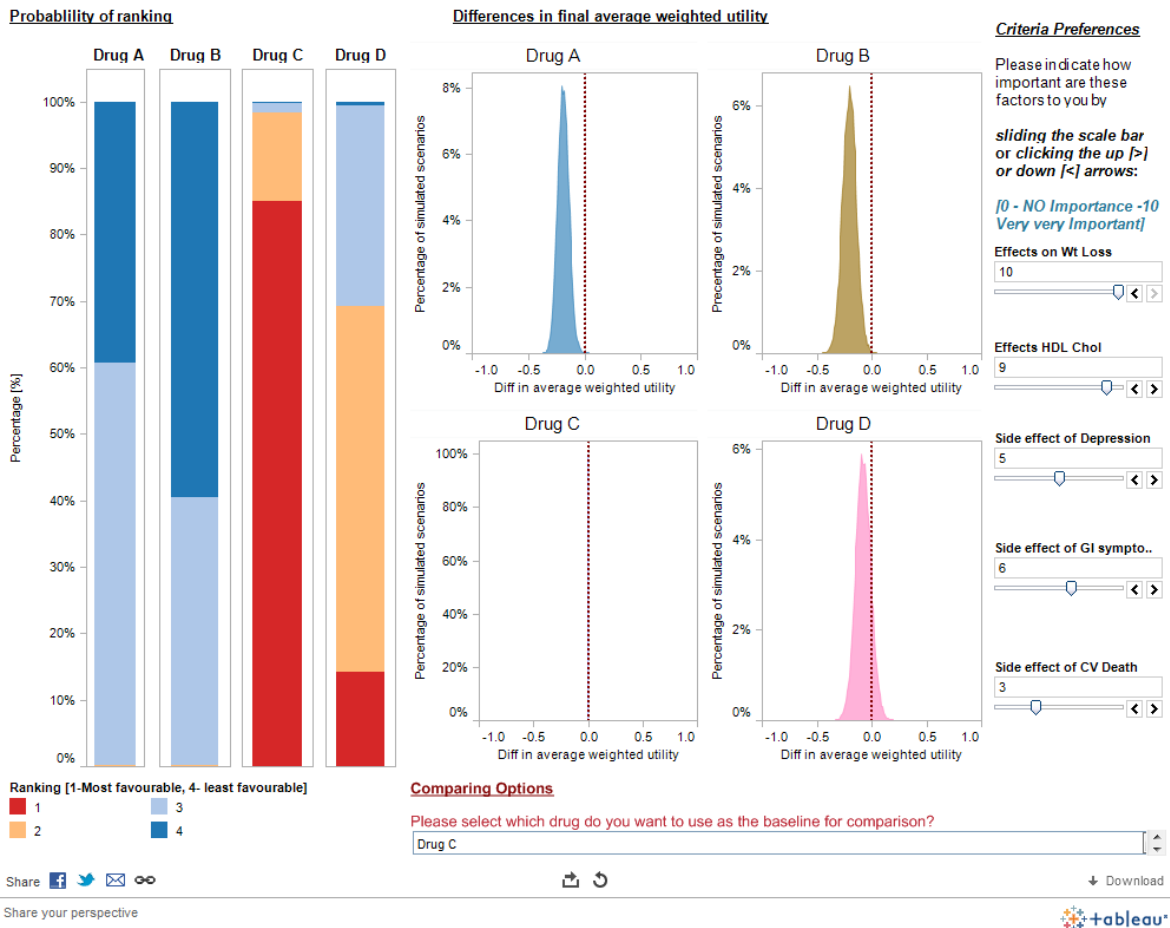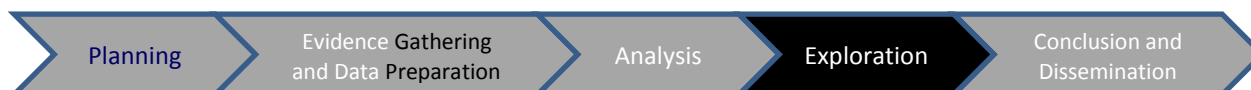
| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

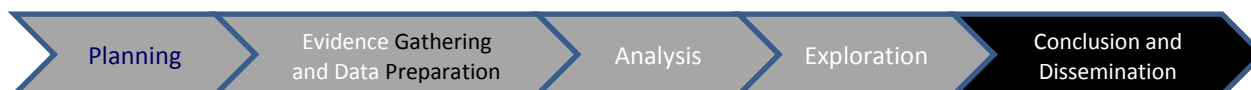**Figure 26 Interactive SMAA-based dashboard from the Wave 2 rimonabant case study**



The [Stage 2 Visual Review](#) (Mt-Isa et al., 2013b) contains some guidelines and further references on the creation of interactive dashboards.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|----------|----------------------------------------|----------|-------------|------------------------------|

## Exploration: Summary of key points

- Subsequent to the main analysis, the strength of the conclusion needs to be assessed for robustness and sensitivity to the various assumptions and sources of uncertainties.

- Presenting an uncertainty analysis is central to ensuring proper interpretation of the results of decision models. It serves as a reminder that the "single number" output of quantitative models is dependent on multiple uncertain inputs.

- PROTECT WP5 recommends that the appropriate statistical, modelling, and clinical expertise is engaged at this stage.

- Relative to the exploration of the benefit-risk analysis, PROTECT WP5 recommends that bias, mismatched populations, sampling error, missing data, and disagreements amongst assessors, factors which contribute to uncertainty in the conclusions, are acknowledged and factored into the final conclusion of the benefit-risk assessment.

- It is recommended that decision makers explore how the uncertainties affect the benefit-risk balance by applying sensitivity analysis, such as repeating the analysis using the same model with different sets of input values. Other methods are Indirect/Mixed Treatment Comparison (ITC/MTC), Probabilistic Simulation Method (PSM), and Stochastic Multi-criteria Acceptability Analysis (SMAA).

- The following visual types were found to be useful in the Exploration stage: Box plot (box and whisker diagram), distribution plots, forest/interval plot, tornado diagram, scatter graph, and the interactive versions of these visualisations.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

## Part 5: Conclusion and Dissemination

The final stage in the benefit-risk assessment process is the point at which a conclusion is reached and the results and consensus are communicated to a wider audience. This last stage makes it explicit that the findings of the benefit-risk assessment have logically led to a conclusion that could influence future actions. It emphasises the need for a transparent audit trail of the whole assessment process from the Planning stage to the Exploration stage. In other words, this last stage of the process brings everything together and sets the stage for action to be taken.

By this point, it is easy for those who are closely involved in the benefit-risk assessment process to have become so absorbed in the details of the analysis that they lose sight of the bigger picture – as the saying goes, they may find themselves "unable to see the woods for the trees."  It is helpful at this point to take a step back and consider, with a fresh pair of eyes, what the overall aims of the assessment are and whether the approaches used to achieve those aims have been adequately documented.  This may involve asking questions such as those set out below.

---

What is the benefit-risk assessment conclusion, on what was the conclusion based, and how is it documented?

What are the critical limitations which apply to the conclusion?

How is the conclusion communicated and to whom?

Has an audit trail been provided so that the benefit-risk assessment can be understood and reproduced by others?

What would be the trigger to re-evaluate this benefit-risk conclusion (i.e., when does the process get re-initiated)?

What are PROTECT WP5's recommendations regarding the use of visualisations at the Conclusion and Dissemination stage?

---

### What is the benefit-risk assessment conclusion, on what was the conclusion based, and how is it documented?

The decision maker should be able to answer the following questions:

- What question(s) was the benefit-risk assessment aimed at addressing?
- What answer(s) were found?
- Is/are the answer(s) highly sensitive to the treatment effects data, the choice of analysis method, or the preference data?
- What is the supporting information on which the conclusion is based?

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
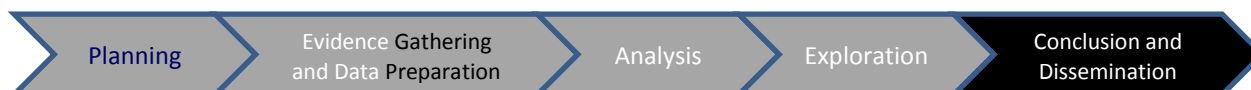
## What are the critical limitations which apply to the conclusion?

It is vital to document any known limitations of the benefit-risk assessment. Two limitations that were encountered in several of the case studies are discussed below.
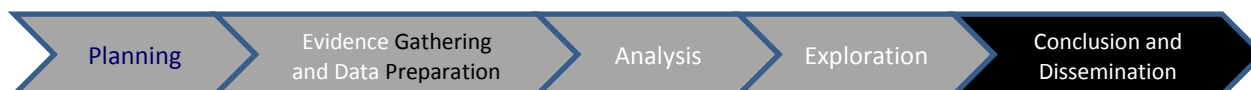
Lack of data was frequently cited as a limitation by the case study teams, who were largely restricted to using publicly available information. This meant that data could not be found on certain benefits and risks of interest for all comparators (a problem in both the Wave 1 efalizumab case study and the Wave 1 natalizumab case study), and the distribution of effects data within a population of patients could not be easily determined (with both the Wave 2 natalizumab case study and the Wave 2 rosiglitazone case study having to make "the best use possible of the statistical summaries to infer the population distributions"). The Wave 2 warfarin case study was unique in that patient-level data was used; however, there were issues with the quality of the data: "Discrimination between haemorrhagic stroke and ischaemic stroke in CPRD is limited as frequently non-specific codes are used."

Another frequently cited limitation concerned the generalisability of the results to real-world populations. For example, the Wave 1 efalizumab case study noted that "measures made on a clinical trial population may not reflect…off label use, misuse…in a post-marketing setting." Several case study teams questioned whether the preferences they had elicited could be replicated by real patients. The Wave 1 rimonabant case study report makes clear that the results are dependent on "the explicit weighting and utility function set by selected decision makers. This raises the question if the results can be applied in the wider population." Given more resources, it was generally felt that this problem could be overcome to some extent. The Wave 2 rosiglitazone case study team were aware that "the value functions and weights elicited for the MCDA model were based on the individual preferences of only a few medical experts in the PROTECT WP5 team" but pointed out that it would be possible "to involve more medical experts in the weighting process to generate a more representative set of weights and thus improve the model."

## How is the conclusion communicated and to whom?

Communication of a conclusion should never focus simply on the analysis results. It is equally important to present the evidence and assumptions on which a conclusion is based. This encourages transparent and robust benefit-risk assessments.

A particular difficulty in the field of benefit-risk assessment is that there are several conflicting versions of the terminology used to describe the concepts and methods involved. A particularly salient example is the word 'risk' itself. In asking 50 European assessors what they or their agency meant by 'risk,' the EMA's Benefit-Risk Project team collected over 50 different words or phrases, several of which were in conflict, e.g., "tolerance of a drug

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
|---|---|---|---|---|

compared to serious side effects," "severity of side effects," "frequency of side effects" (European Medicines Agency, 2009). To avoid confusion and misinterpretations, clear unambiguous language that avoids the use of technical jargon should be used wherever possible. Other examples include the use of "value" versus "utility," and the many alternative terms for benefit-risk assessment itself (e.g., benefit-harm).

The choice of language and visualisations to use when communicating benefit-risk decisions should reflect the level of technical knowledge of the intended audience. The appropriate channels for communication may also vary. As yet, there is little consensus regarding how best to communicate benefit-risk models. PROTECT WP5's Visualisation Reviews (Mt-Isa et al., 2013a; Mt-Isa et al., 2013b) have identified various graphics and classified them based on their complexity and ease of interpretation, but these have not yet been tested out on real-life audiences.

One role for a benefit-risk model is to serve as a communication channel amongst different audiences. For example, a committee of assessors in a regulatory agency might create a model that could be used by the agency's approvals committee to test different perspectives about the clinical relevance of a new drug's effects before making a final decision. Post-marketing, that model could be used by a regulator's pharmacovigilance committee to see if new information tips the benefit-risk balance. Within a pharmaceutical company, a benefit-risk model might begin its life at Phase II and be elaborated as the product profile changes in response to scientific findings. In both of these examples, the model would be used by different people at different times in the drug's life cycle.

## Has an audit trail been provided so that the benefit-risk assessment can be understood and reproduced by others?

An audit trail of the benefit-risk assessment process is clearly desirable for regulators, who are publicly accountable for approving medicines and must communicate their decisions to the public and to drug developers. Companies making submissions to regulators also have a clear incentive to ensure that benefit-risk information is transparently presented; and we would argue that even benefit-risk assessments carried out purely for internal purposes should be documented with a clear audit trail. This may help to ensure consistency of related decisions, to facilitate revisiting assessments in the light of new information, and in the event that bad decisions are made, to unpick what has gone wrong.

Use of formal methodologies like those reviewed by PROTECT WP5 – especially the benefit-risk assessment frameworks – naturally helps to create an audit trail.

Continuous use of a model throughout its life would require periodic documentation as specific milestones are reached so that it can be understood and used effectively following each milestone. Presumably the model would remain in the proprietary ownership of the pharmaceutical company, but once the drug is approved, it could be desirable to place the model in the public domain, either by the company or by the regulator, so that the reasons for approving—or disapproving—the drug are clearly communicated to the public.
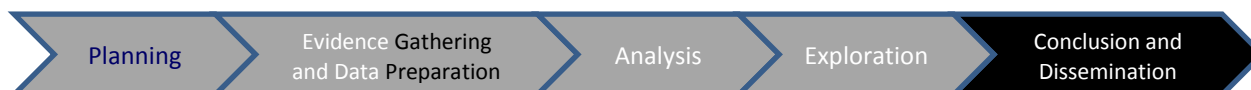
| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |

### What would be the trigger to re-evaluate this benefit-risk conclusion (i.e., when does the process get re-initiated) and what aspects of the assessment would need to be revisited?

During development of a medicinal product, the body of evidence regarding its effects and relevant indications changes rapidly. The benefit-risk balance may frequently need to be re-evaluated in light of any new information.

Updates may also be required from time to time after a product has been brought to market. For example, regulatory authorities typically require marketing authorisation holders (MAH) to submit periodic re-assessments of the benefit-risk balance (such as PBRERs).

In addition to periodic updates, there may be specific events in the post-marketing setting that trigger a repeat of a benefit-risk assessment, including the arrival of new comparators onto the market; safety signals or other emerging data; and overcoming the identified decision limitations.

#### Arrival of new comparators onto the market

The benefit-risk balance of a medicine is not considered in isolation but relative to alternative treatment options, i.e., the comparators. As time passes and new products are brought to market, it may be appropriate to expand the list of comparators to include any new drugs that are relevant to the indication. These may have side effects that were not included in previous benefit-risk assessments; in which case, a re-definition of the value tree may be necessary. Techniques such as ITC/MTC may be employed to aggregate evidence on the various treatments.

#### Safety signals or other emerging data

From time to time in the post-marketing setting, new evidence regarding a treatment's effects may become available. In some cases, the new evidence may simply provide revised estimates of the benefit and risk measures that have been included in a benefit-risk assessment; and in such cases, updating the assessment to reflect the new evidence should be fairly straightforward.

A more challenging scenario arises when previously unknown benefits or risks are revealed. For example, this may arise from long-term observational follow-up studies or spontaneous safety reports. Incorporating data from such sources into a quantitative benefit-risk assessment may not be straightforward. Spontaneous reports in particular may be perceived as unreliable, with potential bias in the reporting process and uncertainty regarding the extent of exposure to the treatment.

Revisiting a benefit-risk assessment in order to include additional outcomes may require some remodelling of the value tree for the reasons discussed earlier in this report. Any preference weights used in a quantitative analysis may no longer be valid in the presence of new outcomes and, therefore, will need to be re-elicited.

| Planning | Evidence Gathering and Data Preparation | Analysis | Exploration | Conclusion and Dissemination |
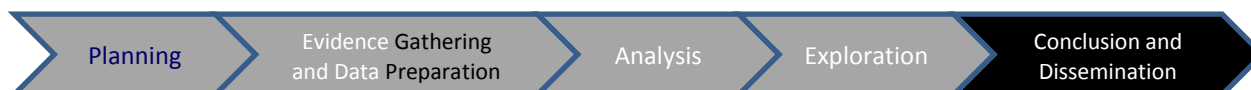
*Overcoming the identified decision limitations*

If any critical limitations of a benefit-risk assessment (e.g., a lack of relevant data) can subsequently be overcome, then clearly a more robust decision will be achieved by updating the benefit-risk assessment to reflect the latest information.

If there are any known limitations to the current benefit-risk assessment that could be expected to be overcome in the near future, such as lack of data on a particular risk, then these should be highlighted in the conclusion documentation as priorities for future study.

## What are PROTECT WP5's recommendations regarding the use of visualisations at the Conclusion and Dissemination stage?

The PROTECT project found that currently the use of graphics and visual aids for communicating the benefit-risk of drugs is very limited, and what they did find was mainly in scientific journals. Some patient leaflets illustrate procedures, such as placing a patch or injecting a substance, but only words are used to communicate benefit-risk. There are currently many initiatives in the field of risk visualisation, but these are neither specifically for visualising benefit-risk balance or trade-off, nor specifically linked to the benefit-risk assessment approaches (Cammax Limited, 2011; Gapminder, 2011; IBM, 2011; Spiegelhalter, 2010). Quantitative benefit-risk models provide many opportunities for displaying results in easily-understood graphic form, so we present here the major recommendations from our review displays that are relevant to benefit-risk.

The aspects to be presented depend on the audience's level of technical knowledge, as well as being dependent on their interests. Therefore, the first step in generating visuals is to determine the intended audience. It is difficult to say which stakeholders should be presented with which information, but a survey might be able to give some information on the average visual preferences.

It is recommended to consider Wickens' Principles of Display Design, which are principles for human perception and information processing, to aid a better design of visual displays for human use. The Wickens principles provide a set of 13 principles, which include principles to promote perception and attention, and principles based on that the individual interprets visuals based on existing experience and knowledge of a visual or the world. The principles are set out in full in the Stage 2 Visual Review (Mt-Isa et al., 2013b). The Wickens' principles are concretised in GlaxoSmithKline (GSK) Graphics Principles, and it is recommended to use these as guidelines when designing graphs to communicate numerical information. The GSK Graphics Principles can be found online (https://ctspedia.org/do/view/CTSpedia/BestPractices) and are also set out in the Stage 2 Visual Review (Mt-Isa et al., 2013b).

If possible, the use of interactive displays is also recommended.

105

# Section 3   Discussion

Over the course of the IMI-PROTECT project, PROTECT WP5 has carried out extensive academic and practical investigations into benefit-risk assessment methods.  Of particular note are the following achievements, where innovative approaches have been used to evaluate methodologies and communicate our findings:

- A new taxonomy of methods for benefit-risk assessment;
- An up-to-date review of those methods, covering both matters of principle and practical application;
- Comparative testing of methods on real-life scenarios;
- A review of visualisation methods, considering the application of graphical display principles to benefit-risk problems;
- Confirmation of the principle that quantitative benefit-risk modelling of medicinal products is possible and desirable;
- A review of the role of patient and public involvement in the benefit-risk assessment process including comparative evaluation of preference elicitation methods;
- Effective collaboration amongst pharmaceutical companies, regulators, and academics, working together in teams to arrive at a common consensus, ensuring a variety of viewpoints are represented.

There are a number of recent or ongoing initiatives examining the role of formal benefit-risk assessment methods, both in practical applications for organisations such as regulatory agencies, and common frameworks to be applied in multiple venues.  For example, the EMA benefit-risk project followed a similar approach to PROTECT WP5, with a review of existing methodologies and evaluation via case studies.  However, the scope of that review was somewhat more restricted, with a focus on EU regulatory processes (European Medicines Agency, 2013a).

The Unified Methodologies for Benefit-Risk Assessment (UMBRA) initiative has an altogether different focus, as it seeks to develop a standard platform for benefit-risk assessments.

The importance of adopting formal approaches to benefit-risk assessment is increasingly recognised by regulatory agencies, some of whom have already issued guidelines for benefit-risk assessment.  However, whilst we expect regulators to specify the requirements that must be met, we consider it unlikely that they will provide extensive guidance on how to meet those requirements.  We hope that these PROTECT WP5 recommendations provide practical advice that will play a useful role in bridging that gap.

It is worth emphasising that, although many of the tested methodologies are very useful, not every benefit-risk assessment will require advanced quantitative techniques.  Indeed, it is likely that the majority of benefit-risk assessments (particularly during the early phases of drug development) will be very clear cut, and a qualitative analysis of the benefits and risks will suffice.  However, there will be a minority of assessments that are extremely complex – along the lines of our case studies – and a more demanding analysis will be required.

A related point is that complex benefit-risk assessments can take a significant amount of time, perhaps more than working timelines allow. Therefore, it may be the case that it is not easy to incorporate the more advanced methods into current practice; however, with careful future planning, we believe such methods can be used to improve the transparency and generalisation of benefit-risk assessments.

Also, we recognise that our teams' experience at the start of using the methods was limited or non-existent; we learned as we applied the methods, and we found that the Wave 2 case studies were completed more expeditiously than the Wave 1 cases.  As might have been expected in using unfamiliar technology, experience in using the methods reduced the time it took to complete an assessment.

The case studies that aimed to model regulatory decisions were based on drugs that had already been brought to market, and as such, there is a possibility that the results were influenced by hindsight bias.  In other words, the analyses may have been based on a post hoc understanding of the evidence and could not have been used as the basis for real-world regulatory decisions.  However, by using data that would only have been available at the relevant time and making use of regulatory experience within the work package, we believe this bias has been been minimised.

Throughout PROTECT WP5's case studies, there are two limitations that have been encountered time and time again, regardless of the particular methodologies being employed.  These relate to the benefit-risk time horizon and the extent of publicly available data.  We highlight these issues again here in the hope that solutions might be found in future.

None of the methodologies are designed to quantify changes in benefits and risks over time (except perhaps some health indices for specific disease areas, e.g., QALYs, but even these handle time in a rigid, pre-defined way).  The standard approach is simply to focus on a particular time period of interest, and update the analysis when warranted by additional information or at designated time periods.  This approach is probably sufficient for many purposes, but it would be interesting to see if methods can be developed that explicitly model the dynamic nature of benefits and risks over time.

Our researchers were frequently frustrated by the lack of publicly available effects data, particularly at the individual patient level.  Changing this situation may require significant political will; the Wave 2 rosiglitazone case study team went as far as to recommend "that the European Commission investigate this issue of data availability and take steps to ensure that patient-level data about clinical studies of medicinal products are properly archived and made accessible."  Related to this is the problem of heterogeneity of the outcome measures reported in clinical trials.  Establishment of a standard reporting template that facilitates the extraction of data, including measures of uncertainty, would be a great step forward.

Most recently, there have been a number of steps forward with regards to data availability.  An Institute of Medicine workshop on sharing clinical research data took place in October 2012, and a summary of the findings are now

available.  The European Medicines Agency published a draft policy on the publication and access to clinical trial data in June 2013.   The European Federation of Pharmaceutical Industries and Associations (EFPIA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) launched their joint principles for responsible clinical trial data sharing in July 2013.  Efforts being undertaken by GlaxoSmithKline to provide access to anonymised patient-level data has also been recently published (Nisen and Rockhold, 2013).

## Challenging scenarios for structured benefit-risk assessment

PROTECT WP5 selected its case studies based on benefit-risk decisions that were of particular interest for exploring or testing systematic approaches to benefit-risk (Table 1) and (at least for the Wave 1 case studies) for which sufficient data were judged to be available.   However, this finite number of case studies cannot explore all possibilities, and new challenges beyond those seen to date will inevitably arise.  Therefore, we list below various challenges that we anticipate assessors may have to address, and have placed these under two sub-headings corresponding to the pre- and post-marketing stages of the product life cycle.  We hope that the discussion below may provide some useful guidance.

In general, the more that decision makers face difficulties such as those below, the more difficult it will be to derive reliable numerical estimates of the favourable and unfavourable effects of treatment.  The appropriate response to this depends upon the importance of the outcomes in question and the extent to which quantitative modelling is employed.  Assessors using quantitative decision models should take particular care to consider the impact of uncertainty of the treatment effects.  Those employing a more narrative/descriptive approach to benefit-risk should ensure that the sources of evidence, and any associated uncertainty, are clearly communicated.

### Pre-marketing / Licensing

### *Limited Evidence*

Before a treatment is marketed, many factors influencing its benefit-risk balance may remain unknown.  Early clinical studies may be too small in size, short in duration, or narrow in focus to capture data on all of the key benefits and risks that may eventually occur with chronic usage in the wider patient population.  An example of this is the use of metoclopramide for gastrointestinal disorders, nausea, and/or vomiting, which is associated with the emergence of tardive dyskinesia as an adverse reaction in the long term (EMA, 2013c).

This does not mean that there is no value in carrying out a benefit-risk assessment during the early stages of the product life cycle.  Benefit-risk assessment is always a dynamic process to be undertaken throughout the use of a treatment, rather than as a single determination.  A decision regarding the benefit-risk balance should be based on the best evidence available at that point in time and may later change as new evidence becomes available.

### *Relevance of outcome measures*

The outcomes recorded in clinical trials may not have been measured in a way that is optimal for the purpose of a benefit-risk assessment.  Surrogate measures may have been used in place of the long-term outcome measures of

interest.    In most cases, surrogate markers provide a good proxy but can result in less precision on the relation between the intervention and the primary outcome.  In cases where hard endpoints are needed, the evidence generated from such trials may not be sufficient.  There are also concerns over varying definitions and quality of measurements for different outcomes collected in clinical trials.  Where such concerns exist regarding the relevance of outcome measures, these should be documented carefully and fully so that the decision makers using this evidence can make informed decisions on the relevance of certain outcome measures. This documentation should also be revised and addressed in future periodic assessments.

## Post-marketing

### Long term follow-up data

Where trials have followed up participants beyond the original trial period, they can provide a useful source of data on a treatment's long term effects.  However, analysts and reviewers of a benefit-risk analysis will recognise that the controlled nature of a clinical trial breaks down at the end of the original study period, and data from that point on is more akin to that from an observational study.   Assuming the trial had a positive result, the control subjects will often have been switched to the active treatment after the end date, meaning that long-term control data may not be readily available.  Such extensions to comparative clinical trials may also encounter more issues with compliance and confounders.  As with any analysis, the sources and degrees of uncertainty, and their likely impact should be clearly documented.

### New evidence of efficacy and safety

If a company becomes aware of new efficacy or safety evidence relating to an approved indication for one of its products, it is obliged to consider this new information in terms of its impact on the benefit-risk balance (ICH, 2012), documenting this assessment, e.g., in the periodic benefit-risk evaluation reports (PBRER) to the regulatory authorities, and applying the appropriate risk minimisation measures such as labeling, as needed.

Where the new evidence has come from a study that is not sponsored by the company and does not fall under the EMA's clinical data transparency regime, only the published summary results may be available.  Integrating this information into a benefit-risk assessment based mainly on the company's own data may present challenges.  Meta-analytical techniques such as ITC/MTC may be required in order to allow for factors such as heterogeneity between study populations.  Bayesian modelling, which allows the distributions of summary data to be incorporated as prior information, may also be a theoretically sound and viable option.

Where the new evidence specifically relates to a different patient group from that for which the product was originally licensed, a separate benefit-risk assessment may be required for these patients.  This is not simply a case of changing the data in the existing assessment; for different patient groups, the decision context will vary and so the entire assessment should be revisited from the bottom up.   Assessors will need to consider whether it is appropriate to assume that the efficacy and safety profile is similar between the different groups.

*Observational / surveillance data*

Epidemiological studies, registry reports, and spontaneous reports may provide important data on emerging risks. As is true of clinical trials, there is potential bias associated with each source. The source of data can also be considered in terms of quality of evidence, e.g., CDC hierarchical system. Aggregating the evidence with that observed in clinical trials may also be problematic. Statistical methods may exist to deal with these issues, but this remains a relatively specialised field and not all assessors may have the resources for such approaches (or consider it appropriate to use such complex techniques for the decision at hand). Observational data will not contribute to the same extent in reducing uncertainty on benefit-risk balance as compared to randomised controlled trials. As has been noted elsewhere in this document, it is recommended that the complexity of the assessment be sufficient to answer the question of benefit-risk balance, with any limitations and sources of uncertainty appropriately noted, along with their potential impact.

*Well-established products*

Products with a long history on the market have the advantage of cumulative data. Information collected over time provides some of the answers to the questions described above regarding the impact of longer-term treatment. As noted, there are challenges regarding combining the data if a single quantitative database is needed. But a sufficiently flexible benefit-risk framework that accommodates multiple data sources offers the potential for a multi-faceted view of many aspects of the treatment and its favorable and unfavorable effects. As described for other treatments, the benefit-risk assessment of mature products should begin with robust framing to understand the questions that need to be addressed, followed by a consideration of the data sources that are appropriate to answer the questions, and the implications of including and excluding, or weighting other sources of data, e.g., those from other indications. As in other contexts, one of the advantages of using a benefit-risk framework is the transparency afforded around the construction of the analysis and the reporting of the results.

Another challenge with mature products is missing information. The regulatory paradigm was likely not as robust as it is today, resulting in less comprehensive documentation of evidence at time of approval. In addition, there are practicalities, such as the loss of archived information, that impact the ability to introduce data into a benefit-risk assessment. Some sort of sensitivity analysis may deal with this issue, but the best practices for this scenario are still evolving.

## Summary

In summary, we believe that the work of PROTECT WP5 has the potential to strengthen the monitoring of the benefit-risk balance of medicines via the application of its experience and recommendations on the integration and presentation of benefit and risk data. The scope of our review of benefit-risk methodologies makes what we believe is a unique contribution to other efforts. As with all such endeavours, it is naturally limited by the extent of the work package's resources and the timing of the review. It has not been possible to examine the entire universe of benefit-risk approaches in detail, and new methodologies and evolutions of the frameworks are constantly emerging. We nevertheless hope that these recommendations serve as a valuable guide for readers who are new to the world of benefit-risk assessment, as they highlight key issues and considerations that are common to many approaches.

# Section 4   Acknowledgements

# Section 5   Glossary

| Term | Description |
| --- | --- |
| Approach | The system of methods and principles used in a particular discipline |
| Aspect ratio | The ratio of the lengths of the two axes on a graph; a square graph has an aspect ratio of 1 |
| Benefit | The positive results of a given treatment for an individual or a population (i.e., efficacy, convenience, or even quality of life) |
| Benefit-risk assessment | An evaluation of medical product either quantitatively or qualitatively taking both benefits and risks of the product into account |
| Benefit-risk model | A formal way to analyse benefit and risk consequences and their balances from a set of actions and to aid making choices amongst actions when risk aversion and preferences are specified |
| Bias | The systematic tendency of any factors associated with the design, conduct, analysis, and evaluation of the results of a benefit-risk assessment to make the estimate of a treatment effect deviate from its true value |
| Clinical trial | A research study of a patient population to answer specific questions of medical interest through intervention |
| Cognition | The mental action or process of acquiring knowledge and understanding through thought, experience, and the senses |
| Confounding factors | Factors that affect the outcome together with other factors / predictors |
| Conjoint analysis | An umbrella term which refers to techniques that look at the individual contribution of attributes to overall value; such exercises may be ranking, rating, or choice based exercises |
| Criterion | A standard by which the performance of a methodology and the alternatives can be judged or decided |
| Effectiveness | The extent to which an intervention provides a therapeutic benefit when given under the usual circumstances |
| Efficacy | The extent to which an intervention provides a therapeutic benefit under ideal circumstances |

| Term | Description |
|---|---|
| Elicitation | The process through which relevant notions for a problem of interest are made explicit |
| Extension Clinical Study (Long Term) | The follow-up (long-term of 1-2 years) of short-term comparative studies, where the comparator's arm is abandoned (e.g., in some designs, the patients in the comparator's arm may move to the active medication arm) |
| Framework | A structured stepwise approach to perform a task |
| Graphical methods/ representation | The principles and procedures to present some numerical features or relations by a graph |
| Greyscale | The shades in the black and white spectrum with no other colours |
| Health technology assessment | An analysis of the medical, economic, social, and ethical implications of the incremental value, diffusion, and use of a medical technology in health care |
| Hue | The dominant colour; higher hue of a primary colour gives the perception that the object appears with the shades of that colour |
| Incidence | The frequency of the first occurrence of an event or a condition in a specified period |
| Line pattern | The look of a line which could be, e.g., solid, dash, dot |
| Measurement | A process of establishing the correspondence between a property of the world and a number system |
| Methodology | The system of methods and principles used in a particular discipline |
| Metric | A system of measurement |
| Perception | The way in which something is regarded, understood, or interpreted, i.e., the translation of sense impressions into meaningful experiences of the outside world |
| Pharmacoepidemiology | The study of the use and effects of drugs in well-defined populations |

| Term | Description |
|------|-------------|
| Preference values | A quantitative measure of the extent to which an outcome achieves an objective, as judged by an individual or group<br><br>The value or utility associated with a score; preference values or utilities are judged by assessors to reflect the clinical relevance of effects or outcomes |
| Qualitative benefit-risk assessment | In a qualitative benefit-risk assessment, the clinical relevance of the evidence and the trade-offs between the safety and efficacy effects may be judged but are not quantified |
| Quantitative | Involving considerations of amount or size; capable of being measured |
| Quantitative benefit-risk assessment | In a partially quantitative benefit-risk assessment, the clinical relevance of the evidence, and the trade-offs between the favourable and unfavourable effects are quantified.<br><br>A fully quantitative benefit-risk assessment goes a step further by mathematically aggregating the favourable effects, the unfavourable effects, and the trade-off values into a measure of the benefit-risk balance. |
| Rates | The relative frequency of an event in a given time period |
| Reference point | An anchor on the visual, usually refers to meaningful values on the scale to aid information extraction |
| Reproducibility | A process or a set of results/decisions is reproducible if the steps involved and parameters used in the process are clearly defined and stated so that the process can be repeated by someone else |
| Revealed preference | An approach which observes and explores preferences as indirectly revealed by an individual's action(s) within real life situations |
| Risk | The negative results (adverse outcomes) of a given treatment for an individual or a population in terms of probability of occurrence having considered the magnitude of severity |
| Safety | The safety of a medical product concerns the medical risk to the subject (e.g., assessed in a clinical trial by laboratory tests (including clinical chemistry and haematology), vital signs, clinical adverse events (diseases, signs, and symptoms), and other special safety tests) |

| Term | Description |
|---|---|
| Saturation | The purity of primary colours in relation to the wavelengths; narrower wavelengths are more saturated than wider wavelengths |
| Score | The numeric values with fixed minimum and maximum (bounded scales) assigned to distinguish, e.g., magnitude, severity, performance, preference<br><br>A measure of a real world effect or outcome |
| Stated preference | An approach which asks individuals to state their preferences within hypothetical scenarios |
| Uncertainty | Uncertainty may refer to:<br><br>1. Randomness, the possibility of different outcomes from an action, which cannot be foreseen for sure in advance<br>2. Uncertainty in estimation due to insufficient sampling<br>3. Discrepancy in evidences from different sources of data<br>4. Measurement error or quality of data (e.g., data not measured by proper means or poor equipment) |
| Utility | A subjective measurement that describes a person's or group's preferences (e.g., satisfaction, risk attitude) for an effect or outcome |
| Value function | A function which converts the input data (parameters) in all criteria into preference value or utility for the options under evaluation |
| Value judgement | A subjective assessment for appropriateness of values or utility in a decision making problem |
| Value tree | A visual map, in a hierarchical diagram, of the benefits and risks that are being considered for the analysis; also referred to as an attribute or effects tree |
| Visual methods / representation | The principles and procedures to present some numerical features or relations by a graph |
| Weight | The scaling constants assigned to criteria such that the units of scaled preference values across all criteria are equal |

## Section 6   List of abbreviations

| Abbreviation | Description |
|---|---|
| ABS | Acute bacterial sinusitis |
| ACS | Acute coronary syndrome |
| ADR | Adverse drug reaction |
| AE | Adverse event |
| AECB | Acute exacerbation of chronic bronchitis |
| AE-NNT | Adverse Event adjusted Number Needed to Treat |
| AHP | Analytic Hierarchy Process |
| ALT | Alanine transaminase |
| ASF | Ashby and Smith Framework |
| BBN | Bayesian belief network |
| BLRA | Benefit Less Risk Analysis |
| BM | Beckmann Model (aka Evidence Based Model) |
| BP | Blood pressure |
| BR, B-R, B/R | Benefit-Risk |
| BRAT | Benefit-Risk Action Team |
| BRR | Benefit-Risk Ratio |
| CA | Conjoint Analysis |
| CCDS | Company Core Data Sheets |
| CDS | Cross Design Synthesis |
| CHF | Congestive heart failure |
| CHMP | Committee for Medicinal Products for Human Use |
| Chol | Cholesterol |
| CI | Confidence Interval |
| CIN | Case Impact Number |
| CIRS | Centre for Innovation in Regulatory Science |

| Abbreviation | Description |
|---|---|
| cm | Centimetre |
| CMR | Centre for Medicines Research |
| CMR CASS | Centre for Medicines Research Health Canada, Australia's Therapeutic Goods Administration, SwissMedic, and Singapore Health Science Authority |
| CMV | Cytomegalovirus |
| CNS | Central nervous system |
| COBRA | Consortium for Benefit-Risk Assessment |
| CPM | Confidence Profile Method |
| CPRD | Clinical Practice Research Datalink |
| CUI | Clinical Utility Index |
| CV | Cardiovascular |
| CV | Contingent Valuation |
| DAG | Directed Acyclic Graphs |
| DALY | Disability Adjusted Life Years |
| DCE | Discrete Choice Experiment |
| DI | Desirability Index |
| DIN | Disease Impact Number |
| DLQI | Dermatology Life Quality Index |
| DM | Decision Maker |
| DSUR | Development Safety Update Report |
| EBV | Epstein-Barr virus |
| EFPIA | European Federation of Pharmaceutical Industry Association |
| e.g. | *exempli gratia*, for example |
| EMA | European Medicines Agency  (formerly EMEA) |
| EPAR | European Public Assessment Report |
| et al. | Et alii (and others) |

| Abbreviation | Description |
|---|---|
| Etc | etcetera |
| EU | European Union |
| FDA | Food and Drug Administration |
| FDA BRF | FDA Benefit-Risk Framework |
| FT | Final treatment |
| GBR | Global Benefit-Risk |
| GBS | Guillain-Barre syndrome |
| GCP | Good Clinical Practice |
| GSK | GlaxoSmithKline |
| HALE | Health Adjusted Life Expectancy |
| HbA$_{1c}$ | Glycosylated haemoglobin |
| HDL | High-density lipoprotein |
| ICH | International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use |
| i.e. | *id est*, that is |
| IMI | Innovative Medicines Initiative |
| INB | Incremental Net Benefit |
| INHB | Incremental Net Health Benefit |
| ITC | Indirect Treatment Comparison |
| KBRS | Key benefit-risk summary |
| Kg | Kilograms |
| LDL | Low-density lipoprotein |
| MACBETH | Measuring Attractiveness by a Categorical Based Evaluation Technique; also referred to as M-MACBETH |
| MAH | Marketing Authorisation Holder |
| MAR | Maximum Acceptable Risk |
| MCDA | Multi-Criteria Decision Analysis |

| Abbreviation | Description |
|---|---|
| MCE | Minimum Clinical Efficacy |
| MDP | Markov Decision Process |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MI | Myocardial infarction |
| mmHg | Millimetres of mercury |
| mmol/ml | Millimole per millilitre |
| MTC | Mixed Treatment Comparison |
| n.a., N/A | Not Applicable |
| NCB | Net Clinical Benefit |
| NEAR | Net Efficacy Adjusted for Risk |
| NEPP | Number of Events Prevented in the Population |
| n/N | Number/Total Number |
| NNH | Number Needed to Harm |
| NNT | Number Needed to Treat |
| OLS | Overall lesion severity |
| OR | Odds ratio |
| OR | Overall risk |
| PADER | Periodic Adverse Drug Experience Report |
| PASI | Psoriasis Area and Severity Index |
| PBRER | Periodic Benefit-Risk Evaluation Report |
| PGA | Physicians Global Assessment |
| Ph. Alk | Alkaline phosphatase |
| PhRMA | Pharmaceutical Research and Manufacturers of America |
| PIN | Population Impact Numbers |
| PIN-ER-$t$ | Population Impact Numbers of Eliminating a Risk Factor over time $T$ |
| PML | Progressive multifocal leucoencephalopathy |

| Abbreviation | Description |
|---|---|
| PPI | Patient and Public Involvement |
| PRAC | Product Review Advisory Committee |
| PrOACT-URL | Problem, Objectives, Alternatives, Consequences, Trade-offs, Uncertainty, Risk, and Linked decisions framework |
| PROTECT | Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium |
| PSM | Probabilistic Simulation Method |
| PSUR | Periodic Safety Update Report |
| Pts | Patients |
| Ptyrs | Patient years |
| QALY | Quality Adjusted Life Years |
| QoL | Quality of Life |
| Q-TWiST | Quality-adjusted Time Without Symptoms and Toxicity |
| RCT | Randomised Controlled Trial |
| RDiff | Risk difference point estimates |
| REL | Time of Relapse to death |
| RMP | Risk Management Plan |
| RR | Relative risk point estimate |
| RV-MCE | Relative Value adjusted Minimum Clinical Efficacy |
| RV-NNH | Relative Value adjusted Number Needed to Harm |
| RV-NNT | Relative Value adjusted Number Needed to Treat |
| SABRE | Southeast Asia Benefit-Risk Evaluation group |
| SAE | Serious adverse event |
| SBRAM | Sarac's Benefit-Risk Assessment Method |
| SMAA | Stochastic Multi-criteria Acceptability Analysis |
| SPM | Stated Preference Method |
| STEMI | ST-elevated myocardial infarction |

| Abbreviation | Description |
|---|---|
| Tox | Toxicity |
| TOX | Time subject to toxicity effect |
| TURBO | Transparent Uniform Risk Benefit Overview |
| TWiST | Time Without Symptoms and Toxicity |
| UK | United Kingdom |
| ULN | Upper limit of normal |
| UMBRA | Unified Methodologies for Benefit-Risk Assessment |
| URTI | Upper respiratory tract infection |
| USA | United States of America |
| UT-NNT | Utility- and Time-adjusted Number Needed to Treat |
| wNCB | Weighted Net Clinical Benefit |
| WP | Work Package |
| Yr | Year |

# Section 7   References

Ades, A.E. and Sutton, A.J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(1): 5-35.

Ashby, D. and Smith, A.F. (2000). Evidence-based medicine as Bayesian decision-making. Statistics in Medicine, 19(23): 3291-3305.

Attia, J., Page, J., Heller, R.F., and Dobson, A.J. (2002).  Impact numbers in health policy decisions.  Journal of Epidemiology and Community Health, 56(8): 600-605.

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

Belton, K.J. (1997). Attitude survey of adverse drug reaction reporting by health care professionals across the European Union. The European Pharmacovigilance Research Group. European Journal of Clinical Pharmacology, 52(6): 423-427.

Belton, V. and Stewart, T.J. (2002). Multiple Criteria Decision Analysis: An integrated approach. United Kingdom: Springer.

Beyer, A.R., Fasolo, B., Phillips, L.D., deGraeff, P.A., and Hillege, H.L. (2013). Risk perception of prescription drugs: Results of a survey among experts in the European Regulatory Network. Medical Decision Making, Dol: 10.1177 / 0272989X12472397.

Boada, J.N., Boada, C., Garcia-Saiz, M., Garcia, M., Fernandez, E., and Gomez, E. (2008). Net efficacy adjusted for risk (NEAR): A simple procedure for measuring risk:benefit balance. PLoS One, 3(10): e3580.

Boada, J., Boada, C., Garcia, M.M., Rodriguez, C., Garcia, M., and Fernandez, E. (2009). Net efficacy adjusted for risk: Further developments. Expert Opinion on Drug Safety, 8(6): 649-654.

Cammax Limited (2011). The drugs box – easy-to-use educational tool about drugs and alcohol. http://www.thedrugsbox.co.uk/

Carswell, C.M. (1992). Choosing specifiers: An evaluation of the basic tasks model of graphical perception. Human Factors, 34(5): 535-554.

CHMP / EMA (2005). Guideline on risk management systems for medicinal products for human use. http://eudravigilance.emea.europea.eu/human/evriskmanagement.asp.

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

Chuang-Stein, C., Entsuah, R., and Pritchett, Y. (2008). Measures for conducting comparative benefit: Risk assessment. Drug Information Journal, 42(3): 223-233.

Chuang-Stein, C., Mohberg, N.R., and Sinkula, M.S. (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. Statistics in Medicine, 10(9): 1349-1359.

CIOMS Working Group IV (1998). Benefit-risk balance for marketed drugs: Evaluating safety signals. Geneva: Council for International Organizations of Medical Sciences.

CIRS Workshop Synopsis on Benefit-Risk (2012). Building the benefit-risk toolbox: Are there enough common elements across the different methodologies to enable a consensus on a scientifically acceptable framework for making benefit-risk decisions? http://cirsci.org/system/files/private/ CIRS_June2012_Workshop_Synopsis.pdf, 20-6-2012, Washington, DC: Centre for Innovation in Regulatory Science.

Clarkson, A. and Choonara, I. (2002). Surveillance for fatal suspected adverse drug reactions in the UK. Archives of Disease in Childhood, 87(6): 462-466.

Coplan, P.M., Noel, R.A., Levitan, B.S., Ferguson, J., and Mussen, F. (2011). Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. Clinical Pharmacology and Therapeutics, 89(2): 312-315.

Darwiche, A. (2010). Bayesian networks. Communications of the ACM, 53(12): 80-90.

Deal, L., Gold, B.D., Gremse, D.A., Winter, H.S., Peters, S.B., Fraga, P.D, Mack, M.E., Gaylord, S.M., Tolia, V., and Fitzgerald, J.F. (2005). Age-specific questionnaire distinguish GERD symptom frequency and severity in infants and young children: development and initial validation. Journal of Pediatric Gastroenterology and Nutrition, 41(2): 178-185.

Dickersin, K., Chan S., Chalmersx, T.C., Sacks, H.S., and Smith Jr., H. (1987). Publication bias and clinical trials. Controlled Clinical Trials, 8(4): 343-353.

Dodgson, J., Spackman, M., Pearman, A., and Phillips, L.D. (2000). Multi-Criteria Analysis: A Manual. London: Department of the Environment, Transport and the Regions.

Dolan, J.G., Qian, F., and Veazie, P.J. (2012). How well do commonly used data presentation formats support comparative effectiveness evaluations? Medical Decision Making, Nov, 32(6): 840-850.

Droitcour, J., Silberman, G., and Chelimsky, E. (1993). A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. International Journal of Technology Assessment in Health Care, 9(3): 440-449.

Easterbrook, P.J., Berlin, J.A., Gopalan, R., and Matthews, D.R. (1991). Publication bias in clinical research. Lancet, 337(8746): 867-876.

Eddy, D.M. (1989). The confidence profile method: A Bayesian method for assessing health technologies. Operational Research, 37(2): 210-228.

Eddy, D.M., Hasselblad, V., McGivney, W., and Hendee, W. (1988). The value of mammography screen in women under age 50 years. JAMA: The Journal of the American Medical Association, 259(10): 1512-1519.

Edwards, R., Wiholm, B.E., and Martinez, C. (1996). Concepts in risk-benefit assessment: A simple merit analysis of a medicine? Drug Safety, 15(1): 1-7.

Eichler, H.G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L.L., Leufkens, H., Rowland, M., Schneider, C.K., and Bloechl-Daum, B. (2011). Bridging the efficacy-effectiveness gap: A regulator's perspective on addressing variability of drug response. Nature Reviews Drug Discovery, 10: 495-506.

European Federation of Pharmaceutical Industries and Associations (EFPIA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) Principles for Responsible Clinical Trial Data Sharing: Our Commitment to Patients and Researchers (2013).
http://transparency.efpia.eu/responsible-data-sharing
http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf

European Medicines Agency (2009). Work Package 1 Report: Description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network (pp.23). London: Access at www.ema.europa.eu, Specific topics, Benefit-risk methodology.

European Medicines Agency (2012). Benefit-risk methodology project: Report on risk perception study module. http://www.ema. Europa.eu/docs/en_GB/document_library/Report/2012/02/ WCS00123226.pdf, 1-68.

European Medicines Agency (2013a). The Benefit-risk methodology project documents currently available on EMA website at:
http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/document_listing/document_listing_000314.jsp&mid=WC0b01ac0580223ed6

European Medicines Agency (2013b).  Draft Policy on Publication and access to clinical-trial data.
http://www.ema.europa.eu/docs/en_GB/document_library/Other/2013/06/WC500144730.pdf

European Medicines Agency (2013c). European Medicines Agency recommends changes to the use of metoclopramide, 26Jul2013, currently available on EMA website at:

http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2013/07/news_detail_001854.jsp&mid=WC0b01ac058004d5c1

Few, S. (2004). Dashboard confusion. Intelligent Enterprise, 20-3-2004.

Figueira, J., Greco, S., and Ehrgott, M. (eds.) (2005). Multiple Criteria Decision Analysis: State of the art surveys. NY: Springer Science + Business Media Inc, ISBN 0-387-23067-X.

French, S., Maule, J., and Papamichail, N. (2009). Decision behaviour, analysis and support. Cambridge University Press.

Frey, P. (2012). Benefit-risk considerations in CDER: Development of a Qualitative Framework. http://www.fda.gov/downloads/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/UCM317788.pdf, 28-6-2012, Philadelphia, PA: US Food and Drug Administration.

GAO/PEMD (1992). Cross design synthesis. A new strategy for medical effectiveness research. Washington, DC: United States General Accounting Office / Program Evaluation and Methodology Division, Mar 17, Report No. B-244808.

Gapminder (2011). Gapminder: Unveiling the beauty of statistics for a fact-based world view, http://www.gapminder.org/

Garrison, L.P., Towse, A., and Bresnahan, B.W. (2007). Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. Health Affairs (Millwood), 26(3): 684-695.

Gelber, R.D., Cole, B.F., Gelber, S., and Aron, G. (1995). Comparing treatments using quality-adjusted survival: The Q-Twist method. The American Statistician, 49(2): 161-169.

Goldhirsch, A., Gelber, R.D., Simes, R.J., Glasziou, P., and Coates, A.S. (1989). Cost and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. Journal of Clinical Oncology, 7(1):36-44.

Guyatt, G.H., Sinclair, J., Cook, D.J., and Glasziou, P. (1999). Users' guide to the medical literature: XVI. How to use a treatment recommendation. Evidence-Based Working Group and the Cochrane Applicability Methods Working Group. JAMA: The Journal of the American Medical Association, 281(19): 1836-1843.

Hammond, J.S., Keeney, R.L., and Raiffa, H. (1999 / 2002). Smart choices: A practical guide to making better decisions. Boston, MA: Harvard Business School Press.

Havet, N., Morelle, M., Remonnay, R., and Carrere, M.O. (2011). Cancer patients' willingness to pay for blood transfusions at home: Results from a contingent valuation study in a French cancer network. The European Journal of Health Economics, 1-12.

Hawley, S.T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T., and Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. Patient Educational Counseling, Dec, 73(3): 448-455.

Heller, R.F., Buchan, I., Edwards, R., Lyratzopoulos, G., McElduff, P., and Leger, S.S. (2003). Communicating risks at the population level: Application of population impact numbers. BMJ, 327(7424): 1162-1165.

Heller, R.F., Dobson, A.J., Attia, J., and Page, J. (2002). Impact numbers: Measures of risk factor impact on the whole population from case-control and cohort studies. Journal of Epidemiology and Community Health, 56(8): 606-610.

Holden, W.L. (2003). Benefit-Risk analysis: A brief review and proposed quantitative approaches. Drug Safety, 26(12): 853-862.

Holden, W.L., Juhaeri, J., and Dai, W. (2003a). Benefit-risk analysis: A proposal using quantitative methods. Pharmacoepidemiology and Drug Safety, 12(7): 611-616.

Holden, W.L., Juhaeri, J., and Dai, W. (2003b). Benefit-risk analysis: Examples using quantitative methods. Pharmacoepidemiology and Drug Safety, 12(8): 693-697.

Hunink, M., Glaszious, P., Siegel, J., Weeks, J.C., Pliskin, J.S., Elstein, A.S., and Weinstein, M.C. (2001). Decision making in health and medicine: Integrating evidence and values. Cambridge: Cambridge University Press.

IBM (2011). Many Eyes, http://www.958.ibm.com/software/data/cognos/manyeyes/

ICH (2012). ICH E2C (R2). Periodic Benefit-Risk Evaluation Report (PBRER) E2C (R2). International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/periodic-benefit-risk-evaluation-report.html

Jenkins, J.A. (2010). A United States regulator's perspective on risk-benefit considerations. Rockville, MD: Shady Grove Conference Center.

Jensen, F.V. (1996). An introduction to Bayesian Networks, First Edition. Secaucus, NJ: Springer-Verlag New York Inc.

Jensen, F.V. and Nielsen, T.D. (2007). Bayesian Networks and Decision Graphs. New York: Springer.

Johnson, F.R., Ozdemir, S., Mansfield, C., Hass, S., Siegel, C.A., and Sands, B.E. (2009). Are adult patients more tolerant of treatment risks than parents of juvenile patients? Risk Analysis, 29(1): 121-136.

Keeney, R.L. (1992). Value-focused thinking: A path to creative decision making. Boston, MA: Harvard University Press.

Keeney, R.L. and Raiffa, H. (1976). Decisions With Multiple Objectives: Preference and Value Tradeoffs. New York: John Wiley.

Korting, H. and Schafer-Korting, M. (1999). The benefit-risk ratio: A handbook for the rational use of potentially hazardous drugs. Boca Raton: CRC Press LLC.

Krantz, D. H., Luce, R.D., Tversky, A., and Suppes, P. (1971). Foundations of measurement Volume I: Additive and polynomial representations. New York: Academic Press.

Lahdelma, R., Hokkanen, J., and Salminen, P. (1998). SMAA – Stochastic multiobjective acceptability analysis. European Journal of Operational Research, 106(1): 137-143.

Laupacis, A., Sackett, D.L., and Roberts, R.S. (1988). An assessment of clinically useful measures of the consequences of treatment. New England Journal of Medicine, 318(26): 1728-1733.

Levitan, B.S., Andrews, E.B., Gilsenan, A., Ferguson, J., Noel, R.A., Coplan, P.M., and Mussen, F. (2011). Application of the BRAT framework to case studies: Observations and insights. Clinical Pharmacology and Therapeutics, Feb2011, 89(2): 217-224.

Liberti, L., McAuslane, N., and Walker, S. (2010). Progress on the development of a benefit/risk framework for evaluating medicines. Regulatory Focus, 1-6.

Lipkus, I.M. and Hollands, J.G. (1999). The visual communication of risk. The Journal of The National Cancer Institute Monographs, 25, 149-163.

Louviere, J.J., Flynn, T.N., and Carson, R.T. (2010). Discrete choice experiments are not conjoint analysis. Journal of Choice Modelling, 3(3): 57-72.

Louviere, J.J., Hensher, D.A., and Swait, J. (2000). Stated choice methods: Analysis and application. Cambridge: Cambridge University Presss.

Lu, G. and Ades, A.E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine, 23(20): 3105-3124.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. Statistics in Medicine, 21(16): 2313-2324.

Lynd, L.D., Marra, C.A., Najafzadeh, M., and Sadatsafavi, M. (2010). A quantitative evaluation of the regulatory assessment of the benefits and risks of rofecoxib relative to naproxen: An application of the incremental net-benefit framework. Pharmacoepidemiology and Drug Safety, 19(11): 1172-1180.

Lynd, L.D., Najafzadeh, M., Colley, L., Byrne, M.F., Willan, A.R., Sculpher, M.J., Johnson, F.L., and Hauber, A.B. (2010). Using the incremental net benefit framework for quantitative benefit-risk analysis in regulatory decision-making – a case study of alosetron in irritable bowel syndrome. Value Health, 13(4): 411-417.

Lynd, L.D. and O'Brien, B.J. (2004). Advances in risk-benefit evaluation using probabilistic simulation methods: An application to the prophylaxis of deep vein thrombosis. Journal of Clinical Epidemiology, 57(8): 795-803.

Minelli, C., Abrams, K.R., Sutton, A.J., and Cooper, N.J. (2004). Benefits and harms associated with hormone replacement therapy: Clinical decision analysis. BMJ, 328(7436): 371.

Mitchell, R.C. and Carson, R.T. (2005). Using surveys to value public goods: The contingent valuation method. Washington, DC: The Johns Hopkins University Press.

Mt-Isa, S., Hallgreen, C.E., Asiimwe, A., Downey, G., Genov, G., Hermann, R., Hughes, D., Lieftucht, A., Noel, R., Peters, R., Phillips, L.D., Shepherd, S., Micaleff, A., Ashby, D., and Tzoulaki, I. on behalf of the PROTECT Work Package 5 participants (2013b). Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 2 of 2. London, Report No. PROTECT WP5, Version 1 (15Feb2013).

Mt-Isa, S., Peters, R., Phillips, L.D., Chan, K., Hockley, K.S., Wang, N., Ashby, D., and Tzoulaki, I. on behalf of PROTECT Work Package 5 participants (2013a). Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 1 of 2. London, Report No. PROTECT WP5, Version 1 (15Feb2013).

Mt-Isa, S., Tzoulaki, I., Callreus, T., Micaleff, A., and Ashby, D. (2011). Weighing benefit-risk of medicines: Concepts and approaches. Drug Discovery Today: Technologies, 8: e29-e35.

Mt-Isa, S., Wang, N., Hallgreen, C.E., Callreus, T., Genov, G., Hirsch, I., Hobbiger, S., Hockley, K.S., Luciani, D., Phillips, L.D., Quartey, G., Sarac, S.B., Stoeckert, I., Micaleff, A., Ashby, D., and Tzoulaki, I. on behalf of PROTECT Work Package 5 participants (2012). Review of methodologies for benefit and risk assessment of medication. London, Report No. PROTECT WP5, Version 4 (14Feb2012).

Mussen, F., Salek, S., and Walker, S. (2009). Benefit-Risk Appraisal of Medicines. John Wiley & Sons Ltd.

National Research Council (2013). Sharing Clinical Research Data: Workshop Summary. Washington, DC: The National Academies Press.

Nixon, R.M., Bansback, N., and Brennan, A. (2007). Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. Statistics in Medicine, 26(6): 1237-1254.

Nisen, P. and Rockhold, F. (2013). Access to patient-level data from GlaxoSmithKline clinical trials. New England Journal of Medicine, Aug 1, 369(5): 475-478.

Nutt, D.J., King, L.A., Phillips, L.D., and on behalf of the Independent Scientific Committee on Drugs (2010). Drug harms in the UK: A multi-criteria decision analysis. The Lancelet, 376: 1558-1565.

Ouellet, D. (2010). Benefit-risk assessment: The use of clinical utility index. Expert Opinion on Drug Safety, 9(2): 289-300.

Ouellet, D., Werth, J., Parekh, N., Feltner, D., McCarthy, B., and Lalonde, R.L. (2009). The use of a clinical utility index to compare insomnia compounds: A quantitative basis for benefit-risk assessment. Clinical Pharmacology and Therapeutics, 85(3): 277-282.

Payne, J.T. and Loken, M.K. (1975). A survey of the benefits and risks in the practice of radiology. CRC Critical Reviews in Clinical Radiology and Nuclear Medicine, 6(3): 425-439.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Phillips, L.D. (1984). A theory of requisite decision models. Acta psychological, 56 (1-3): 29-48; ISSN 0001-6918.

Philllips, L.D. (1986). Decision analysis and its application in industry. In Mitra, Gautam (ed.), Computer assisted decision making expert systems, decision analysis, mathematical programming, North-Holland, ISBN 9780444878878.

Phillips, L.D., Fasolo, B., Zafiropoulos, N., and Beyer, A. (2010). Benefit-risk methodology project work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. London: European Medicines Agency, Aug 31, Report No.: EMA/549682/2010.

Potter, K. (2006). Methods for presenting statistical information: The box plot. Salt Lake City, UT: University of Utah, School of Computing.

Raiffa, H. (1968). Decision analysis: Introductory lectures on choices under uncertainty. McGraw-Hill.

Renard, D., Wu, K., Wada, K., and Flesch, G. (2009). Using desirability indices for decision making in drug development. Population Approach Group Europe (PAGE), Jun 23, St. Petersburg, Russia: Novartis Pharma AG, Basel, Switzerland.

Ried, W. (1998). QALYs versus HYEs –what's right and what's wrong: A review of the controversy. Journal of Health Economics, 17(5): 607-625.

Riegelman, R. and Schroth, W.S. (1993). Adjusting the number needed to treat: Incorporating adjustments for the utility and timing of benefits and harms. Medical Decision Making, 13(3): 247-252.

Ryan, M., Bate, A., Eastmond, C.J., and Ludbrook, A. (2001). Use of discrete choice experiments to elicit preferences. Quality in Health Care, 10: I55-I60.

Ryan, M., Gerard, K., and Amaya-Amaya, M. (2008). Using Discrete Choice Experiments to Value Health and Health Care. Dordrecht, The Netherlands: Springer.

Ryan, M. and Hughes, J. (1997). Using conjoint analysis to assess women's preferences for miscarriage management. Health Economics, 6(3): 261-273.

Sacristan, J.A., Gomez, J.C., Bandia, X., and Kind, P. (2001). Global index of safety (GIS): A new instrument to assess drug safety. Journal of Clinical Epidemiology, 54(11): 1120-1125.

Sarac, S.B., Rasmussen, C.H., Rasmussen, M.A., Hallgreen, C.E., Soeborg, T., Colding-Jorgensen, M., Christensen, P.K., Thirstrup, S., and Mosekilde, E. (2012). A comprehensive approach to benefit-risk assessment in drug development. Basic Clinical Pharmacology and Toxicology, 111(1):65-72.

Sassi, F. (2006). Calculating QALYs, comparing QALY and DALY calculations. Health Policy Plan, 21(5): 402-408.

Schulzer, M. and Mancini, G.B. (1996). 'Unqualified success' and 'unmitigated failure': Number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events. Internal Journal of Epidemiology, 25(4): 704-712.

Smith, R.D. (2003). Construction of the contingent valuation market in health care: A critical assessment. Health Economics, 12(8): 609-628.

Sonnenberg, F.A. and Beck, J.R. (1993). Markov models in medical decision making: A practical guide. Medical Decision Making, 13(4): 322-338.

Spiegelhalter, D. (2010). Understanding uncertainty website, http://understandinguncertainty.org.

Spiegelhalter, D., Abrams, K., and Myles, J.P. (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chicester, West Sussex: John Wiley & Sons Ltd.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Dong, F. (2000). Methods for meta-analysis in medical research. John Wiley & Sons, Chichester, ISDN 0 471 49066 0, Published October 2000.

Sutton, A.J., Cooper, N.J., Abrams, K.R., Lambert, P.C., and Jones, D.R. (2005). A Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. Journal of Clinical Epidemiology, 58(1): 26-40.

Tervonen, T. and Figueira, J.R. (2008). A survey on stochastic multicriteria acceptability analysis methods. Journal of Multi-Criteria Decision Analysis, 15(1-2): 1-14.

Tervonen, T., van Valkenhoef, G., Buskens, E., Hillege, H.L., and Postmus, D. (2011). A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. Statistics in Medicine, 30(12): 1419-1428.

Thompson, J.P., Noyes, K., Dorsey, E.R., Schwid, S.R., and Holloway, R.G. (2008). Quantitative risk-benefit analysis of natalizumab. Neurology, 71(5): 357-364.

van Staa, T.P., Smeeth, L., Persson, I., Parkinson, J., and Leufkens, H.G. (2008). What is the harm-benefit ratio of Cox-2 inhibitors? International Journal of Epidemiology, 37(2): 405-413.

Walker, S., McAuslane, N., Liberti, L., and Salek, S. (2009). Measuring benefit and balancing risk: Strategies for the benefit-risk assessment of new medicines in a risk-averse environment. Clinical Pharmacology and Therapeutics, March, 85(3): 241-246.

Wickens, C.D., Lee, J., Liu, Y.D., and Gordon-Becker, S. (2004). An introduction to human factors engineering (Second Edition). Upper Saddle River, NJ: Pearson Prentice-Hall.

# Section 8   Appendices

### Appendix 1 Criteria for Benefit-Risk Methodology Appraisal
Source:  Methodology Review (Mt-Isa et al., 2012)

| Dimension | Specific Evaluation Criteria |
|---|---|
| Fundamental principle | 1)  Is the method logically sound? This will be determined by the underlying mathematical/empirical reasoning used to build the models, and in the results e.g. the point estimates and construction of associated confidence intervals. <br> 2)  Does the method offer increased transparency in the assessment allowing reproducibility of the results? We will determine, descriptively, how the methods enforce transparency and whether any insufficient disclosure of the steps taken in the process prohibits reproducibility. <br> 3)  Does the method also produce statistical uncertainty estimates around the point estimates (using the standard models)? This is satisfied when the method has a technique to produce confidence intervals which are mathematically sound. Otherwise, we will describe whether the methods provide any guideline on how uncertainty is to be dealt with. <br> 4)  Can the method incorporate other sources of uncertainty in the input parameters? This is assessed by how the approach elicits the input parameters allowing for uncertainty in the response. <br> 5)  Can the principles of the methods be easily understood by the end users? We will describe to what extent the principles are thought important to be understood before a decision maker can build decision models or interpret the results from a particular method. <br> 6)  Does the approach appropriately incorporate value judgements, either explicitly or implicitly? Stakeholders' involvement in providing preference value is needed to satisfy this criterion. <br> 7)  How does the approach handle multiple options? Often in a decision making, more than two options (e.g. drug treatments) would be considered. We describe how an approach handles this, and whether there is a natural extension to the approach when it comes to multiple options. |

| Dimension | Specific Evaluation Criteria |
|---|---|
| Features of respective approaches | 1) Does the method appropriately allow balancing of the benefit-risk profile either numerically or visually? We will also describe whether the assessment of benefits and risks are done separately or simultaneously.<br>2) Can the model flexibly include several benefits and risks criteria? We shall also describe whether the method has a technique to handle multiple benefits and risks evidence simultaneously.<br>3) Can the model flexibly include multiple sources of evidence? We shall describe whether the method can incorporate pieces of evidence from different sources of data.<br>4) Does the method naturally allow sensitivity analysis? We will address the feasibility of conducting a sensitivity analysis for each method and what has been suggested e.g. to investigate the best and worst scenarios.<br>5) Can the method incorporate time dimension? We will describe how time variables are dealt with.<br>6) Is the model ready to be formally updated with new/additional data/assumptions? We will describe how feasible it is for a model built to be modified to take into account new evidence or changes in the input parameters.<br>7) Is there any unique feature of a particular method? We will describe any unique feature of a method that gives an added advantage to other methods. Additionally, we will also describe any fatal flaw, if any, of models built from a particular method. Available computer programmes and/or manuals relevant to the methods will also be described. |
| Visual representation of model | 1) Does the model propose potential visualisations of the results? We will describe the proposed visualisation techniques and what are they intended to represent. |
| Assessability and accessibility | 1) Are the parameters and results acceptable and easily interpretable (from the perspective of a non-statistician)? This shall include any interim results, if any, before the final results are reached. We will describe how the methods ensure consistency in the input parameters, if any. We will also describe where we see there are potential misinterpretations of the results.<br>2) How practical is the method when used in real-life decision making? This will address the economic aspects of the methods in terms of their complexity, the time to set up, the (monetary) cost involved if directly applicable, and the ease of rerunning/modifying the models.<br>3) Which perspective are the methods useful for e.g. for regulators, physicians, patients, stakeholders, etc.? We will also address whether a model built to take on one perspective can be easily modified into another.<br>4) In what respect the use of the approach can lead to make better decision making? |

## Appendix 2  Specific criteria used for the discussion section of all Wave 1 case studies

### *Appropriate frame*
*This should cover the following:*

   i.   *the rationale for the chosen approach in terms of the context of the problem, who the decision maker is, the timing, and the available expertise;*

   ii.  *the ease of implementation in terms of the technical demand, available clear documentations on the approach, any software that you used (please name) to simplify application, and the anticipated time to be spent and level of a new person to learn the approach;*

   iii. *the stage of product life cycle decision is to be made for – consider other scenarios that you could apply the approach including separate time points, continuous monitoring, change in indications etc.;*

   iv.  *the appropriateness of making comparison to creative doable alternatives e.g. some approaches do not allow comparison;*

   v.   *what are the limitations in your chosen approach when considering benefits and risks criteria;*

### *Meaningful reliable information*
*This should include the following:*

   i.   *what was the rationale for including or excluding the benefit and risk criteria? Can the approach deal with criteria other than efficacy and safety;*

   ii.  *to what extent are valid data available for favourable and unfavourable effects?*

   iii. *were clinical judgements about the effects available, and if so, at what stage were they considered? Please also consider at what stage should consumers e.g. the patients should be involved, if applicable, and why.*

### *Clear values and trade-offs*
*This should include the following:*

   i.   *does the approach make judgements of value explicit?  If so, how were these judgements obtained;*

   ii.  *were the favourable and unfavourable effects defined clearly in the approach, including the use of a common scale which allows for benefits to be directly compared to risks; and*

   iii. *are the final results easily interpretable? Please state any concerns you might have, e.g. unclear or no direct trade-offs, or results are still unclear for a decision to be made.*

### *Logically correct reasoning*
*This should include the following:*

   i.   *can the approach handle any form of data – continuous or discrete, qualitative or quantitative, objective or subjective?*

   ii.  *how was uncertainty in the data accommodated?*

     iii.     *what was the theoretical justification for combining effects (e.g., weighted averages justified on the basis of mutual preference independence of criteria; multiplying values or utilities by probabilities justified on the basis of the coherence of the expected utility model)?,*

     iv.     *does the approach have any apparent technical flaws that hinder correct reasoning and/or interpretation, and how did you overcome them; and*

     v.     *how did you deal with differences of opinion arising from the application of the approach amongst team members?*

## *Commitment to action*

*This should include the following:*

     i.     *did the approach develop insight and promote learning;*

     ii.     *to what extent are the final results directly relevant to the decision to be taken;*

     iii.     *to what extent are the final results easily communicable, transparent and easily understood;*

     iv.     *to what extent does the approach provide a clear audit trail so that all aspects of the benefit-risk evaluation can be traced; and*

     v.     *would your team recommend the approach forward, and why? Please also specify any disagreement.*

## Appendix 3   Methodologies – Short Descriptions

*Descriptive Frameworks*

**PrOACT-URL** (Hunink et al., 2001; Hammond, Keeney, Raiffa, 2002) is a generic decision making guide. The acronym PrOACT-URL represents the steps of this framework: (1) determine the decision context and frame the *Problems*; (2) establish *Objectives* and identify criteria; (3) identify options and *Alternatives*; (4) evaluate the expected *Consequences* of the options for each criterion; (5) assess the *Trade-offs* of benefit and risk; (6) report the *Uncertainty* in benefit and risk, and assess the impact of uncertainty on the benefit-risk balance; (7) judge the relative importance and the *Risk* attitude of the decision maker and assess how this affects the benefit-risk balance; and (8) consider the decision's consistency with other *Linked decisions*, both in the past and its impact on future decisions.

**ASF** (Ashby and Smith framework) (Ashby and Smith, 2000; Mt-Isa et al., 2011) is a simple framework for evidence-based medical decision making addressing five aspects: the decision maker, the possible actions, the uncertain consequences, the sources of evidence, and the utility assessments.

**BRAT** (Benefit-Risk Action Team) (Coplan et al., 2011; Levitan et al., 2011) standardises and supports the decision and communication of a benefit-risk assessment between pharmaceutical companies and the regulators through a 6-step process: define decision context, identify outcomes, identify data sources, customise framework, assess outcome importance, and display and interpret key benefit-risk metrics.

**FDA BRF** (Benefit-Risk Framework) (Frey, 2012; Jenkins, 2010) provides the "big picture" to "tell the story" by summarising evidence and addressing their implications for decision in a table for five decision factors: analysis of condition, unmet medical need, benefit, risk, and risk management.

**CMR CASS** (Canada, Australia, Switzerland, and Singapore) (Walker, 2009) was intended to be quantitative and meant for small regulatory agencies to address the benefit-risk throughout product lifecycle and the post-approval assessment challenges. It has been superseded by COBRA (Consortium on Benefit-Risk Assessment) (CIRS, 2012) with a mission to develop a semi-quantitative framework to reflect the actual practice, but no details are yet published.

**SABRE** (Southeast Asia Benefit-Risk Evaluation) (CIRS, 2012) is another recent regional initiative in Southeast Asia to promote better assessment of the benefits and risks of medicines, but details have not yet been published.

**UMBRA** (Unified Methodologies for Benefit-Risk Assessment) (CIRS, 2012) works with the PhRMA BRAT, COBRA, and SABRE initiatives to establish a unified benefit-risk framework with common elements, currently addressed in a 4-stage, 8-step process: (1) framing the decision – decision context; (2) identifying benefits and risks – building value tree, refining the value tree; (3) assessing benefits and risks – relative importance of benefits and risks, evaluating the options; and (4) interpretation and recommendations – evaluating uncertainty, concise presentation of results, and expert judgement and communication.

*Quantitative Frameworks*

**BLRA** (Benefit Less Risk Analysis) (Chuang-Stein, 1994) deals with multiple criteria decision problems using individual-level data. It organises observed adverse events into body functions for benefit-risk assessment. Benefits and risks are balanced by a defined proportionality constant $f$ in the expression $\sum \text{benefits} - f \times \sum \text{risks}$.

**NCB** (Net Clinical Benefit) (Sutton et al., 2005) is a quantitative framework that compares the overall change in the benefits and risks of a drug over a comparator. The framework is divided into three steps; (1) define the decision problem and data sources; (2) establish the functional form of the NCB equation; and (3) estimate the NCB, which is the sum of the change in expected benefits minus the change in expected risks as a result of treatment. Once the functional form of the NCB has been established, the benefits and risks must be placed on a common scale, such as health-state related utilities. The expected benefit is calculated by multiplying the benefit, assuming it is realised by the patient, by the probability of it being realised, with a similar calculation for expected risks.

A **decision tree** is a horizontal tree diagram, with decisions as roots, uncertain events with their outcomes, and further decisions and outcomes as branches, ending with consequences (Hunink et al., 2001; Raiffa, 1968; Spiegelhalter, Abrams and Myles, 2004). The expected utility rule is applied repeatedly and added at each node ('averaging out'), and then 'folded back' for the highest utility decision.

**MDP** (Markov Decision Process) (Sonnenberg and Beck, 1993; Thompson et al., 2008) is multi-stage decision making with finite states and options. The probability of being in the subsequent state only depends on the current state as defined by a set of transitions. The aim is to find an option that maximises the expected utility of the entire process.

**MCDA** (Multi-Criteria Decision Analysis) (Mussen, Salek, and Walker, 2009; Dodgson et al., 2000; Keeney and Raiffa, 1976) is a process derived from decision theory that quantifies the overall performance of two or more alternatives. As applied to the benefit-risk balance of a drug and its comparators, performance of the alternatives on the favourable and unfavourable effects are judged for their clinical relevance, and all effects are weighted to create a common unit of preference value or utility. Summing those common units of benefit and risk provides an overall benefit-risk preference value or utility for each alternative, enabling calculation of the difference of the drug against the comparators.

**SMAA** (Stochastic Multi-criteria Acceptability Analysis) (Tervonen and Figueira, 2008; Tervonen et al., 2011; Lahdelman, Hokkanen, and Salminen, 1998) is a multi-criteria decision method dealing with statistical uncertainty. It combines the distributions of scores $\xi$ and weights $w$ for each option across all criteria in $b = \int_{\xi} f(\xi) \int_{w} f(w) \, dw d\xi$ which then calculates the probabilities of an option being a certain rank.

**SBRAM** involves eight successive steps: (1) establishment of the decision context; (2) identification of benefit-risk criteria; (3) weighting of criteria; (4) scoring of criteria; (5) evaluation of uncertainty; (6) calculation of weighted scores; (7) discussion of results; and (8) formulation of an overall conclusion (Sarac et al., 2012). To allow

comparisons across different benefit-risk categories, different criteria are weighted on a scale of 1 (low), 2 (medium), and 3 (high) according to their importance. In order to reduce the impact of subjective judgments, scores are assigned to each criterion on the basis of the data wherever possible. Scoring charts are used to visualise the data under the SBRAM's scoring technique. Weights and scores are multiplied, and the results are visualised in a tornado-like diagram. Uncertainties are dealt with qualitatively from the scoring charts or quantitatively using bootstrapping.

**CUI** (Clinical Utility Index) and **DI** (Desirability Index) (Ouellet et al, 2009; Ouellet, 2010; Renard et al., 2009) provide a general framework in assessing the benefit-risk balance of drugs under development when measured over a range of doses or time. CUI and DI are defined over (0,1) range and calculated as $\sum(\text{weight} \times \text{utility})$ and $(\prod \text{utility})^{\frac{1}{\text{weight}}}$ respectively.

*Threshold metric indices*

**NNT** (Number Needed to Treat) (Holden, Juhaeri, and Dai, 2003; Laupacis, Sackett, and Roberts, 1988) is derived from the probabilities of a favourable effect for the treatment and comparator groups. The difference between the two probabilities, $p_t$ and $p_c$, gives the increase in certainty, $p_t - p_c$. NNT is then calculated as the reciprocal of this difference, $1/(p_t - p_c)$, and can be interpreted as the number of patients that need to be treated (on average) for one event to be observed as a result of treatment. A parallel but opposite metric, number needed to harm (NNH), is defined similarly but based on the probabilities of unfavourable effects. AE-NNT (Adverse Event Adjusted-NNT) (Schulzer and Mancini, 1996) penalises NNT for the occurrence of AEs in the same patient. RV-NNH (Relative Value adjusted NNH) (Guyatt et al., 1999) incorporates stakeholders' value preferences on the importance of AEs into NNH.

**Impact numbers** (Attia et al., 2002; Heller et al., 2002; Heller et al., 2003) are a group of metrics that generalise the NNT concept to the population level instead of focusing on only those patients who receive treatment. By considering the baseline event probabilities in the population of interest, estimates of the number of individuals that will be affected by a disease and/or an intervention can be derived. Therefore, these metrics describe the 'impact' of treatments from the public health perspective.

**MCE** (Minimum Clinical Efficacy) (Holden, Juhaeri, and Dai, 2003a; Holden, Juhaeri, and Dai, 2003b) determines the minimal therapeutic benefit for a treatment to be worth considering, accounting for the event probability when untreated. RV-MCE (Relative Value adjusted MCE) incorporates stakeholders' value preferences on the importance of AEs into MCE.

**MAR** (Maximum Acceptable Risk) (Johnson et al., 2009) is analogous but opposite to MCE. MAR assumes mutually exclusive benefit and risk events, and estimates the metric as $Pr(\text{Risk}) \times U(\text{Risk}) + (1 - Pr(\text{Risk})) \times U(\text{Benefit})$.

**NEAR** (Net Efficacy Adjusted for Risk) (Boada et al., 2008; Boada et al., 2009) uses benefit or risk event and non-event count data for two comparative treatments in a $2 \times 2$ table. NEAR estimates NEAR odds ratio (OR) or relative risk (RR) using the standard formulae for OR and RR.

*Health indices*

**Health indices** are validated and standardised QoL indicators. Provided they are derived from the same assessment tool, they are comparable across different populations and treatments. Four health indices are described below.

**QALY** (Quality Adjusted Life Years) (Sassi, 2006; Ried, 1998) is the most used health index, where the time spent in a particular health state is multiplied by the QoL score in that state. The total QALY is simply the sum of all QALYs in all health states. The summary of all individual QALYs in a population is known as the **HALE** (Health Adjusted Life Expectancy).

**DALY** (Disability Adjusted Life Years) (Sassi, 2006) is a parallel extension of QALY and is an index quantifying number of years lost from treatment compared to the national life expectancy.

**Q-TWiST** (Quality-adjusted Time Without Symptoms and Toxicity) (Gelber et al., 1995; Goldhirsch et al., 1989) is in principle a QALY metric, with explicit definitions of the discrete health states in cancer therapy: toxicity, time without symptoms and toxicity, and relapse.

*Trade-off metric indices*

**UT-NNT** (Utility and Time adjusted NNT) (Riegelman and Schroth, 1993) adjusts the benefit-risk event probabilities in NNT for the time saved or lost due to treatment and the utilities associated with the treatment.

**INHB** (Incremental Net Health Benefit) (Garrison, Towse, and Bresnahan, 2007; Lynd, Najafzadeh, et al., 2010; Minelli et al., 2004) calculates the difference in the "incremental" change of benefits to that of risks. INHB uses QALY specifically to characterise benefits and risks, but other metrics can be used and generalises as INB (Incremental Net Benefit) (Lynd, Marra, et al., 2010).

**BRR** (Benefit Risk Ratio) (Chuang-Stein, Entsuah, and Pritchett, 2008; Korting and Schafer-Korting, 1999; Payne and Loken, 1975) is a simple trade-off metric which divides benefits by risks, and, therefore, assumes equal importance of benefits and risks.

**GBR** (Global Benefit Risk) (Chuang-Stein, Entsuah, and Pritchett, 2008; Chuang-Stein, Mohberg, and Sinkula, 1991) refers to three trade-off metrics constructed around individual patients' outcomes in clinical trials: linear, ratio, and

composite ratio scores. The application of GBR requires a pre-determined proportionality constant to rescale risk to the same unit as benefit.

**Principle of three** (Mussen, Salek, and Walker, 2009; Edwards, Wiholm, and Martinez, 1996) assesses benefit-risk balance using three criteria (disease, effectiveness, and adverse drug reactions) on three attributes (seriousness, duration, and incidence), each scored on three levels 1-3.

**TURBO** (Transparent Uniform Benefit Risk Overview) (Mussen, Salek, and Walker, 2009; CIOMS, 1998) is a benefit-risk concept which prioritises two most important benefit and risk criteria. Primary criteria are scored on a scale of 1-5 and secondary criteria on a scale of 1-2. The final $T$ (Transparent) composite score is determined from an arbitrary grid.

**Beckmann Model** (Mussen, Salek, and Walker, 2009; Beckmann, 1999) scores benefit of treatment as efficacy $\times$ response rate $\times$ evidence, and scores risk as seriousness $\times$ incidence $\times$ evidence. The categories of 'efficacy' and 'seriousness' in the expressions are determined arbitrarily.

*Estimation techniques*

**DAG** (Directed Acyclic Graph) (Pearl, 1988; Jensen, 1996; Jensen and Nielsen, 2007; Darwiche, 2010) is a graphical structure where nodes are connected through directed edges. It relies on conditional independences which allow decomposition of information on strength of associations in benefit-risk models into distinct probability distributions.

**PSM** (Probabilistic Simulation Method) (Lynd and O'Brien, 2004; van Staa et al., 2008) is a statistical technique for exploring the impact of uncertainty in data on a model's results. In applying PSM to benefit-risk assessment, statistical summaries of data in a quantitative model are replaced with probability distributions related to the patient-level data. The overall benefit-risk balance is then calculated a large number of times with different input data drawn from the probability distributions in proportion to their likelihood of being chosen. This generates a probability distribution over the difference between benefits and risks for the drug, and another distribution for the comparator. The same process can then be applied to determine the probability that the benefit-risk balance of the drug is more than that of the comparator. PSM provides a good means to quantify and explore the uncertainty of the benefit-risk balance. PSM can accommodate any type of metric as well as the correlations between favourable and unfavourable effects, if it is known.

**CPM** (Confidence Profile Method) (Eddy et al., 1988; Eddy, 1989; Ades and Sutton, 2006) uses conditional probabilities as arbitrarily specified in a "chain of evidence," similar to DAGs. A benefit-risk metric is calculated from single link chains for direct evidence and from multiple link chains for linking together indirect evidence.

**ITC** (Indirect Treatment Comparison) and **MTC** (Mixed Treatment Comparison) (Lumley, 2002; Lu and Ades, 2004; Nixon, Bansback, and Brennan, 2007) are meta-analytic methods to synthesise different pieces of evidence into a coherent set of estimates of treatment effects. In the absence of direct comparative evidence between two treatments, ITC can infer their relationship through a common comparator; for example, both treatments may have been directly compared against a placebo. Indirect comparisons are subject to greater statistical uncertainty than direct comparisons, and this effect on uncertainty is captured by ITC. MTC generalises the concept by providing a method to integrate both direct and indirect evidence.

**CDS** (Cross Design Synthesis) (Droitcour, Silberman, and Chelimsky, 1993; GAO/PEMD, 1992; Deal et al., 2005) combines randomised clinical trials evidence with evidence from clinical databases or observational data. CDS is intended to improve benefit-risk evidence by eliminating biases and complementing the weaknesses of one study design with another's strengths.

*Utility Survey Techniques*

**SPM** (Stated Preference Method) (Ryan, Gerard, and Amaya-Amaya, 2008) explores how stakeholders respond to decision problems in hypothetical scenarios. The response to real scenarios is known as the revealed preference method.

**CV** (Contingent Valuation) (Smith, 2003; Mitchell and Carson, 2005; Havet et al., 2011) is an SPM approach where stakeholders are asked about their willingness to pay for a more beneficial option. Conversely, CV asks about willingness to accept compensation for a less beneficial option.

**CA** (Conjoint Analysis) (Louviere, Hensher, and Swait, 2000; Louviere, Flynn, and Carson, 2010) breaks down hypothetical scenarios into a set of characteristics and attributes to ease the utility elicitation process, and then mathematically combines them to produce the overall expected utility.

**DCE** (Discrete Choice Experiment) (Ryan, Gerard, and Amaya-Amaya, 2008; Ryan and Hughes, 1997; Ryan et al., 2001) uses exactly the same principles as CA with a more structured guideline to generating the hypothetical scenarios to be used in the elicitation process.

## Appendix 4 Quantitative Framework Appraisal

Source: Methodology Review (Mt-Isa et al., 2012)

**Original Source Table 4 Comparative overview and justifications for recommendations: Descriptive frameworks**

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | **Reasons** | **Specific use** |
| PrOACT-URL | • Strongly emphasises uncertainties in input values and value judgements as well as the importance of sensitivity analysis<br>• Proposes 'effects table' as snapshot of evidence | • Missing the importance of identifying appropriate sources of evidence and immediate parties involved<br>• Has been extended in EMA benefit-risk Methodology working group 2<br>• Does not address communication | Yes | • Address the necessary elements in dealing with decision problems<br>• Forms basis for other frameworks | • To structure decision problems using 8-step process<br>• To allow transparency |
| BRAT | • Value tree model build-up<br>• Does not integrate benefit and risk<br>• Optional weights assignment to benefit and risk criteria<br>• Summarise criteria as tables and forest/dot plots | • Summarises evidence and communicates them but does not assess B-R<br>• Can be exhaustive | Yes | • Developed by PhRMA<br>• Accessible to those not familiar with complex statistical models<br>• Offers graphical presentation of results in the form of a forest plot | • To structure decision problems using 6-step process<br>• To allow transparency<br>• To aid B-R communication |
| ASF | • Based on evidence and reiterates decision making | • Suitable for physicians and patients | No | • Features addressed in the extended PrOACT-URL | • To structure decision problems using 5-step process |
| CMR-CASS | • Consider product life-cycle<br>• Consider post-approval phase | • Targeting small regulatory agencies | No | • Still under development | nil |
| FDA BRF | • Template for facilitate BR decision and communication | nil | No | • Still under development | • To provide decision makers with the "big picture" of decision problems |

## Appendix 5 High Level Overview to Determine Type of Benefit-Risk Method Utilised

The figure below provides a graphic statement of the distinctions in this report we used to classify a benefit-risk method or approach as qualitative, partially quantitative, or fully quantitative. All three categories presume that data have been collected and summarised, and the relevant benefits and risks have been identified. The subsequent paths in the diagram, ending in the ovals, depend on whether or not judgments about clinical relevance and trade-offs have been quantified, and judgements aggregated using a mathematical rule.

**NOTE:** *The graphic serves only to make explicit our working principles for classifying the methods we examined; It does not represent the full process of making a benefit-risk assessment of a medicinal product.*
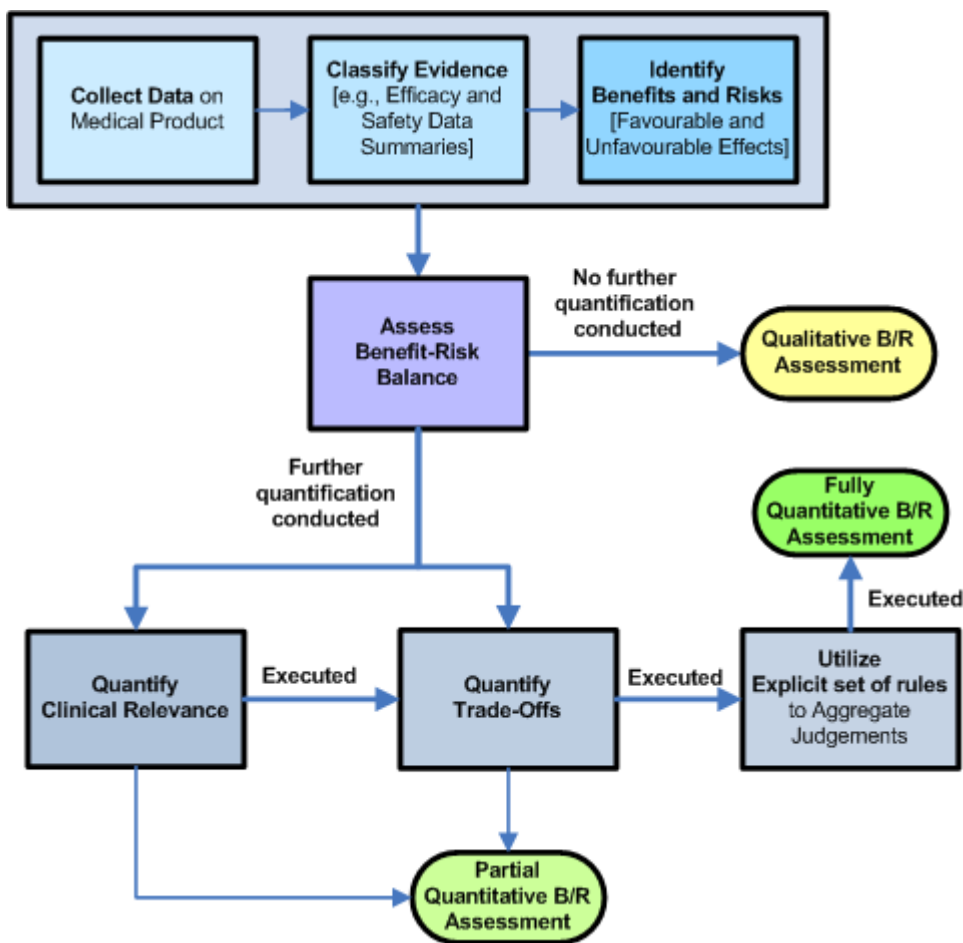
# Appendix 6  Sample tree diagrams

## Appendix 6.1 Decision Tree for rimonabant –Medical/Regulatory prospective with weighted results from trials
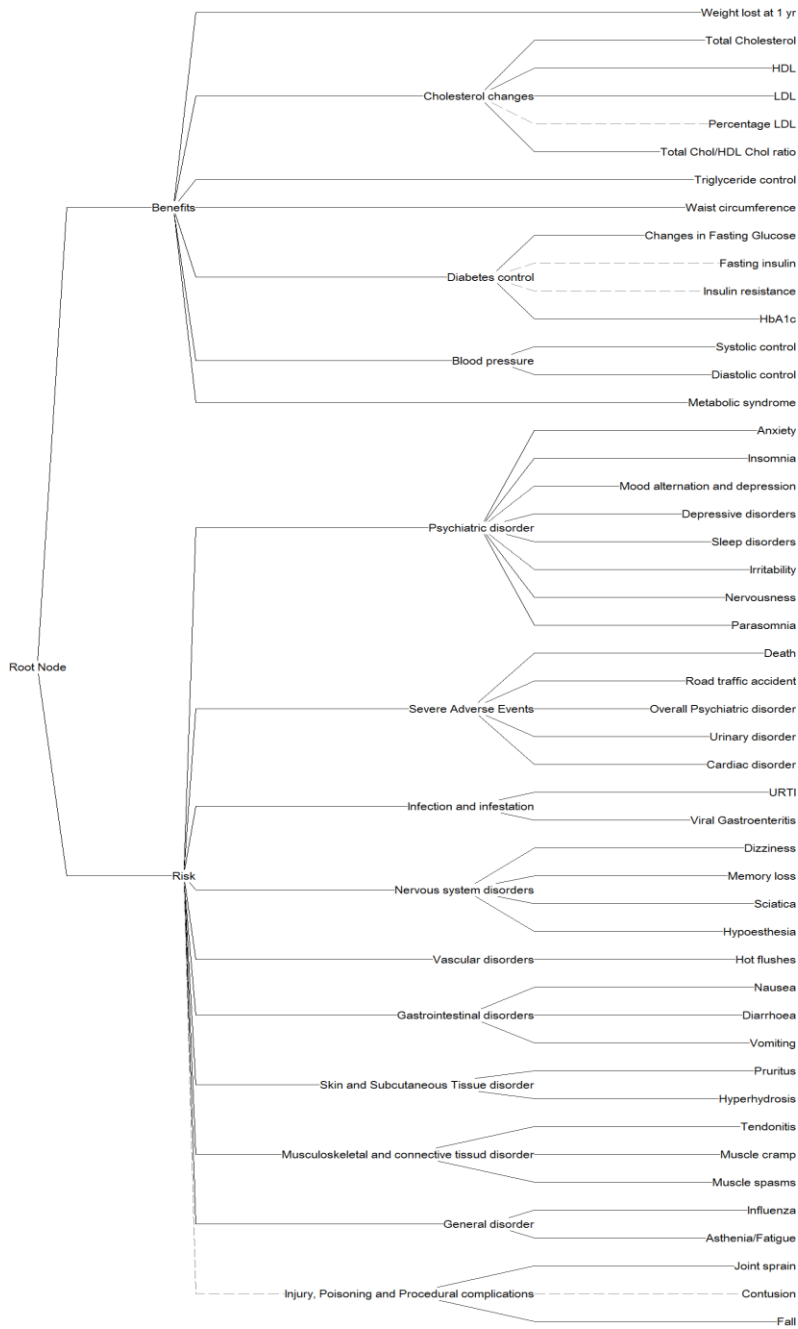
## Appendix 6.2 Decision Tree for rimonabant using Microsoft Word SmartArt graphics



The decision tree shows "Acomplia" branching into two categories:

**Benefits:**
- 10% weight lost at 1 year
- Reduction in incidence of metabolic syndrome
- Total Cholesterol changes
- HDL Cholesterol changes
- LDL Cholesterol Changes
- Total Cholesterol/HDL Ratio changes
- Triglycerides changes
- Waist circumference changes
- Fasting Glucose
- Insulin resistance
- HbA1c changes
- Systolic blood pressure changes
- Diastolic blood pressure changes

**Risks:**
- Infection & infestation
- Psychiatric disorder
- Nervous system disorder
- Vascular disorder
- Gastrointestinal disorder
- Skin and subcutaneous tissue disorder
- Muscular and connective tissue disorder
- General disorder
- Injury, Poisoning, Procedure related complication
- Severe Adverse events

**Appendix 6.3 Value tree for possible rimonabant risks and benefits (using the FreeMind software)**
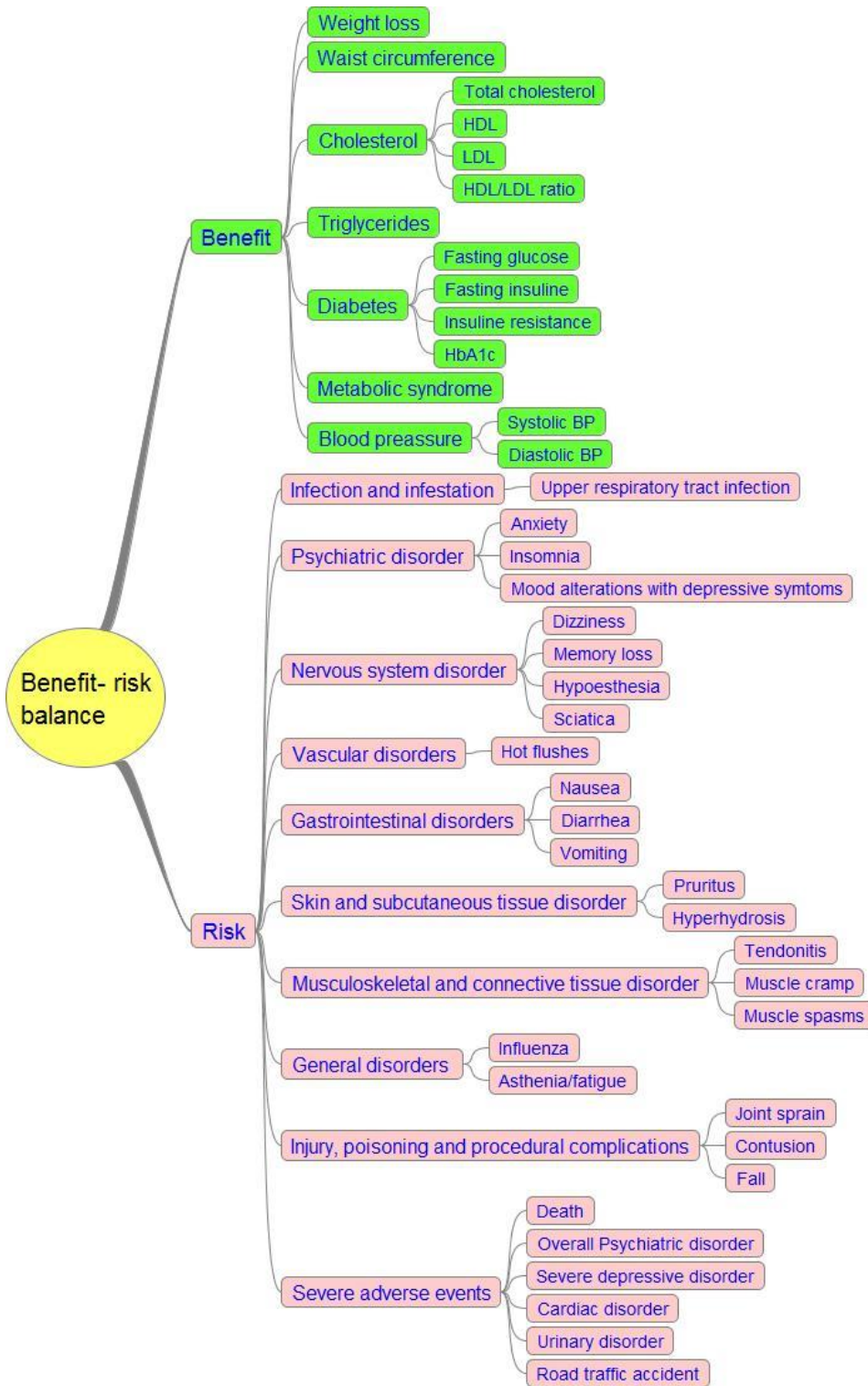
## Appendix 6.4 Value tree for possible rimonabant risks and benefits (using the BRAT software)

## Appendix 6.5 A reduced value tree for rimonabant (using the BRAT software)

```
                    ┌─ Weight loss ──────────────→ Weight loss - >10%
                    │
        Benefits ───┼─ Waist circumfence ────────→ Waist circumfence - cm reduced
        │           │
        │           └─ Systolic blood pressure ──→ Systolic blood pressure - reduction mmHg
Benefit-Risk
Balance │
        │           ┌─ Psychiatric disorder ─────→ Psychiatric disorder - percent
        │           │
        Risks ──────┼─ Nervous system disorder ──→ Nervous system disorder - percent
                    │
                    └─ Severe adverse events ────→ Severe adverse events - percent
```
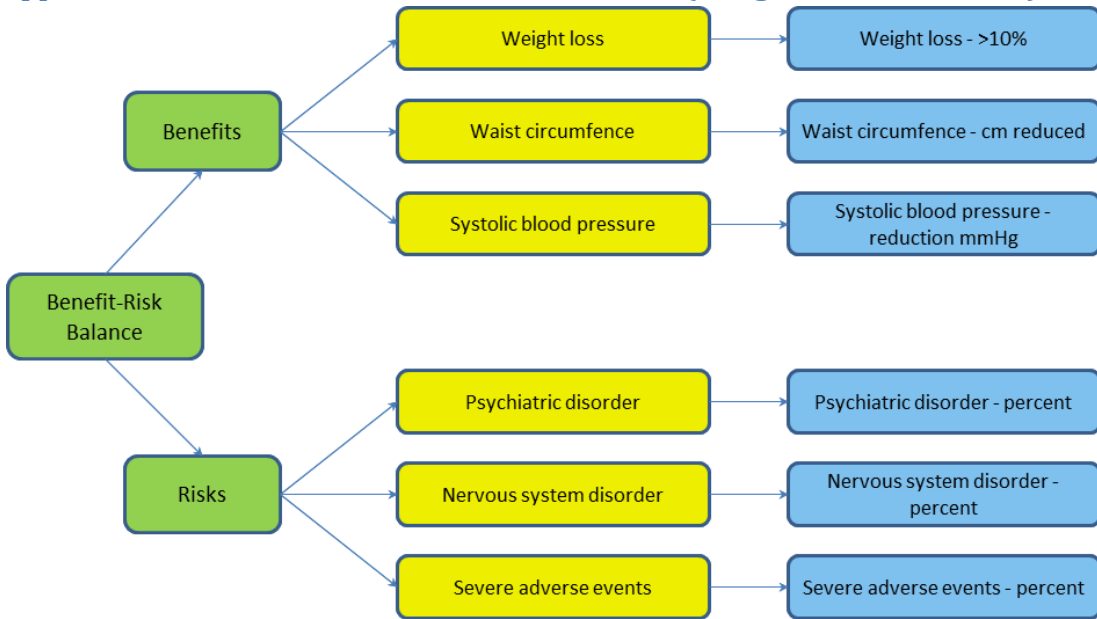
## Appendix 6.6 Value Tree for efalizumab (using the BRAT software)
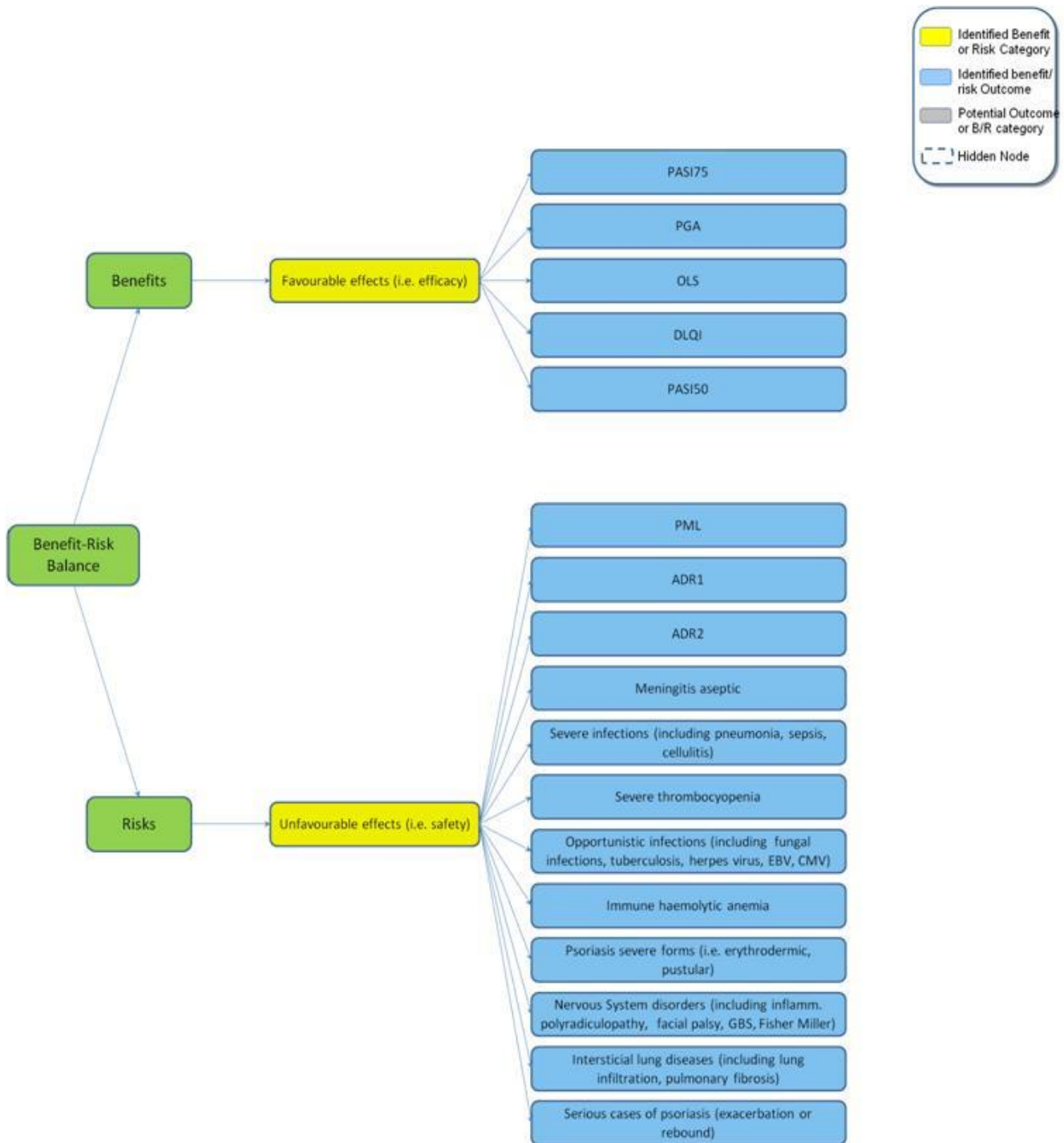
## Appendix 7  Estimation Techniques Comparison

Source: Methodology Review (Mt-Isa et al., 2012)

**Original source Table 15 Comparative overview and justifications for recommendations: Estimation techniques**

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| DAGs | • Graphical method<br>• Uses principles from Bayes Network<br>• Network diagrams as visuals | • Graphical feature can help in communicating the characteristics of the underlying decision<br>• Similar to MDP | No | • Similar to MDP | • To establish the relationship between evidence<br>• To support the application of benefit-risk assessment<br>• To assess benefit and risk |
| PSM | • Uses Monte-Carlo simulation or re-sampling form original data<br>• Can be applied to any type of data<br>• Highly flexible | • Can be applied in combination with most quantitative benefit-risk approaches | Yes | • Can be applied to most of the quantitative benefit risk methods<br>• Flexibility | • To support the application of benefit-risk assessment<br>• To assess benefit and risk<br>• To deal with uncertainties |
| CPM | • Deals with multiple benefit-risk criteria<br>• Deals with multiple sources of evidence<br>• Evidence easily updated under the Bayesian framework | • Mathematically exhaustive<br>• Requires extensive mathematical modelling expertise | No | • May be difficult to apply being mathematically exhaustive<br>• Somewhat difficult for routine use<br>• Similar to MTC | • To support the application of benefit-risk assessment<br>• To assess benefit and risk<br>• To deal with uncertainties and varieties of data |
| ITC | • A meta-analytic approach<br>• Allows comparison of two options indirectly through common denominator when direct evidence is unavailable<br>• Flexible | • Offer increased transparency in terms of clarifying the sources of evidence, bias and uncertainties<br>• Otherwise incomparable options can be compared | No | • Superseded by MTC | • To support the application of benefit-risk assessment<br>• To assess benefit and risk<br>• To deal with uncertainties<br>• To compare options where there is no direct evidence |

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | **Reason** | **Specific use** |
| MTC | • Generalisation of ITC<br>• Include both direct and indirect evidence<br>• Flexible to deal with complex structures | • Similar to ITC<br>• Collapsed to ITC when there is no direct evidence<br>• Requires statistical modelling expertise but fairly straightforward to understand<br>• Computer codes to implement MTC are available | Yes | • Flexible to accommodate many aspects of evidence synthesis<br>• Flexible to deal with complex structures<br>• Computer codes to implement are available | • To support the application of benefit-risk assessment<br>• To assess benefit and risk<br>• To deal with uncertainties<br>• To compare options where there is no direct evidence<br>• To improve inference using both direct and indirect evidence |
| CDS | • A meta-analytic approach<br>• Focuses on potential biases form study designs weaknesses<br>• Focus on synthesising the evidence instead of comparing the outcomes.<br>• Does not integrate benefit and risk | • Benefits and risks evidence from one population are used to predict benefit and risks in a slightly different population<br>• Principles can be adopted using MTC | No | • Principles can be adopted using MTC | • To support the application of benefit-risk assessment<br>• To assess benefit and risk<br>• To deal with uncertainties and biases<br>• To combine different sources of evidence |

## Appendix 8 Metric Indices Appraisal

Source: Methodology Review (Mt-Isa et al., 2012)

**Original Source Table 10 Comparative overview and justifications for recommendations: Threshold metric indices**

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| NNT and NNH | • Can only include one criteria<br>• The reciprocal of absolute risk reduction<br>• Benefit and risk are described separately as NNT or NNH<br>• Implies equal weight between benefit and risk when direction comparing NNT to NNH | • Undefined with no treatment effect<br>• CI's are problematic when absolute risk reduction includes zero (CI includes infinity)<br>• Values of NNT for different conditions are not comparable<br>• Timeframes must be considered carefully since they are used implicitly in the calculations | Yes | • Widespread use in clinical literature<br>• Simple<br>• Easy to understand | • To describe results in terms of number of people<br>• To facilitate communication to lay persons |
| UT-NNT | • Extension of NNT to incorporate utility and time | • It is trade-off index but falls directly within NNT family | No | • Similar to NNT and INHB<br>• May lead to implausible interpretations | • To incorporate utility and time factors into NNT analysis<br>• Also see NNT |
| AE-NNT | • Extension of NNT based on marginal probabilities<br>• Integrates one benefit with multiple risks | • Can only be used correctly when individual level data of treatment are available | No | • Similar to NNT<br>• Individual level data may be difficult to obtain | • To integrate benefit and risk in NNT analysis<br>• Also see NNT |
| RV-NNH | • Extension of NNH to include utilities<br>• Can include multiple risks by the reciprocal sum of the products of absolute risk difference and their relative values | • Has no upper limit, means that RV-NNH measures approaches zero, which contribution to implausible interpretation<br>• Incorporation of utilities changes the definition of the reciprocal | No | • Similar features to NNH<br>• May lead to implausible interpretations | • To account for subjective judgements on criteria for NNT<br>• To integrate multiple benefits and risks<br>• Also see NNT |

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| Impact numbers | • Similar to NNT and based on classical epidemiological metrics<br>• Several impact numbers were proposed for different purposes<br>• Provide population perspective<br>• Does not integrate benefits and risks<br>• Two of the metrics (NEPP and PIN-ER-*t*) do not require reciprocation | • Emphasise importance of justifying data sources<br>• Not suitable for rare idiosyncratic reactions<br>• Better interpretation to the general audience<br>• NEPP and PIN-ER-*t* do not suffer disadvantages of NNT | Yes | • Similar to NNT<br>• Taken population of interest data into account<br>• Relatively new concept with potential in benefit-risk assessment particularly in epidemiology | • To provide population perspective based on the number of people in the population of interest<br>• To characterise benefit-risk balance in specific populations<br>• Also see NNT |
| NEAR | • Avoids null point and "sign" problem of NNT<br>• Presents results as relative risks or odds ratios<br>• Compares one benefit and one risk<br>• Tables and forest plots as visuals | • There is extension to deal with intention-to-treat and per protocol analyses<br>• Uses expected frequencies hence does not need marginal probabilities | No | • Similar to NNT and AE-NNT<br>• No clear advantage compared to other NNT-related metrics | • To integrate one benefit and one risk<br>• To characterise B-R balance using OR and RR concepts |
| MCE | • Based on point estimates and unable to handle uncertainty in the measure of benefit and risks<br>• Integrates one benefit and one risk | • Requires comparison of two active treatments<br>• Statistical properties are not well-studied | No | • Similar to NNT<br>• Statistical properties are not well-studied | • To assess the threshold at which efficacy can be established |
| RV-MCE | • Extension of MCE to include utilities<br>• Integrates multiple benefits and risks | • Requires comparison of two active treatments<br>• Statistical properties are not well-studied | No | • Similar to RV-NNH and MCE | • To account for subjective judgements on criteria in MCE analysis<br>• Also see MCE |
| MAR | • Compares one risk to multiple benefits individually<br>• Closely linked to utility survey techniques<br>• Bar and antenna graphs as visuals | • Assumes benefit only occurs when risk does not (mutually exclusive events) | No | • Similar to SPM<br>• Mutually exclusive events assumption does not normally apply | • To assess the threshold at which risk becomes no longer acceptable |

Original Source Table 11 Comparative overview and justifications for recommendations: Health indices

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| QALY | • Measure of life time with quality of life incorporated<br>• Instruments used to derive quality of life are usually validated<br>• Integrates benefit and risk and includes time dimension<br>• Scatter plots as visuals | • Can be derived in a number of ways<br>• The most appropriate health instrument for deriving QALYs is subjective in some areas<br>• Validation for health instruments may not be in the population of interest | Yes | • Provides a measure of time trade-off with life quality<br>• Established in many areas of medicine | • To assess B-R balance after taking quality of life into account |
| DALY | • Years lost compared to national life expectancy, accounting for years lost due to disability<br>• Conceptually opposite of QALY<br>• Can act as a population measure | • See QALY | No | • Similar to QALY<br>• Not used as often as QALY | • To assess B-R balance after taking quality of life into account using population perspective |
| HALE | • Benefit and risk criteria affect disability weights | • Simply a summary of QALY in a concerned population | No | • Similar to QALY | • To summarise QALY |
| Q-TWiST | • Integrates benefit and risks and incorporates time dimension<br>• Confined to survival endpoints only<br>• Sensitivity analysis can be performed on choice of utility<br>• Visualisation by stratified survival curve for one treatment only | • Developed for oncology<br>• Easy to understand<br>• Health states are defined | Yes | • QALY modified for cancer therapy | • To assess B-R of cancer therapy using defined health states |

Original Source Table 12 Comparative overview and justifications for recommendations: Trade-off metric indices

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| INHB | • Commonly used with health indices<br>• Implicitly assumes equal weights for benefits and risks<br>• Compares two options each time | • Assumption of equal weights for benefit and risks can be overcome by establishing a common metrics for benefits and risks before using INHB<br>• Easy to perform and understand | Yes | • Simple and intuitive<br>• Uses established health indices such as QALYs | • To assess and integrate benefits and risks when described by health indices<br>• Also see QALY |
| BRR | • Can only deal with one benefit and one risk<br>• Assumes equal weighting of benefit and risks<br>• Similar to NNT | • Not transparent for benefit-risk assessment when used in its simplest form<br>• May be derived using thorough evidence synthesis and statistical modelling<br>• Should only be used with high quality data<br>• Must be presented together with their absolute or baseline rates | Yes | • Intuitive<br>• Simple to calculate<br>• Commonly used with other indices | • To assess and integrate benefit and risk<br>• To characterise the equilibrium point when benefit equals risk |
| GBR | • Integrates benefit and risks<br>• Multiple benefit and risks are not differentiated, but regarded as a collective criteria | • Does not explicitly distinguish the extend of severity of seriousness of adverse events<br>• Collectively analysing criteria may result in loss of information<br>• Three functional forms of GBRs were proposed | No | • May not be easily understandable or interpretable without the knowledge of the three measures which in this case is not well known.<br>• They are also not directly comparable to other measures | • To assess and integrate multiple benefits and risks |
| Principle of three | • Simple multi criteria model<br>• Employs three criteria: "disease", "effectiveness" and "ADRs"<br>• Benefit and risk are not integrated<br>• Tables as visuals | • Simple<br>• Low discriminative scoring<br>• Not suitable for complex situations<br>• Does not account for the relative importance of different criteria | No | • Does not provide a robust metric for benefit risk assessment | • To provide a snapshot of simple benefit and risk assessment<br>• May be used for preliminary assessment |

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| TURBO | • Simple multi-criteria decision making approach<br>• Accommodates two benefit criteria and two risk criteria<br>• Second criteria is regarded as correction factor to the first and scored on a shorter scale<br>• Grids as visuals | • Too simple to be transparent or to be used in drug benefit-risk decision making | No | • Does not provide a robust metric for benefit risk assessment<br>• Limited to two criteria | • To provide a snapshot of simple benefit-risk assessment<br>• May be used for preliminary assessment |
| Beckmann Model | • Simple multi-criteria model<br>• Quality of data contributes to the scores<br>• Does not integrate benefit and risk | • Easy to perform<br>• Does not take into account relative importance of benefit and risk criteria<br>• Does not incorporate uncertainties | No | • Does not provide a robust metric for benefit risk assessment<br>• Similar to principle of three | • To provide a snapshot of simple benefit and risk assessment<br>• May be used for preliminary assessment |

## Appendix 9 Quantitative Framework Appraisal
Source: Methodology Review (Mt-Isa et al., 2012)

**Original Source Table 5 Comparative overview and justifications for recommendations: Quantitative frameworks**

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| BLRA | • Organises adverse events into body functions for analysis<br>• Sensitivity analysis involves varying proportionality constant | • Burdensome for limited evaluations and simple problems<br>• More medical knowledge is required compared to others | No | • Very similar to MCDA | • To structure and analyse decision problems using 7-step process<br>• To allow transparency<br>• To integrate benefits and risks |
| NCB | • Naturally allows evidence update as Bayesian models but can also be used in a frequentist setting.<br>• Line graphs and distribution plots as visuals | • Functional form for quantifying benefit risk trade-off is not specific | No | • May require extensive statistical modelling expertise<br>• Similar to MCDA | • To structure and analyse decision problems using 3-step process<br>• To allow transparency<br>• To integrate benefits and risks |
| Decision tree | • Represents the expected utility rule visually<br>• Tree and tornado diagrams as visuals | • Utilities for the nodes on decision tree may be influenced by another | No | • Other methodologies recommended have incorporated the principles<br>• Can be highly complex with large "trees" | • To structure and analyse decision problems using 5-step process<br>• To allow transparency<br>• To integrate benefits and risks<br>• To investigate whether certain data are worth obtaining |
| MDP | • Similar to decision tree<br>• Markov chain combined with decision tree | • Not all medical decision problems can be described by MDP's dynamic nature<br>• The structure can be very complex with many criteria<br>• Use may be limited due to assumption that past history does not matter in future decisions | No | • Use may limited due to assumption that past history does not matter in future decisions<br>• Transition probabilities may be difficult to establish<br>• Explicit steps are unclear<br>• Similar to decision tree | • To structure and analyse decision problems<br>• To allow transparency<br>• To integrate benefits and risks |

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| MCDA | • Explicit value judgements<br>• High discriminative scoring system<br>• Multiple sources of evidence can be defined as criteria<br>• Multiple objectives can be addressed simultaneously<br>• Allows any data types<br>• Value tree diagram, line graphs, bar graphs (including difference diagram), and area graphs (frontier plot) as visuals | • Burdensome for limited evaluations and simple problems<br>• Software: Hiview, V.I.S.A. Intelligent decision system, Logical Decisions, etc.<br>• Hiview provides some useful visual representations<br>• Does not account for uncertainties in data | Yes | • Highly structured<br>• Can deal with multiple objectives simultaneously<br>• Combining multiple criteria is easy<br>• Several software to implement are available eliminating the need for mathematical knowledge of decision theory | • To structure and analyse decision problems using 8-step process of PrOACT-URL<br>• To allow transparency<br>• To integrate benefits and risks |
| SMAA | • Features similar to MCDA<br>• Uncertainties in evidence data are taken into account<br>• Can deal with missing or partially-missing value preferences<br>• Bar and line graphs as visuals | • Extends MCDA<br>• Software: JSMAA<br>• Requires extensive mathematical and computational knowledge<br>• There are several specialist metrics/concepts in SMAA<br>• There are variations of SMAA | Yes | • Complement MCDA<br>• Provide better estimates than MCDA when evidence are unknown, uncertain, or when their distributions are skewed | • To structure and analyse decision problems<br>• To allow transparency<br>• To integrate benefits and risks<br>• To account for uncertainties<br>• To better reflect situations encountered in real-life decision problems |
| SBRAM | • Does not integrate benefit and risk<br>• Incorporate value judgements implicit in the scoring in an objective manner<br>• Clinical data evaluated and scored based on descriptive statistic<br>• Low discriminative scoring system<br>• Tornado-like diagram as visuals | • Only compares two options at a time<br>• To evaluate multiple options additional scoring of data is required and analysis is compared visually<br>• No software available for data analysis (scoring)<br>• Underlying statistical analysis might be difficult to understand for layman<br>• Developed for drug development | No | • Similar to but lacks comprehensiveness of MCDA<br>• More specific to drug development<br>• Requires greater knowledge of statistical inferences | • To structure and analyse decision problems using 8-step process<br>• To allow transparency<br>• To eliminate ambiguous absolute judgement in the final results |

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| CUI/DI | • Not model specific, therefore the final metrics can be compared across different assessments<br>• Data must be from controlled clinical trials with same indication<br>• Does not require alternatives<br>• Line graphs, surface and contour plots as visuals | • Transparency hampered from having a rather too general framework<br>• More useful when utility index is expressed as a function of dose | No | • More specific to drug development<br>• Requires data that are difficult to obtain<br>• Limited applicability | • To structure and analyse decision problems using 4-step process<br>• To allow transparency<br>• To integrate benefits and risks<br>• To establish the benefit-risk balance without a comparator |

## Appendix 10  Utility Survey Techniques Comparison
Source: Methodology Review (Mt-Isa et al., 2012)

**Original Source Table 18 Comparative overview and justifications for recommendations: Utility survey techniques**

| Approach | Features | Comments | Has it been recommended that this be taken forward to the next stage? | | |
|---|---|---|---|---|---|
| | | | | Reason | Specific use |
| SPM | • Benefit and risk described through a hypothetical scenario<br>• Accommodates multiple benefits and risks which has the potential to vary over time<br>• Can collect large amounts of data with moderate cost<br>• Can examine proposed changes from a stakeholder perspective prior to implementation | • Methods to conduct may vary greatly since there is no standard way to implement | No | • There is no standard way to implement<br>• Similar to DCE | • To elicit preference values through a hypothetical scenario |
| CV | • Places monetary value on trade-offs using the willingness to pay concept<br>• Assumes benefits and risks in medicine behave like market goods<br>• Similar to SPM | • Known bias for being to over-sensitive if stakeholder directly affected and opposite if not<br>• Monetary valuations differ between people<br>• Treatments may be available for free thus CV becomes inappropriate | No | • Willingness-to-pay using money for trade-off can be biased to how people perceive money<br>• Treatments may be free<br>• Similar to DCE | • To elicit preference values through a hypothetical scenario<br>• To use money as trade-off currency |
| CA | • Hypothetical scenario as in SPM is broken down to a specific number of attributes before evaluated | • More robust than SPM | No | • Similar to DCE | • To elicit preference values through hypothetical scenarios |
| DCE | • Provides structured framework to elicit utilities<br>• Based on random utility theory and statistical experimental design<br>• Hypothetical scenario is broken down to a specific number of attributes before evaluated<br>• Anything can be defined as an attribute including time<br>• Minimises bias in response | • Requires statistical expertise in experimental designs<br>• Takes time<br>• There are some debates on internal validity, consistency and test-retest reliability<br>• Can be used to investigate how specific attributes may be viewed differently by different stakeholders | Yes | • Well-structured approach<br>• Most comprehensive of the utility survey techniques reviewed<br>• Provides transparency in the elicited preference values | • To elicit preference values through hypothetical scenarios<br>• To make robust inference on preference values<br>• To collect preference values for use with other benefit-risk approaches |

## Appendix 11 "Baseball cards" for visualisations (based on the natalizumab case study)

### Flow chart 1

**Literature searching results**
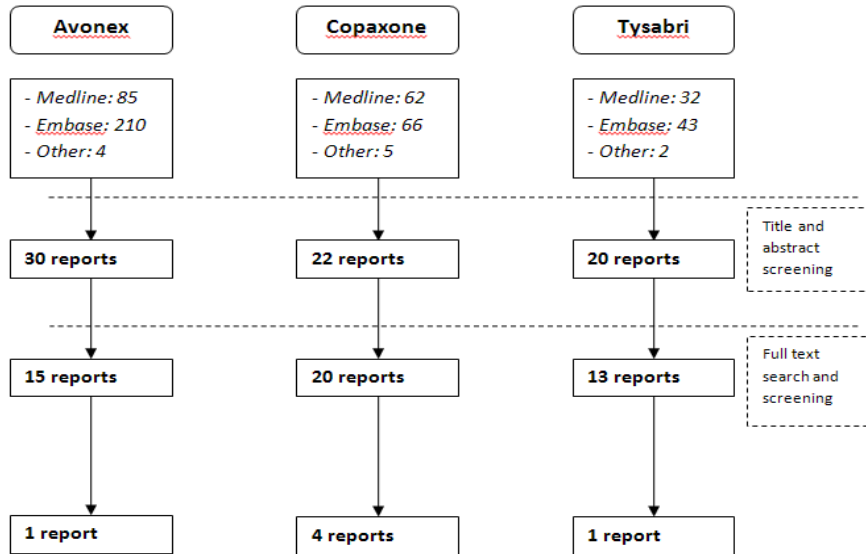


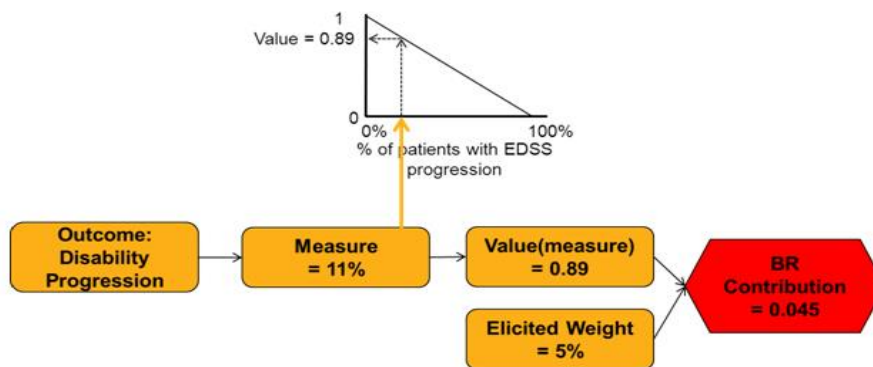| Name/rubric: | Flow chart showing systematic review literature screening |
|---|---|
| Created in: | Microsoft Word (drawing facility) |
| Intended audience: | Statisticians and regulators. Not for patients. |
| Message: | To visualise the flow of a literature search. Figure shows the sources and process of extracting evidence data. It also shows the amount of relevant data available in terms of number of articles/reports. |
| Knowledge required: | Some knowledge on systematic review process and quality of database sources. |
| Unintentional message: | N/A |
| Message not communicated: | N/A |
| Proposed improvement: | Provide number of excluded articles and brief reasons for exclusion at each screening stage |
| Comment | The visual is not taken into Phase II of visual methodology work since it is too general and too simple to benefit from interactive visualisation. |

*Flow chart 2*

**Benefit-risk calculations**



| | |
|---|---|
| **Name/rubric:** | Flow chart |
| **Created in:** | Microsoft Powerpoint |
| **Intended audience:** | Statisticians, Regulators, not for Physicians or Patients |
| **Message:** | To visualise the MCDA concept of scoring. The figure shows benefit risk contribution of one isolated parameter (Disability); relationship between value and parameter, modelled as a linear function. |
| **Knowledge required:** | In depth knowledge of MCDA and value functions. Knowledge of parameter/endpoint on the horizontal axis and the plausible values. |
| **Unintentional message:** | Unclear. Also, there is no inclusion of sensitivity or uncertainty aspects of the value function. |
| **Message not communicated:** | It is also beneficial to focus further on the likely range of the values on the horizontal axis? |
| **Proposed improvement:** | It might help some users understand the concept better if similar figures are produced for all key parameters. The figure would benefit from annotations explaining more clearly why a linear function was chosen. The terminologies used are unfamiliar to many users, therefore would require more explanations e.g., the difference between value and weight, and why MCDA requires them. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too general and varies from one approach to another. Explaining concepts of benefit-risk approach through visualisation is also beyond the scope of this case study. |

*Table 1*

**Descriptions of key data sources**

| Drug | Author | Year | Exposure | Comparator | Details |
|------|--------|------|----------|------------|---------|
| Tysabri | Polman CH | 2006 | 28 Months | Placebo | RRMS. Nov 2001: 942 patients - multiple centres and countries. 856 patients (91%) completed the 120-week trial |
| Avonex | Jacobs LD | 1996 | 24 Months | Placebo | RR MS. Early 1993: 301 patients - 4 clinical centers US. |
| Copaxone | Johnson KP | 1995 | 24 months | Placebo | RRMS. October 1991-May 1992: 251 randomized. 215 completed 2-year trial. US population. |

| | |
|---|---|
| **Name/rubric:** | Simple descriptive table on included study articles |
| **Created in:** | Microsoft Word (table facility) |
| **Intended audience:** | Statisticians, regulators, and physicians. Not for patients |
| **Message:** | To lay out descriptive (or numerical) information in a grid structure. The figure describes the study characteristics; showing larger, more recent, and longer trial was carried out for natalizumab compared to Beta-interferon and Glatiramer acetate. All active drugs used placebo as comparator. |
| **Knowledge required:** | Low statistical and low medical knowledge |
| **Unintentional message:** | The table implicitly assumes that the quality of evidence for all studies is similar and for comparison to be made, and that the populations are also comparable. This may not always be the case. |
| **Message not communicated:** | Characteristics of patients who took part in the trials are not highlighted, implicitly assuming they are comparable. |
| **Proposed improvement:** | Display demographics information for patients who took part in the trials and highlight similarities or dissimilarities. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too general and too simple to benefit from interactive visualisation. |

*Table 2*

**Master data summary table (only partly-shown here)**

| GROUP | CATEGORY | OUTCOME | MEASURE | DRUG | COMMON PLACEBO DRUG | EST | RELATIVE VALUE DESCRIPTION | EST | OUTCOME ON VALUE SCALE EST |
|---|---|---|---|---|---|---|---|---|---|
| Benefit | Relapse | Relapse | 2 year relapse rate | Placebo | Tysabri PBO | 0.73[1] | rate ratio | 1 | 1.46 |
| | | | | Tysabri | | 0.73 | | 0.32 | 0.47 |
| | | | | Avonex | | 0.73 | | 0.82 | 1.19 |
| | | | | Copaxone | | 0.73 | | 0.71 | 1.04 |
| | Disability progression | Disability progression | 6-month confirmed % progressing after 2 years | Placebo | Tysabri PBO | 0.17 | hazard ratio | 1.00 | 0.23 |
| | | | 6-month confirmed % progressing after 2 years | Tysabri | | 0.17 | | 0.46 | 0.11 |
| | | | 6-month confirmed % | | | | | | |

| | |
|---|---|
| **Name/rubric:** | Table as used in BRAT framework (only a part of full table is shown) |
| **Created in:** | Microsoft Word (table facility) |
| **Intended audience:** | Statistician, regulators and physicians. Not for patients. |
| **Message:** | To lay out descriptive (or numerical) information in a grid structure The figure shows a summary of the master data consisting of benefits and risks criteria used, the way they are measured, and their magnitudes from all studies for all drugs being compared. |
| **Knowledge required:** | Some familiarity with the units of measurements and terminologies used – percentage (%), rate ratio, hazard ratio, value scale. Some medical knowledge on the outcomes and epidemiology of the drug-disease. |
| **Unintentional message:** | The table header labelling is confusing, particularly in the sixth column when it is labelled with "Common Placebo" but listed natalizumab in subsequent rows in that column. The colour-coding and variation are meaningless since they only represent hierarchy of criteria which does not add value. |
| **Message not communicated:** | N/A. Note: This is part of a larger table with risks also defined. |
| **Proposed improvement:** | Colour-coding should be done by row instead of column to be meaningful. Colour-coding by rows imply grouping of benefits and risks criteria therefore making the measurements more easily interpretable. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too general and too simple to benefit from interactive visualisation. |

*Table 3*

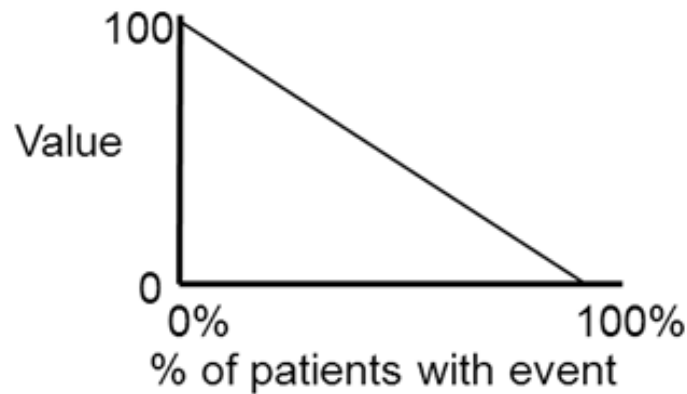**natalizumab versus Placebo (Comparator) at time of CHMP re-evaluation**

| | | Outcome | Tysabri Risk / 1000 pts | Comparator Risk / 1000 pts | Risk Difference (95% CI)/ 1000 pts | |
|---|---|---|---|---|---|---|
| Benefits | Convenience Benefits | Convenience (weight 0.6%) | - | - | - | (-, -) |
| | Medical Benefits | Relapse (weight 3.9%) | 280 | 540 | -260 | (-326, -195) |
| | | Disability Progression (weight 5.6%) | 110 | 230 | -120 | (-, -) |
| Risks | Infection | Reactivation of serious herpes viral infections (weight 6.7%) | 80 | 70 | 10 | (-26, 45) |
| | | PML (weight 55.9%) | 2 | 0 | 2 | (-, -) |
| | Liver Toxicity | Transaminases elevation (weight 11.2%) | 50 | 40 | 10 | (-16, 38) |
| | Reproductive Toxicity | Congenital abnormalities (weight 5.6%) | - | - | - | (-, -) |
| | Neurological Disorders | Seizures (weight 5.6%) | 0 | 0 | 0 | (-, -) |
| | Other | Infusion/Injection reactions (weight 2.8%) | 236 | 180 | 56 | (6, 114) |
| | | Hypersensitivity reactions (weight 1.1%) | 90 | 40 | 50 | (20, 82) |
| | | Flu-like reactions (weight 1.1%) | 399 | 400 | -1 | (-114, 114) |

Higher for Tysabri
Higher for Comparator

| | |
|---|---|
| **Name/rubric:** | Key benefit risk summary table as used in BRAT framework |
| **Created in:** | BRAT Excel Tool table with edit in Microsoft PowerPoint |
| **Intended audience:** | Physicians and regulators. Not for patients |
| **Message:** | To lay out descriptive (or numerical) information in a grid structure. The figure shows numerical key benefits and risk with comparison to alternative treatments. The absolute numbers on the table are supportive to allow better judgements. |
| **Knowledge required:** | Low statistical knowledge and some medical knowledge of the outcomes. |
| **Unintentional message:** | Risks have higher weights than benefits from the number of criteria used in the model. The legend is misleading when the colours in the last column are coded as 'Higher for natalizumab' instead of coding in terms of 'Favors natalizumab', and likewise for Comparator. |
| **Message not communicated:** | The criteria weights are displayed in the table, but the role of the weights is not clear. It is unclear whether the results presented have taken into account weights or whether the weights should be considered separately. This in turns could make deciding on the benefit risk balance much more critical. |
| **Proposed improvement:** | Comparison to other comparators than placebo may also help in the decision making. A table can be more complex than a graph. The use of blue and yellow scheme should be replaced with the traditional green and red scheme. If weights are to be presented, the ordering should reflect the weights. Weights should be removed or incorporated into the incidences and other calculations since presenting individual weights is not self-explanatory. Labelling of placebo should be explicit as "Placebo" instead of "Comparator." There should also be a text note to emphasise that data come from clinical trials. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too general and too simple to benefit from interactive visualisation. |

## *Line graph 1*

**Value function used for proportions**



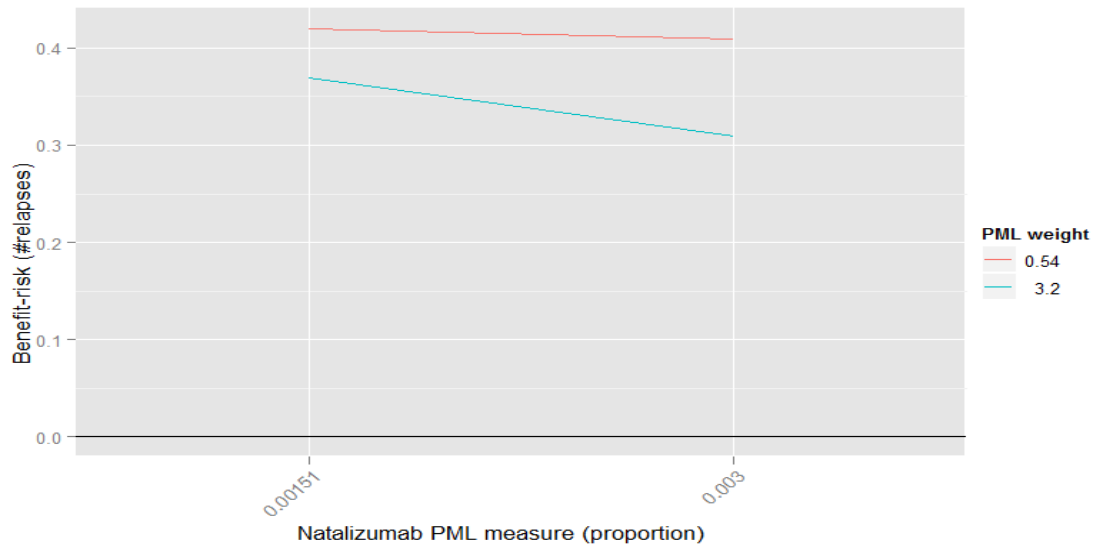| | |
|---|---|
| **Name/rubric:** | Line graph of preference values against a binary event (proportions of patients with an event) |
| **Created in:** | Microsoft Word Draw facility. Can also be created in most packages |
| **Intended audience:** | Statisticians, Regulators, not for Physicians or Patients |
| **Message:** | To show the relationship between preference value and data for one isolated parameter, under linear map assumption of the value function. The representation is unclear, especially when compared to a similar figure which includes flow chart for more context and information. |
| **Knowledge required:** | Needs in depth knowledge of MCDA and value functions. Knowledge of the parameter/endpoint on horizontal axis and its plausible values are also required. |
| **Unintentional message:** | There is no inclusion of sensitivity or uncertainty aspects which are likely to be associated with individual's preference. The aspect ratio of less than 1 (longer vertical axis than horizontal axis), may be perceived as increase in 1% of patients with event translates to less than 1 unit increase in preference value when they actually equal to the same amount. |
| **Message not communicated:** | Value function for other benefits and risks criteria are not shown but could easily be produced. There is no explanation whether the value function refers to active treatments or placebo, or is the same for all treatments. It is beneficial to focus further on the likely range of values on the horizontal axis (data values) and the justifications for range of values choice. |
| **Proposed improvement:** | The value function should be provided for all key parameters. A text explanation to justify the use of a linear value function is needed for transparency. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too simple to benefit from interactive visualisation |

## Line graph 2

**Value function used for convenience (administration route)**
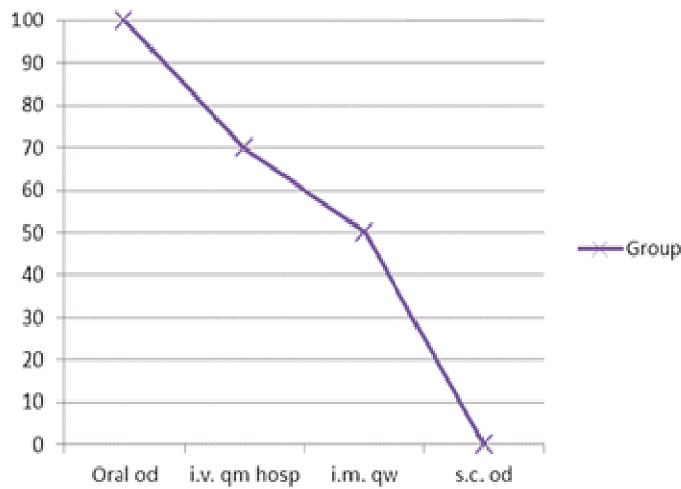


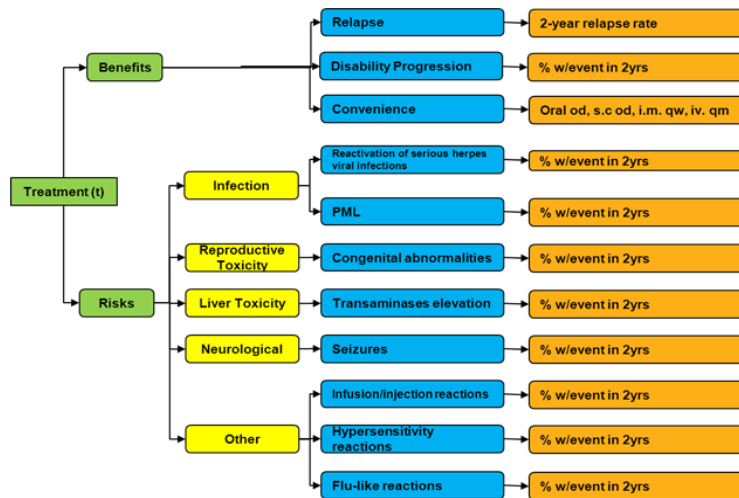| Name/rubric: | Value function for convenience (administration route) |
|---|---|
| Created in: | Microsoft Excel©. Can also be created in most software |
| Intended audience: | Statisticians, Regulators, not for Physicians or Patients |
| Message: | Preference values of a categorical variable (administration route), ordered semi-arbitrarily from high preference to low preference. |
| Knowledge required: | The terminology used, in this case route of administration, and the abbreviations, need to be understood to make sense of the graph. |
| Unintentional message: | Erroneously displays a categorical variable as if it were continuous and for the untrained eye seems to assume a natural ordering, which might in reality be considered post-hoc. |
| Message not communicated: | The role of the legend "Group" is not communicated and is misleading since there is no grouping to be seen on the graph. Also, see below and above. |
| Proposed improvement: | Use bar chart or dot plot instead, making clear that the variables are categorical not continuous. At the very least, the connecting line needs to be removed. |
| Comment | The visual is taken into Phase II visual methodology work as an example of the likely improvement for the representation of a discrete/categorical variable. |

## Line graph 3

**Two-way sensitivity analysis plot**



| | |
|---|---|
| **Name/rubric:** | Two-way sensitivity analysis plot. |
| **Created in:** | R |
| **Intended audience:** | Statisticians and regulators. |
| **Message:** | To show how the changes in both the number of patients developing PML and the weight associated with PML affect the benefit-risk score. It shows the sensitivity of benefit-risk balance, for values of a variable (here % of patients with PML) for different weights of that parameter. |
| **Knowledge required:** | Need to understanding of concept of weights, as well as know the probabilities of PML events. Users also need to know that negative BR values represent 'poorer' outcomes. |
| **Unintentional message:** | Low "PML weight" might be interpreted as lower preference to experiencing PML. |
| **Message not communicated:** | The choice of particular weight values is unclear from the graph. There is also no mention that the lines only represent benefit-risk scores against plausible proportions of patients who may experience PML with treatment. |
| **Proposed improvement:** | The graph could be somewhat confusing to someone with a lack of background knowledge on the problem at hand; perhaps more explanation in the way of a title or annotation is required. Explanation of weights choice is also required. More distinct colours should be used to discriminate the effects of different weights. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too simple to benefit from interactive visualisation |

## *Value tree 1*

**Value tree from BRAT framework**



| Name/rubric: | Value tree |
|---|---|
| Created in: | BRAT Lotus notes program plus PowerPoint, but can also be created in PowerPoint alone |
| Intended audience: | Regulator and Physician. Also for patients if background is provided on the medical terms in lay language. |
| Message: | To display qualitative listing of available key benefits and risks criteria for the decision model and the description of their measurements in a hierarchical way. Colours are used to indicate level of criteria. |
| Knowledge required: | Needs Medical knowledge (terminology of the medical terms; for judging whether selected benefits and risks make sense), no statistical knowledge needed. The tree could be amended for the patients as well using lay terminology to be easily readable. |
| Unintentional message: | As usually benefits are predetermined, and on the risk side, potential risks are also included so there is a tendency to show more risks than benefits. For the inexperienced reader this could be perceived as "more" risks if no further qualification of the risks is provided. This could be done by more explicit description of certainties, numbers are not weighted but weights could be added alongside. As presented, each criterion may be seen as being equally weighted. |
| Message not communicated: | The context in which the value tree is created is needed (underlying database, indication, what decision). There are no numerical data and no information on the treatment alternatives. |
| Proposed improvement: | Some background information, such as definitions for the indications should be included via text annotation. Explanation of whether the order of criteria reflects weighting would be useful. The naming of criteria must intuitive reflect their role as a benefit or a risk of treatment, for example write "relapse" as "reduction in relapse rate" to reflect it is a benefit of treatment. Different colours should be used to distinguish benefit and risks. Risks should be characterised in terms of certainty, and lay terms to be added to make the value tree more useful for patients. |
| Comment | The visual is taken into Phase II of visual methodology work to illustrate how text annotations and colour choices may help improve the understanding of a value tree in an interactive visualisation |

*Value tree 2*

**Value tree from BRAT framework with added preference weights**
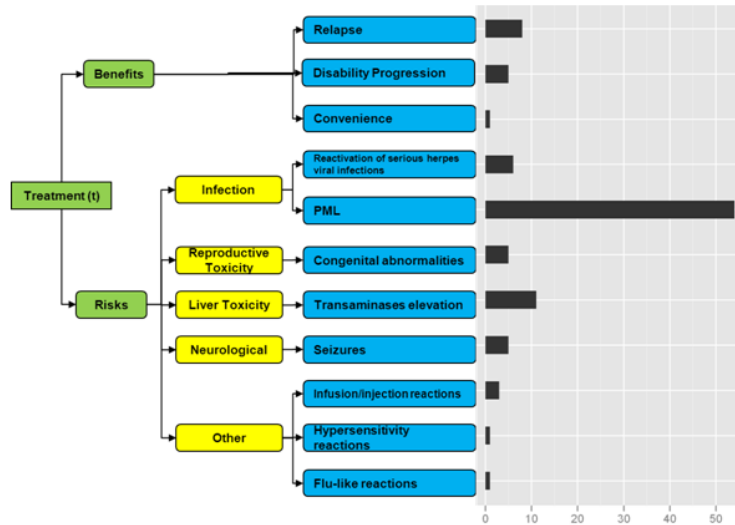


| | |
|---|---|
| **Name/rubric:** | Value tree with preference weights |
| **Created in:** | BRAT Lotus notes program plus Microsoft Power Point |
| **Intended audience:** | Regulators , Physicians, not patients |
| **Message:** | To display qualitative listing of available key benefits and risks criteria for the decision model and the description of their measurements in a hierarchical way. Patient preferences on each criterion are also displayed. Colours are used to indicate level of criteria. |
| **Knowledge required:** | Needs Medical knowledge (terminology of the medical terms; for judging whether selected benefits and risks make sense), no statistical knowledge needed. Understanding of the meaning and role of weights on the value tree is required. |
| **Unintentional message:** | Risks are weighted very high compared to benefits. Since the probability of event is missing, individuals may quickly judge that risk clearly outweighs benefit. |
| **Message not communicated:** | The way parameters were measured in the trial (ALT Units etc.) is unclear (but included in value tree). It is also unclear which treatments are being compared, and whose preference weights are being represented. |
| **Proposed improvement:** | The axis for preference weights should be labelled and the scale is to be made explicit. Whose preference weights are being represented should be made clear through the axis label. The criteria could be ordered by magnitude of weights within the hierarchy. The graph should be accompanied by the frequency of events to allow for appropriate judgement to be made. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too similar to value tree 1, which has been carried forward. |

## *Dot/Forest plot*

**Forest plot from BRAT framework**
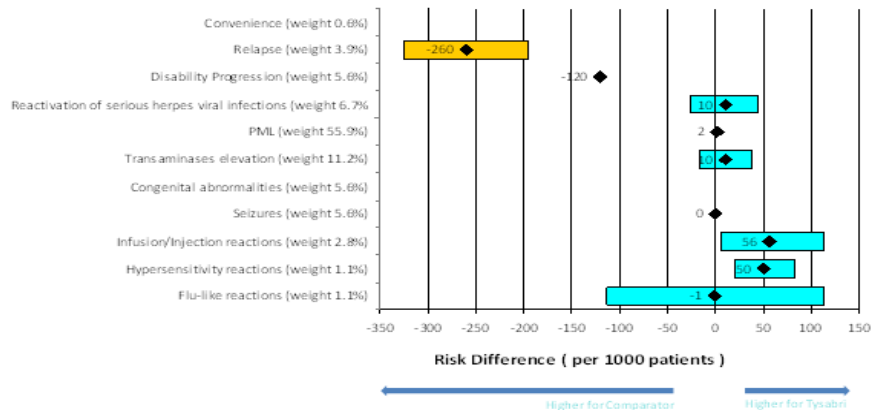


| Name/rubric: | Forest plot on absolute risks |
|---|---|
| Created in: | BRAT Excel Tool adapted in Microsoft PowerPoint, but could also be done in PowerPoint alone. |
| Intended audience: | Statisticians, Physicians, Regulators; not for patients |
| Message: | It shows the risk difference between natalizumab and placebo at time of CHMP re-evaluation on individual key benefits and risks criteria. The forest plot leads to judgement of positive benefit risk balance. Absolute numbers as presented are more supportive to allow judgements when compared to relative values. Statistical uncertainty of the risk difference is also given. The forest plot is a simple visual graph, and is easier to comprehend when compared to a table with the same information. |
| Knowledge required: | Needs medical knowledge, low statistical knowledge |
| Unintentional message: | Risk of very small but important events may be undervalued |
| Message not communicated: | The weights appeared with criteria name are confusing since it is unclear how they are to be taken into account in this graph. "Convenience" and "Congenital abnormalities" criteria appear on the forest plot but are not quantified. In general, non-categorical values are difficult to capture i.e., 6 min walk test, time aspect not covered as for example in Kaplan Meier curves |
| Proposed improvement: | Replace "higher to" by "favours" if the latter can correctly describe the direction, or otherwise describe whether "higher" is desirable or undesirable outcome. Use green and red instead of blue and yellow with clearer text colours (black is likely to be the most visible here). The plot could have dual horizontal axis to represent both continuous and categorical outcomes. A second graph can be added where weights have been incorporated into the measure of differences. |
| Comment | The visual is taken into Phase II of visual methodology work to demonstrate how a forest plot can be used more efficiently as an interactive graph. |

## *Bar chart 1*

**A simple bar chart**
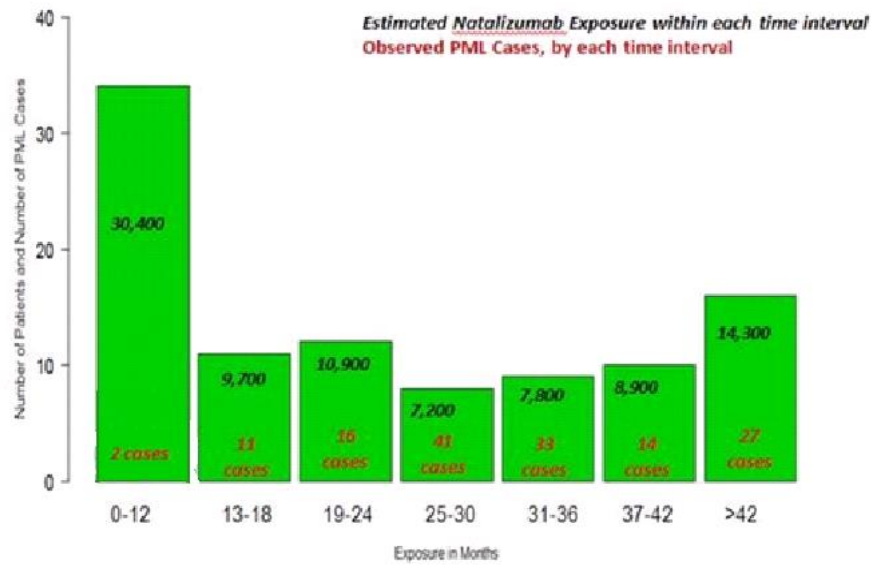


PML case Data: from Dr. B. K. Stanislawski, FDA/EMA PML Workshop, July 25-26, 2011.

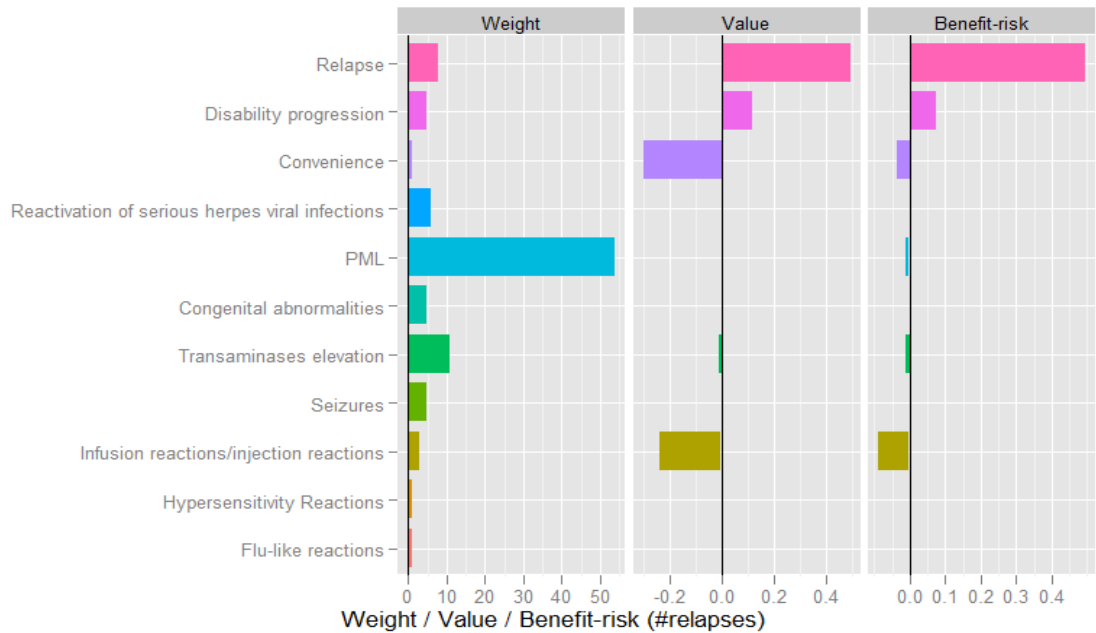| | |
|---|---|
| **Name/rubric:** | Simple bar graph |
| **Created in:** | Extracted from external source and edited in Microsoft Word |
| **Intended audience:** | All |
| **Message:** | The bar chart communicates the distribution of PML cases from post-marketing exposure of natalizumab by number of months of exposure. A short exposure (0-12 months) to natalizumab is the most common. The rate of PML cases increases with longer use up to three years. |
| **Knowledge required:** | Low statistical and medical knowledge. Users need to understand that only the height of the bar is to be interpreted not the area. |
| **Unintentional message:** | The bar value label does not match the values on the vertical axis which would confuse the users. It gives the impression that those in 25-30 months category had the lowest PML rate. The width of the bars may be confused as being of the same range. |
| **Message not communicated:** | It is unclear whether "PML cases" is the number of PML incidence or number of patients who experienced PML regardless of number of PML events. |
| **Proposed improvement:** | Colours of bars and labels should be chosen more carefully to increase contrast so that users would be able to read them easily. Stacked bar graphs (PML + non PML) of the percentage may be more suitable to represent the information in the current graph. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too simple to benefit from interactive visualisation |

## *Bar chart 2*

**Aligned bar chart of utility values by treatment and their difference**
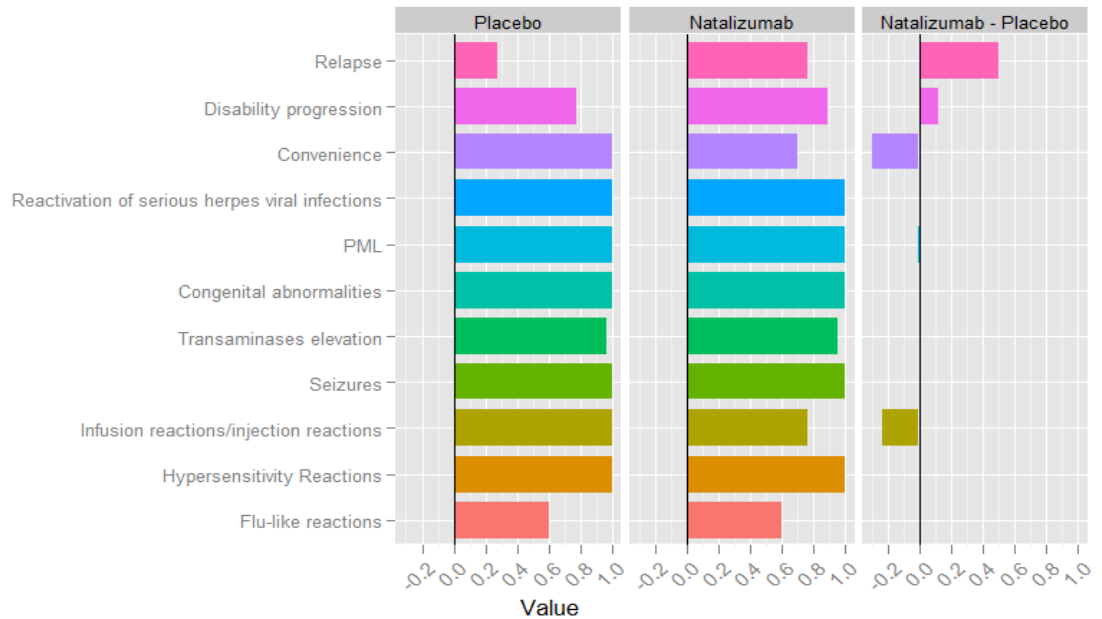


| Name/rubric: | Aligned bar chart with difference display |
|---|---|
| **Created in:** | R (ggplot2 package) |
| **Intended audience:** | Statisticians, Regulators not for Physicians or patients |
| **Message:** | The bar chart shows the numerical key benefits and risks values by criterion and comparative treatments, and the difference between them. The meaning of the values is unclear. The lengths of the bars appear to be the same. |
| **Knowledge required:** | Needs in depth knowledge of MCDA and value functions to understand the message properly. But no specific technical knowledge is required to determine on which criteria natalizumab is valued higher or lower than placebo from the difference display (right-most column) |
| **Unintentional message:** | There is a lack of transparency and gives an impression of complexity. There is very minimal benefit risk balance. The importance of PML may be underestimated from the value difference. |
| **Message not communicated:** | More explanation on whose values and what do the values mean are needed. Statistical uncertainty is not described |
| **Proposed improvement:** | Harmonise colours to give meaningful message. Add sensitivity analysis. The meaning of values and their difference need to be explained to aid interpretation. The horizontal axis on the difference display should be labelled with 'Favours natalizumab' and 'Favours Placebo' appropriately. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since the real message is quite confusing and may not be very useful for benefit-risk assessment. Alternative aligned bar chart is described in Section . |

## Bar chart 3

**Aligned bar chart of the derivation of the difference in weighted values**



| Name/rubric: | Bar Chart (of Weights, Values and Benefit Risk Scores) |
|---|---|
| Created in: | R (ggplot2 package) |
| Intended audience: | Statistician, Regulators, Physicians, Patients |
| Message: | The bar chart shows the effect of combining specific weights with difference in values for the numerical key benefits and risks comparing natalizumab to placebo. |
| Knowledge required: | Needs in depth knowledge of MCDA and value functions or a lot of trust. But no specific technical knowledge is required to determine on which criteria natalizumab performs better or poorer than placebo from the difference display (right-most column) |
| Unintentional message: | Rare severe risks may be underestimated. |
| Message not communicated: | Statistical uncertainty is not communicated making the results seem too certain. Confidence intervals from clinical study data have not been considered; thus results may not reflect the clinical data in whole. |
| Proposed improvement: | Add sensitivity analysis. Confusion by normalisation – relapse was scaled to 1, so needs better explanation. Colours to be harmonised to provide more meaningful message. |
| Comment | The visual is taken into Phase II of visual methodology work since it the representation is found to be useful during decision conferencing in Wave 1 case study, but could be improved further to allow better user comprehension. |

## *Bar chart 4*

**A stacked bar chart of incremental benefit-risk**
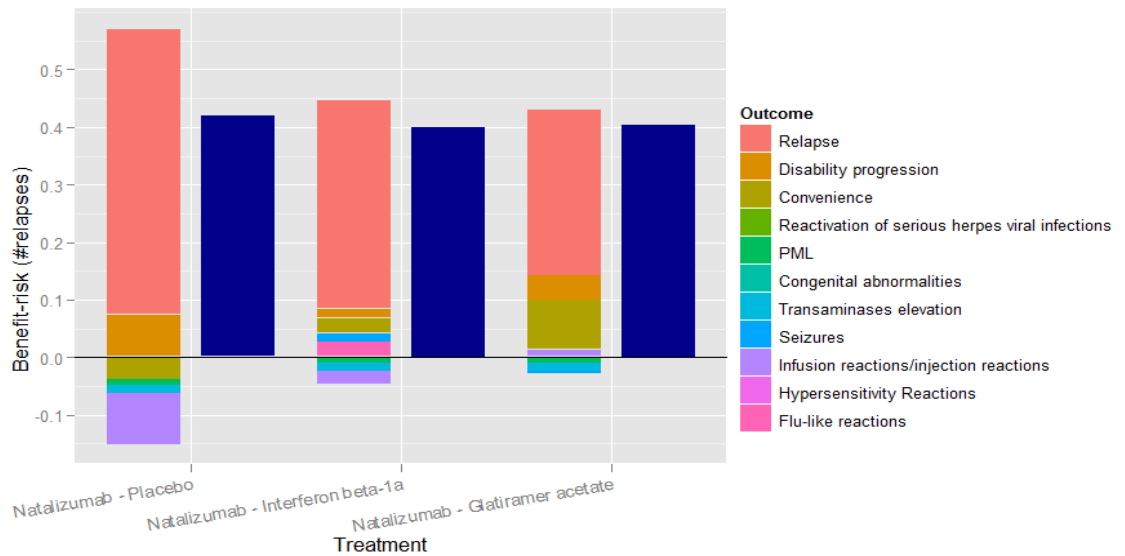


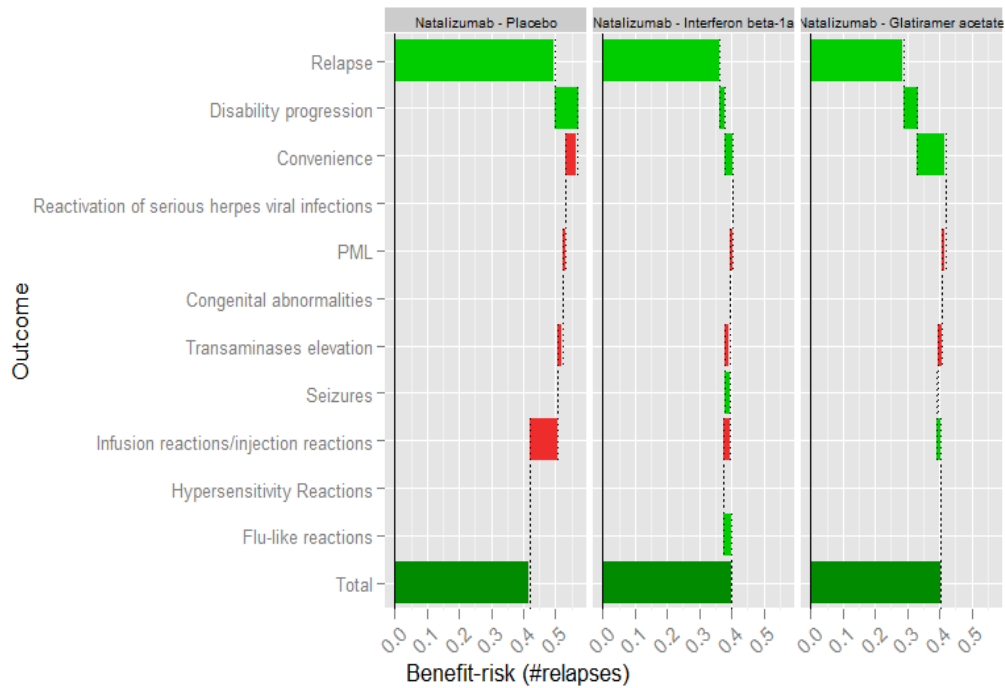| Name/rubric: | Stacked bar chart (of incremental benefit risk) |
|---|---|
| **Created in:** | R (ggplot2 package) |
| **Intended audience:** | All |
| **Message:** | The bar chart shows the incremental benefit-risk contribution by criterion for natalizumab compared to placebo, Beta-interferon and Glatiramer acetate. It displays the magnitude and direction of contributed criteria to the overall score. It also displays the magnitude and direction of the change in benefit values having discounted risk by criteria and treatment (dark blue bar). |
| **Knowledge required:** | No specific technical knowledge is required. Some understanding of the incremental benefit-risk concept i.e. the knowledge that negative values equate to decreased benefit-risk balance, and positive values equate to increased benefit-risk balance. |
| **Unintentional message:** | No uncertainty in the benefit-risk balance is presented. |
| **Message not communicated:** | Very small scores may not be visible, giving the impression that natalizumab exactly equals comparator on the criterion, which may lead to misinterpretation to users. The dark blue bars are not labelled, so the quantity as to what is presented is unknown. |
| **Proposed improvement:** | Dark blue bars to be labelled. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since it is too similar to bar chart in Section 0 which was found to be more useful. |

## Bar chart 5

**Aligned bar chart of the differences between natalizumab and comparators represented as waterfall plot**



| Name/rubric: | Waterfall plot (of incremental benefit risk) |
|---|---|
| **Created in:** | R (ggplot2 package) |
| **Intended audience:** | All |
| **Message:** | The waterfall plot shows the incremental benefit-risk of natalizumab compared to placebo, Beta-interferon and Glatiramer acetate. It displays the magnitude and direction of benefit-risk contribution to the overall score by criterion. |
| **Knowledge required:** | Some understanding of incremental benefit-risk concept. Some knowledge of waterfall plot e.g. which end of the bars to be read (green = right, red = left). The grids act as reference lines for comparison across panels. |
| **Unintentional message:** | Some bars are very narrow – users cannot tell whether information is missing or very small values, or the direction of the difference. |
| | Inexperience users may find it difficult to extract what the final benefit-risk balance is e.g. green at the bottom may be perceived as positive balance, and red as negative balance which may not always be true. |
| **Message not communicated:** | Very small scores may not be visible, giving the impression that natalizumab exactly equals comparator on the criterion, which may not be true since it could be either way and may lead to misinterpretation to users. The overall value is not explicitly communicated. |

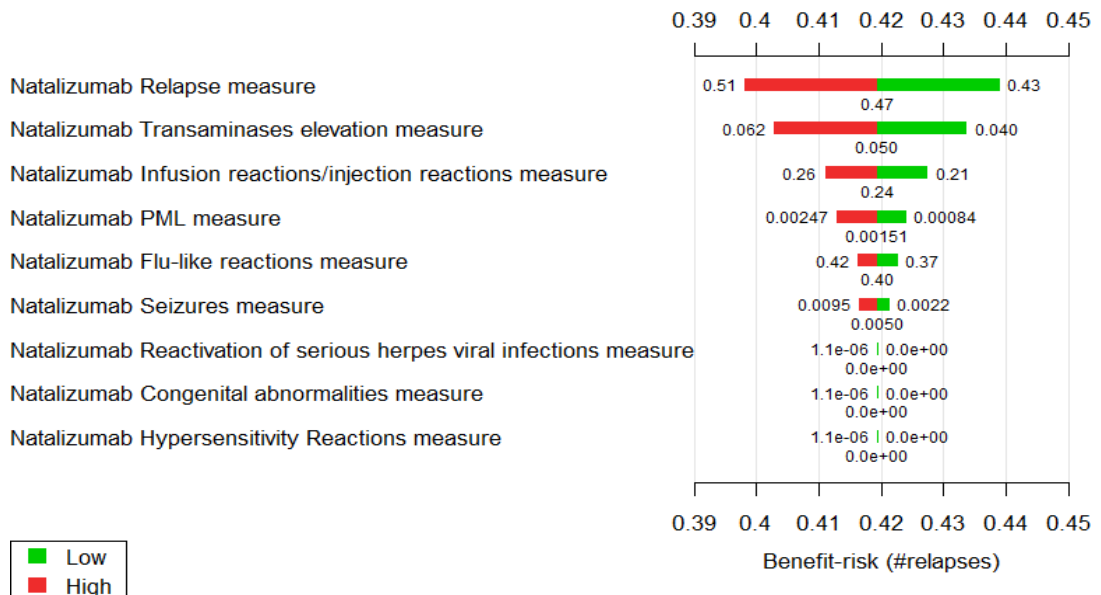| | |
|---|---|
| **Proposed improvement:** | The horizontal axis should be made wider to enhance readability since the benefit-risk values are represented along this axis. Overall scores should be presented. Ordering of criteria should be more meaningful. |
| **Comment** | The visual is taken into Phase II of visual methodology work since it was found to be a useful visualisation tool alternative to stacked bar chart, but require more space to represent the same amount of information. |

## Bar chart 6

**Tornado plot**



| | |
|---|---|
| **Name/rubric:** | Tornado plot |
| **Created in:** | R |
| **Intended audience:** | Statisticians and regulators. Not for physicians and patients. |
| **Message:** | The tornado plot shows how the changes in the natalizumab outcome measure affect the incremental benefit-risk score. It displays the relative importance of criteria via one-way sensitivity analysis of changing a fixed amount of the measured outcomes. |
| **Knowledge required:** | Some knowledge on the use of sensitivity analysis and uncertainty. Some understanding of the incremental benefit-risk concept. Some knowledge on how to extract information from tornado diagrams. |
| **Unintentional message:** | The legend of "high-low" is not intuitive and could be misleading. |
| **Message not communicated:** | It is unclear which of the criteria are benefits and which are risks. The colour-coding is not intuitive and difficult to interpret. |
| **Proposed improvement:** | Horizontal axis should be made wider to accommodate benefit-risk values. To re-label legend items to more intuitive terms. The tornado plot could also be accompanied by text annotations to aid interpretation. |
| **Comment** | The visual is taken into Phase II of visual methodology work as a method to visually represent deterministic uncertainty. |

## *Thermometer scale*

**Thermometer scale**



| Name/rubric: | Thermometer scale |
|---|---|
| **Created in:** | Microsoft Word |
| **Intended audience:** | Any stakeholders |
| **Message:** | Not applicable. It is used to elicit stakeholder's relative importance on criteria |
| **Knowledge required:** | Some medical knowledge of the condition/event, in this case relapse rate and disability progression scoring. |
| **Unintentional message:** | Not applicable. |
| **Message not communicated:** | Not applicable. However, some stakeholders may want to provide values with uncertainty which could be accommodated. |
| **Proposed improvement:** | If the elicitation is not done in a facilitated session, clearer instructions are needed. |
| **Comment** | The visual is not taken into Phase II of visual methodology work since Phase II focuses on displaying the results than for eliciting preference values. |

*[End of Report]*