Pharmacoepidemiological Research on Outcomes
of Therapeutics by a European ConsorTium

Review of methodologies for benefit and risk assessment of medication

Prepared on behalf of the PROTECT Consortium by

Shahrul Mt-Isa[1], Nan Wang[1], Christine E. Hallgreen[2], Torbjörn Callréus[3], Georgy Genov[4], Ian Hirsch[5], Steve Hobbiger[6], Kimberley S. Hockley[1], Davide Luciani[7], Lawrence D. Phillips[4], George Quartey[8], Sinan B. Sarac[2], Isabelle Stoeckert[9], Alain Micaleff[10], Deborah Ashby[1], Ioanna Tzoulaki[1]
On behalf of PROTECT Work Package 5 participants

[1] School of Public Health, Imperial College London, London, United Kingdom
[2] Novo Nordisk A/S, Søborg, Denmark
[3] Danish Health and Medicines Authority, Copenhagen, Denmark
[4] European Medicines Agency, London, United Kingdom
[5] AstraZeneca AB, Macclesfield, United Kingdom
[6] GlaxoSmithKline Research and Development LTD, Middlesex, United Kingdom
[7] Mario Negri Institute for Pharmacological Research, Milan, Italy
[8] Genentech, South San Francisco, USA
[9] Bayer Schering Pharma AG, Berlin, Germany
[10] Merck KGaA, Geneva, Switzerland

Version 4 Date: 14 Feb 2012 Date of any subsequent amendments below	Person making amendments	Brief description of amendments
10 April 2013	Shahrul Mt-Isa	Table 4 was updated with new information

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

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

Executive summary

Background

Pharmacoepidemiological Research on Outcomes of Therapeutics in a European Consortium (PROTECT) is a project, set up under the Innovative Medicines initiative, with the aim of strengthening the monitoring of the benefit-risk of medicines in Europe. 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 and pharmaceutical companies. Information on risk and benefits of drugs comes from diverse sources through the life-cycle of drugs and merging all different sources of data is a challenging task due to the nature of the data, data quality and potential biases. Structured approaches to decision making which may help regulatory decision making have recently gained attention. In parallel, more quantitative statistical approaches have evolved over many decades that may draw together disparate data from the many sources and formally synthesise them with relevant preference values from stakeholders and present them in ways that can aid decision-making. Nonetheless, there is an absence of widely accepted methods for the quantification of benefit-risk and the visual representation of benefit-risk profiles.

Objective

The purpose of this review is to draw together the methods currently proposed for benefit-risk assessment of drugs. In addition, we aimed to show the inter-relationships between different methods, to group methodologies appropriately in order to provide a clear structure for future reference and to evaluate the capacity of each method. Finally, we sought to provide recommendations on which methods merit further consideration for decision-making about medicines and in which contexts. These methods will then be further explored in the next stage of the PROTECT project.

Literature search

Since comprehensive reviews on benefit-risk methodologies have already been undertaken, we performed a review of existing reviews either published in the medical literature or through regulatory agencies or pharmaceutical companies. Qualitative and quantitative methodologies were eligible as well as benefit-risk methodologies within pharmacoepidemiology and clinical trials but also within other fields such as Health Technology Assessment (HTA). Materials were located through PubMed, work carried out by PROTECT members, work from other parallel benefit-risk initiatives and pharmaceutical companies' internal methods. The review team went through the list of benefit-risk approaches extracted from all identified reviews during a face-to-face meeting; and excluded approaches that were not considered relevant for this appraisal.

Appraisal

Appraisal of existing methods was based on four predefined criteria: *Principle* (if the reasoning is theoretically correct); *Features* (number of criteria, number of options, capacity to deal with uncertainty); *Accessibility* (if it is easy to use or not); and *Visualisation* (the proposed visual representation of the results and if there is software to implement them).

Classification

Forty-seven different approaches were extracted from the identified reviews and individual papers. These were grouped into *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 which are descriptive frameworks (non-quantitative) and those that also provide comprehensive quantitative trade-off approaches. Metrics were subdivided into those that we have termed threshold indices, that have been explicitly developed in a various health contexts, termed here health indices, and those that explicitly allow trade-offs. The estimation techniques, which deal with bringing together evidence systematically, range from simple simulations to cutting-edge statistical methods for synthesis of complex data from multiple sources. Finally, the utility survey techniques complement those approaches which need explicit elicitation of values and preferences of various outcomes from the relevant stakeholders.

The associated parameters and elementary features of each benefit-risk approach were summarised in tables of 'operational characteristics' to assist decision-makers making preliminary assessment of the potential and attractiveness of different approaches.

Appraisal of frameworks

Descriptive frameworks are generic stepwise instructions which do not offer much more than reiterate common sense, but do so systematically. PrOACT-URL provides a framework to promote systematic considerations of the important elements in a decision problem. Ashby and Smith framework (ASF) is a simpler framework addressing the basic elements of decision theory including sources of evidence and relevant stakeholders. Collaboration between the pharmaceutical industry and the regulatory agencies has resulted in the Benefit Risk Action Team (BRAT) framework for selecting, organizing, understanding and summarising benefit-risk evidence. BRAT serves to improve communications of benefits and risks between the pharmaceutical companies and the regulators. MCDA framework, as reviewed, applies PrOACT-URL to ensure transparency and trades off benefits and risks criteria through the realisation of value tree and multi-attribute utility theory. However, MCDA lacks the ability to account for sampling variation in the criteria measurements; and its validity can be threatened in the case when preference information is missing or when consensus is not reached. Another variant of MCDA, the stochastic multi-criteria acceptability analysis (SMAA) proposes a way to overcome these limitations through simulations. MCDA – in general – is also the only approach that can formally deal with multiple objectives at the same time.

Appraisal of metric indices

Metric indices are numerical representations of benefits and risks, and their use in benefit-risk decision making must in the context of other approaches and high quality data. The popular concept of the number needed to treat and to harm (NNT and NNH) represents the number of patients to be treated to see one benefited or one harmed, respectively. It has been criticised in the academic literature for not being logically sound, but its popularity among a wider audience has illustrated its appeal in decision making. The NNT concept has been extended to metrics known as the 'impact numbers' to quantify benefits or risks within the population. The interpretation of impact numbers is specific to the circumstances where they are derived from. The health indices such as quality adjusted life years (QALYs) start to trade off benefits and risks with respect to time gained or lost; and are widely used in health research. The derivation of QALYs however can be variable; and mostly derived from generic health states which may not be appropriate in some disease populations. A specialised health index in cancer epidemiology is the quality adjusted time without symptoms and toxicity (Q-TWiST), which are derived from utilities of being in three health states as a result of a cancer treatment. The incremental net health benefit (INHB) index allows the changes in a treatment’s benefits to be discounted by the changes in the treatment’s risks measured by health indices, when compared to an alternative treatment. INHB thus provides the classical trade-off form many people are familiar with. However, a difference metric cannot intuitively represent the trade-off of benefits and risks. The more intuitive metric would be a ratio metric index like the benefit-risk ratio (BRR). BRR is simply the ratio of benefits to risks which are usually measured by the probabilities or rates of the two, and interpreted as the multiplicative effects of benefits over risks.

Appraisal of estimation techniques

Simple benefit-risk models based on limited data summaries require only basic estimation techniques, but more sophisticated methods, or multiple sources of data require more complex estimation techniques. The probabilistic simulation method (PSM) – including the more computer intensive Monte Carlo simulation – allows uncertainties in the input values, characterised by probability distributions, to be propagated through the network of evidence to the end results. Good quality evidence linking the options, actions and consequences are required to reach definitive conclusions on the trade-off values. Direct evidence on treatment comparisons (e.g. a head-to-head trial) can be augmented by indirect evidence through the application of the mixed treatment comparison (MTC). MTC also addresses methods of handling covariance structure in more complex cases when data from trials provide comparisons of three or more treatment options.

Appraisal of utility survey techniques

Utility survey techniques refer to the method of elicitation of preference values from the relevant stakeholders. In this case, they might be the patients, physicians, healthcare providers, pharmaceutical companies or the regulatory agencies. Many techniques of preference elicitation are available, and those reviewed here are a specific group of techniques based on stated preference. The gold standard for this group of techniques is the discrete choice experiment (DCE), which provides structured guidelines for the entire process. In DCE, aspects of experimental designs, scoring systems and methods to combine preference values are addressed.

Conclusion

This review is an extension of many previous reviews on quantitative approaches for benefit-risk assessment. It is different from past reviews in the sense that the approaches are classified based on their characteristics, thus brings to light the dominant features of a particular category of approaches as well as the issues that are associated with their applications. Choosing one approach for every decision problem is difficult and unrealistic because each has their strengths and weakness, and sometimes their pragmatic applications are limited by available evidence and underlying assumptions; and most often are limited by the resources – monetary, time, knowledge – and the ability to effectively communicate the results from a benefit-risk analysis.

The consensus from the PROTECT team is to recommend the following approaches for further considerations and testing for the applications in benefit-risk decision making:

Descriptive Frameworks

(1) PrOACT-URL frameworks. PrOACT-URL provides a framework addressing the necessary elements in dealing with decision problems. PrOACT-URL however does not clearly address the importance of identifying appropriate sources of evidence and immediate parties involved. The application of PrOACT–URL therefore requires improvement in these two aspects, for example by combining with elements of the Ashby and Smith framework (ASF) or other evidence synthesis approach. The extension of PrOACT-URL to drug benefit-risk decision analysis by the EMA Benefit-Risk Methodology working group 2 may, in the future, provide better evolution of the frameworks in this specific domain.

(2) Benefit Risk Action Team (BRAT) framework. BRAT provides guidelines on organising, understanding and summarising evidence of benefits and risks into tabular outputs and graphical summaries. The framework proposes avoiding integration of benefits and risks evidence to make it more accessible and transparent to those not familiar with complex statistical models. The controversial aspect of BRAT is its proposal to use odds ratios as the basis for the decision on benefits and risks balance.

Quantitative Frameworks

(3) Multi-Criteria Decision Analysis (MCDA). MCDA provides structured stepwise instructions in line with PrOACT-URL framework with the capability of assessing and integrating multiple benefits and risks criteria, as well as comparing different options. MCDA is also the only approach that can formally deal with multiple objectives simultaneously. Another appealing feature of MCDA is that specialist several software to perform the analysis are available.

(4) Stochastic Multi-criteria Acceptability Analysis (SMAA). The SMAA families of approaches are potentially serious contenders to the standard MCDA because of the ability to account for sampling variations arising from the type of experimental designs used as well as when there is missing information on preference values. However, the increased complexity of SMAA compared to MCDA may be the major barrier in real-life benefit-risk medical decision making applications. The application of SMAA in medical decision making should be explored further as there is a potential that the same SMAA model could be used to address different stakeholders.

Metric Indices

(5) Number Needed to Treat (NNT) and Number Needed to Harm (NNH). The popularity and its widespread use in clinical literature when describing the benefits or risks of a treatment are well established. These indices, heavily criticised, still provide an attractive feature for benefit-risk assessment – simplicity. They should not be used naively, but be supported by thorough evidence synthesis and suitable modelling of the evidence. We are not recommending the utility-adjusted variants of NNT because the reciprocals of expected utilities do not have the same meaning as the reciprocals of probabilities or rates.

(6) Impact numbers. These metric indices give a different perspective from NNT, and are useful in describing public health burden of a disease, and the potential impact of a treatment. Although other more established epidemiological measures are available and have already been used widely, impact numbers have an intuitive interpretation. Their application in benefit-risk assessment may contribute to making the interpretation more accessible to the general audience.

(7) Quality Adjusted Life Years (QALY). The QALY provides time trade-off with life quality in discrete health states which is absent from other more general trade-off indices. Its use is already established in many areas of medicine particularly in chronic diseases where time factor plays a major role in their assessment of benefits and risks. HALE may be another health index to be considered, but as it is just a summary of QALY in a particular group of people, we will not differentiate it further.

(8) Quality adjusted Time Without Symptoms and Toxicity (Q-TWiST). The discrete health states proposed in Q-TWiST are intuitive and very specific to cancer therapy. Therefore its usefulness is limited to cancer domain but nonetheless is a suitable metric to aid cancer patients to decide on the best acceptable treatment. We are only recommending the use of Q-TWiST within the cancer therapy domain.

(9) Incremental Net Health Benefit (INHB). The INHB builds on health indices like QALY. INHB incorporates time and utility which are desirable elements in benefit-risk assessments. The idea of penalising "incremental" benefits by "incremental" risks in INHB, directly follows the simple intuitive concept of benefit-risk trade-offs.

(10) Benefit-Risk Ratio (BRR). Ratios provide intuitive trade-off metrics as they can be interpreted readily as the multiplicative effect of one relative to the other. The recommendation here is only for the application of BRR when accompanied by high quality evidence data and appropriate statistical modelling, and is presented together with their absolute or baseline rates. The weighting of benefits and risks to be traded off should be carefully taken into consideration when using BRR as the metric for benefit-risk assessment because equal weights assumption may not always, and in most cases do not, hold.

Estimation Techniques

(11) Probabilistic Simulation Method (PSM). The abilities of PSM to deal with statistical adjustments and different kind of uncertainties under different assumptions are its most appealing features. The use of high quality evidence data for the application of PSM is desirable as the end results of the simulations are highly dependent on the input values and the assumptions in the underlying models.

(12) Mixed Treatment Comparison (MTC). MTC is recommended here as the method of choice because it can flexibly accommodate many important aspects of evidence synthesis in different circumstances. Issues of biases, as addressed in the confidence profile method should be considered in an MTC model as should other issues of combining different types of evidence addressed in cross design synthesis.

Utility Survey Technique

(13) Discrete Choice Experiment (DCE). The utility survey techniques from the stated preference methods family are suitable for the purpose of eliciting utilities but we recommend DCE as an approach to elicit utilities. This is simply because these approaches are similar and roughly based on the same principles but DCE provides the most comprehensive instructions on the steps required from designing the elicitation process to the methods of combining the utilities obtained to be used in a benefit-risk assessment model. The use of DCE in drug benefit-risk decision making is less mature than other recommended approaches but its potential needs further exploration. Conducting a DCE is resource-consuming. Therefore alternatives would be required. A structured elicitation is encouraged because ultimately obtaining appropriate value judgments from relevant stakeholders is crucial to the validity of any specific benefit-risk decision analysis.

Contents

Glossary and abbreviations

Term	Description
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 make choice among 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
Confounding factors	Factors that affect the outcome together with other factors
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 does more good than harm when provided under the usual circumstances
Efficacy	The extent to which an intervention does more good than harm under ideal circumstances
Elicitation	The process through which relevant notions for a problem of interest are made explicit
Framework	Structured stepwise approach to perform a task
Graphical methods/ representation	The principles and procedures to present some numerical features or relations by a graph
Health technology assessment	Analysis of the medical, economic, social and ethical implications of the incremental value, diffusion and use of a medical technology in health care
Incidence	The frequency of the first occurrence of an event or a condition in a specified period
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	System of measurement
Pharmacoepidemiology	The study of the use and effects of drugs in well-defined populations
Preference values	A quantitative measure of the extent to which an outcome achieves an objective, as judged by an individual or group.
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.
Quantitative	Involving considerations of amount or size; capable of being measured

Term	Description
Revealed preference	An approach which observes and explores preferences as indirectly revealed by an individual’s action(s) within real life situations
Stated preference	An approach which asks individuals to state their preferences within hypothetical scenarios
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, usually 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
Score	Numeric values with fixed minimum and maximum (bounded scales) assigned to distinguish magnitude, severity, performance, preference etc.
Uncertainty	Uncertainty may refer to Randomness, the possibility of different outcomes from an action, which cannot be foreseen for sure in advance. Uncertainty in estimation due to insufficient sampling. Discrepancy in evidences from different sources of data. Measurement error or quality of data (for example data not measured by proper means, or poor equipment, etc.).
Utility	A subjective measurement that describes a person’s or group’s preferences (satisfaction, risk attitude etc.) for an outcome.
Value function	A function which convert the input data (parameters) in all criteria into preference value or utility for the options under evaluation.
Value judgment	A subjective assessment for appropriateness of values or utility in a decision making problem
Weight	Scaling constants assigned to criteria such that the units of scaled preference values across all criteria are equal.

Methodology abbreviations	Description
AE-NNT	Adverse Event adjusted Number Needed to Treat
ASF	Ashby and Smith Framework
BLRA	Benefit Less Risk Analysis
BM	Beckmann Model (aka Evidence Based Model)
BRAT	Benefit Risk Action Team
BRR	Benefit Risk Ratio
CDS	Cross Design Synthesis
CIN	Case Impact Numbers
CMR CASS	CMR Health Canada, Australia's Therapeutic Goods Administration, SwissMedic, and Singapore Health Science Authority
COBRA	Consortium on Benefit Risk Assessment
CPM	Confidence Profile Method
CUI	Clinical Utility Index
CV	Contingent Valuation
DAG	Directed Acyclic Graphs
DALY	Disability Adjusted Life Years
DI	Desirability Index
DIN	Disease Impact Numbers
ECIN	Exposed Cases Impact Numbers
EIN	Exposure Impact Numbers
FDA BRF	FDA Benefit Risk Framework
GBR	Global Benefit Risk
HALE	Health Adjusted Life Years
INHB	Incremental Net Health Benefit
ITC	Indirect Treatment Comparison
MAR	Maximum Acceptable Risk
MAUT	Multi Attribute Utility Theory
MCDA	Multi Criteria Decision Analysis
MCE	Minimum Clinical Efficacy
MDP	Markov Decision Process
MTC	Mixed Treatment Comparison
NEPP	Number of Events Prevented in the Population
NCB	Net Clinical Benefit
NEAR	Net Efficacy Adjusted for Risk
NNH	Number Needed to Harm
NNT	Number Needed to Treat
PATHS	Preliminary Assessment of Technology for Health Services
PIN	Population Impact Numbers
PIN-ER-*t*	Population Impact Numbers of Eliminating a Risk Factor over time *T*
PrOACT-URL	Problem, Objectives, Alternatives, Consequences, Trade-offs, Uncertainty, Risk, and Linked decisions framework
PSM	Probabilistic Simulation Method
QALY	Quality Adjusted Life Years

Methodology abbreviations	**Description**
Q-TWiST	Quality-adjusted Time Without Symptoms and Toxicity
RV-MCE	Relative Value adjusted Minimum Clinical Efficacy
RV-NNT	Relative Value adjusted Number Needed to Treat
SABRE	Southeast Asia Benefit Risk Evaluation
SBRAM	Sarac's Benefit Risk Assessment
SMAA	Stochastic Multi-criteria Acceptability Analysis
SPM	Stated Preference Method
TURBO	Transparent Uniform Risk Benefit Overview
UMBRA	Unified Methodologies for Benefit Risk Assessment
UT-NNT	Utility- and Time-adjusted Number Needed to Treat

Other abbreviations	**Description**
CIRS	Centre for Innovation in Regulatory Science
CMR	Centre for Medicines Research
EFPIA	European Federation of Pharmaceutical Industry Association
EMA	European Medicines Agency
FDA	Food and Drugs Administration
IMI	Innovative Medicines Initiative
PROTECT	Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium

1 Introduction

1.1 The PROTECT Project

Pharmacoepidemiological Research on Outcomes of Therapeutics in a European Consortium (PROTECT) is a project set up under the Innovative Medicines initiative. Its goal is to strengthen the monitoring of the benefit-risk of medicines in Europe. This will be achieved by developing a set of innovative tools and methods that will enhance the early detection and assessment of adverse drug reactions from different data sources, and enable the integration and presentation of data on benefits and risks. These methods will be tested in 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 between 31 private and public sector partners and is coordinated by the European Medicines Agency (EMA). This report is the first stage of the work on integration and representation of data on benefits and risks.

1.2 Benefits and Risks of Medicine

Evaluation of the balance between benefits and risks of drugs is fundamental to all stakeholders involved in the development, registration and use of drugs. People who need to take drugs for treatment or prevention of disease, and their healthcare providers, need to be able to weigh up the likely benefits and risks of the various therapeutic options before deciding what to take. Regulators in turn must make decisions about benefits and risks of drugs on the basis of trials conducted prior to licensing, at the time of initial licensing review of drug dossiers, and continue to monitor them after approval so that information made available to the public can be updated in the light of emerging benefit-risk information. Pharmaceutical companies have a need to perform benefit-risk assessment throughout the life cycle of the drugs too – pre-marketing and then post-marketing as a basis for decisions on product information changes, and risk mitigation activities to be communicated to regulators, healthcare insurance providers, prescribers and patients. Patients, healthcare providers, regulators and the pharmaceutical industry are all reliant on information from clinical trials (both company sponsored and sponsored by other organisations), while potentially having different values and priorities, so assessments may sometimes lead to apparently conflicting results. Post-marketing, data typically come from spontaneous reporting, industry and/or academia sponsored clinical trials, observational studies or pharmacoepidemiological healthcare databases. It is anticipated that merging the different sources of data can be very challenging due to the nature of the data, where data quality and issue of biases vary.

Balancing benefits and risks has been pivotal to regulatory decision making for many decades but only recently have more structured approaches to decision making gained attention as part of Regulatory Science. In parallel, more quantitative statistical approaches have evolved over many decades that could potentially draw together these disparate data from the many sources mentioned above, synthesise and combine them with a formal assessment of the relevant preference values from stakeholders and present them in ways that can aid decision-making. Such task however is not straightforward; and so far these structured approaches have not been properly developed for use in benefit-risk context in medicine, let alone routinely applied. This absence of widely accepted methods, in particular for quantification of benefit-risk has hampered attempts to integrate both benefits and risks in a clearer or even visual representation of the benefit-risk profile. In this report, we aim to present an overview of existing methodologies for benefit-risk evaluation and to systematically appraise them according to pre-specified criteria. In particular visualisation and communication of benefit-risk are addressed. The purpose of this review is to draw together the methods currently proposed, to show the inter-relationships between them, and to propose which ones are worth further consideration for decision-making about medicines in which contexts. These methods will then be further explored in the next stage of the PROTECT project.

'Benefit' and 'risk' and other key terms used in this report are used with different shades of meaning in scientific and more colloquial contexts. We have outlined in the glossary working definitions of all key terms in relation to benefit risk assessment as used in this document.

1.3 Objectives

The objectives of the benefit-risk methodology appraisal are:

1) to clarify which methods are based on the same principles/characteristics and to what extent the various methodologies are similar.
2) to group methodologies appropriately in order to provide a clear structure for future reference. The grouping strategy will be developed as part of the review.
3) to evaluate the capacity of each method, and to provide readers with suitable references to introductory-level papers, technical papers, and worked example.
4) to decide which methods to take forward to the next stage of the PROTECT project.

1.4 Communication and dissemination

The essence of this project overall is in the communication and dissemination of benefits and risks to relevant stakeholders. Whilst this report reviews the methodologies, it also sets the scene for benefit-risk communications in the form of accessibility, assessability, and particular emphasis on visual representations. The second part of the review will address this aspect more fully.

1.5 Structure of the Report

Section 2 gives details of the methods of the review, including the literature searches for methods proposed for benefit-risk. In section 3 these methods are evaluated against specified appraisal criteria and grouped into categories according to purpose, In sections 4 – 7 , the methods in each of these categories are outlined, compared and contrasted. Fuller details, examples and references are contained in the Appendices. For each, recommendations are made as to which should be taken forward into the next stage of the PROTECT project. Section 8 summarises the review, and collates the recommendations.

2 Methods

2.1 Introduction

This section describes the methods and process used in this review. Section 2.2 starts with addressing other initiatives and benefit-risk methodologies reviews available in the literature, and presents our strategy in searching the required materials. Preliminary assessment of the reviews found through literature search have led to excluding some reviews based on the scope of our charter, and the broad inclusion and exclusion for taking reviews forward are clarified in Section 2.3. Since reviews included may contain several benefit-risk assessment approaches, the strategy for including and excluding certain approaches are discussed in Section 2.4. Approaches that are included in this report are discussed through a set of appraisal criteria which are listed in Section A.3. As we aimed to classify these approaches in order to provide more structured understanding and effortless reading of this report, the strategy for the classifications are then discussed in Section 2.6.

2.2 Literature search strategy

Comprehensive reviews on benefit-risk methodologies have already been undertaken by others. To avoid duplication of effort, we primarily based our methodologies review on identification of existing reviews either published in the medical literature or through regulatory agencies or pharmaceutical companies. We aimed to include both qualitative and quantitative methodologies for benefit risk assessment, and did not restrict the review to benefit-risk methodologies within pharmacoepidemiology and clinical trials but also aimed to include benefit-risk assessment methods from other fields (e.g. Health technology assessment.).

We searched for materials from:

- Published individual assessments or reviews on benefit-risk assessment approaches (see Section 2.1.);
- Other work carried out by PROTECT members on benefit-risk assessment that are not available in the public domain;
- Other work from other parallel benefit-risk initiatives that could not be identified through systematic literature search, for example the EMA Benefit-Risk Report [1]; and
- Pharmaceutical companies' internal methods for benefit risk assessment, obtained via
  - Personal contact
  - FDA Advisory Committee meetings (http://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/default.htm)
  - Other meetings and conferences

2.3 Inclusion of reviews identified

The names of benefit-risk approaches reviewed in each review paper were extracted onto an Excel spreadsheet for further evaluation. Authors' review of the approaches were taken into account in our appraisal, but in most cases the key papers of each B-R approach and related references were referred to in order to gain more understanding on the scientific rigour of the approaches and their applications in benefit-risk assessment. These are referenced throughout.

At the suggestion of our External Advisory Board, we expanded our charter (Appendix A.1) to include consideration of Health Technology Assessment (HTA) methods. This means that we have included more approaches that allow for multiple comparisons between drugs, but as a general rule of thumb, reviews of benefit-risk trading approaches specifically developed for economics evaluation were excluded as economics perspective of benefits and risks are outside the scope of our review as defined in our charter.

2.4 Inclusion of methodologies identified from reviews

The review team went through the list of benefit-risk approaches extracted from all identified reviews during a face-to-face meeting; and excluded approaches that were not considered relevant for this appraisal. These were approaches that are specific to economic and business decisions which had not been excluded in the first pass. All other identified approaches are included for appraisal in this report.

In addition to the approaches found through systematically reviewing the literatures, we also added further approaches that WSB team members were aware of and which were considered relevant to our objective. Some identified approaches were not, in our judgment, benefit-risk assessment approaches in their own rights but, as they are frequently cited in this context, they were still included in this appraisal report to clarify their status in this field.

2.5 Development of appraisal criteria

We developed the criteria for the methodology appraisal on the work in the EMA Benefit-Risk project [1] through a series of discussions in technical meetings. Four dimensions were identified as being the most suitable and useful for the purpose of our appraisal. These are the (1) fundamental principles of the benefit-risk approach, (2) features of the approach, (3) whether there are any existing visual representations associated with the approach, and (4) the assessability and accessibility of the approach.

The dimensions were broken down further into further criteria to address specific issue within each dimension, and are given in details in Appendix A.3. A table template with these criteria was created in Microsoft Excel for each approach to assist reviewers in gathering information for the appraisal.

2.6 Classifications strategy

The benefit-risk assessment approaches identified can be classified into four specific groups based on their characteristics (Appendix A.3). Broadly, the benefit-risk approaches in the literatures can be divided into frameworks and metrics (measures). We based our initial classification strategy on the two groups by studying each approach and categorising them as either a framework (defined as *Structured stepwise approach to perform a task*) or a metric (defined as a *System of measurement*).

Frameworks can be classified into (non-quantitative) descriptive frameworks which are generic, or more comprehensive quantitative frameworks that clearly present the structure and methods of assessing benefits and risks for decision making. Metric indices fall into those that provide thresholds of benefits or risks for the purpose of comparison or as a guide in the decision process, and those that essentially weigh the benefits and risks for benefit-risk assessment. A group of specialised metrics, the health utility indices, is described separately because of they are validated metrics and are widely used in specific areas of medicine.

Some approaches did not fit under the two main categories because they do not provide a metric for benefit-risk and also do not come within a defined framework in their applications. More often than not, these approaches are not standalone and are used in combination with other metrics or within a framework; therefore separate categories are defined to best discriminate them. One group of approaches has been identified as being supporting techniques for making inferences on benefit-risk in decision making. These techniques bring together the evidence in a model that estimates the benefit-risk trade-offs using under certain conditions. This takes place at the modelling stage when evidence, preference data, assumptions and other required parameters are available or, otherwise can be reasonably assumed. Importantly, whilst evidence data are available from sources such as clinical trials, epidemiological studies or even experts' opinions, good preference data are difficult to obtain. Preference data such as utilities are crucial to many underlying decision models. Approaches that deal with elicitation of utilities have also emerged through the systematic review. These are brought into the arena of benefit-risk assessment as the underpinning approaches which are essential to complement the frameworks and quantitative techniques used for estimation.

3 Comparisons and classifications of benefit-risk approaches

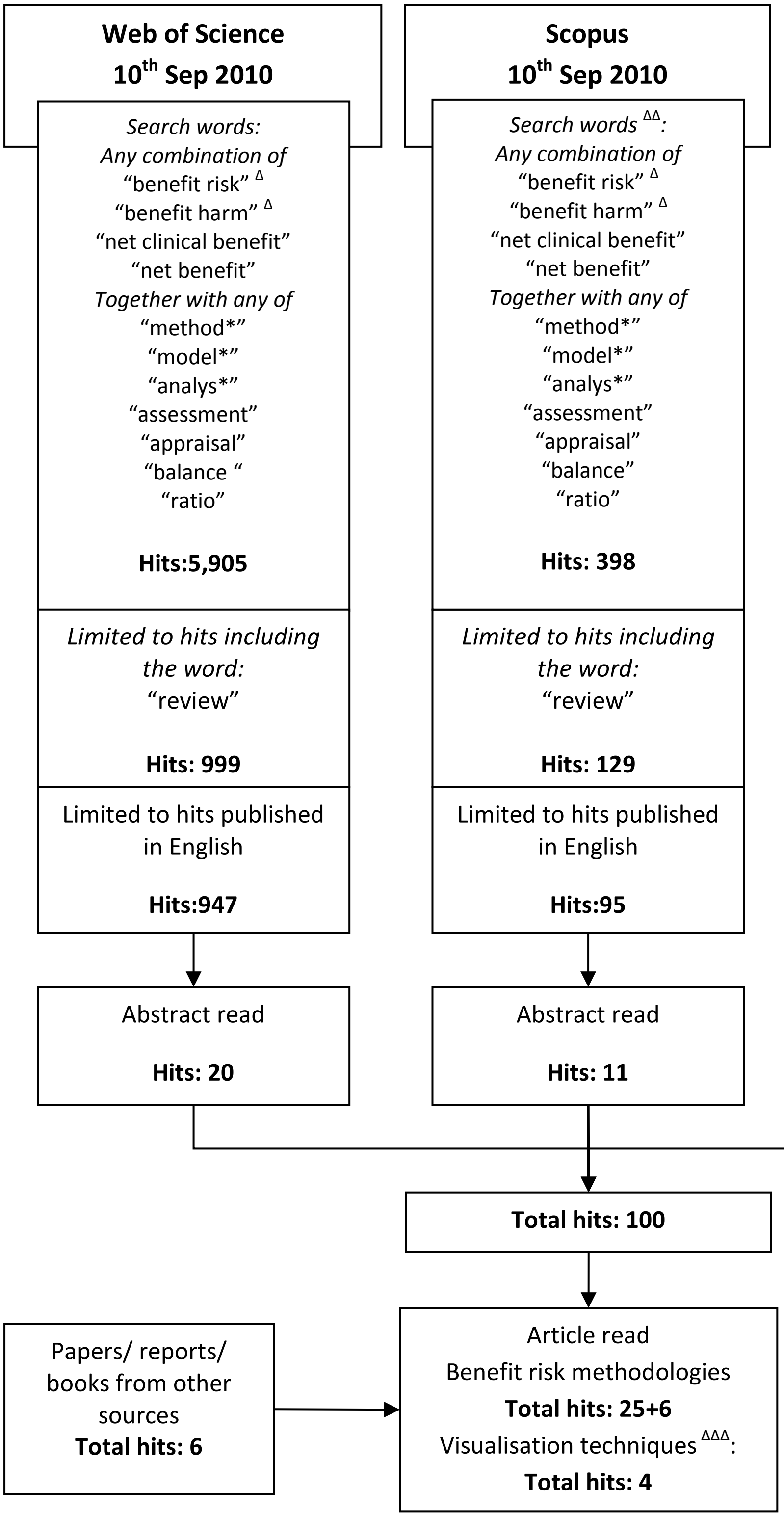
3.1 Introduction

The plethora of techniques that are used in the context of benefit risk decision-making can appear over-whelming. Initially, the results of the literature search that we used as the basis for this review are presented in Section 3.2, then we characterised the methods to make a preliminary assessment, and then subsequently grouped them by purpose, which helps see what aspects of benefit-risk decision-making each technique is intended for (Section 3.3). During the course of writing the review it has been through refinements both of the overall categorisations and of which techniques were listed under each. To aid presentation and make the report easier to navigate, we reverse these steps, presenting this classification in its final form followed by the more detailed characterisations presented according to our classification. We make some remarks about the classifications in Section 3.4.

3.2 Results of the literature search

Following the initial testing and search criteria validation, the full literature search was performed on the 10th of September 2010. The search identified over 45,000 articles using the first level search criteria. Restricting the search to only include reviews published in English, nearly 7,000 articles met the criteria but may still contain duplicates. The abstracts were extracted and read by one reviewer (Christine E. Hallgreen) to arrive at the final list of 100 articles to be extracted in full for the review. Of the 100 articles, 25 met our objectives and are included in this review. A further six sources of reviews from journals, books and internal reports that were not found through the literature search strategy were made aware to the team and were also included. The results of the literature search are summarised in Figure 1.

Figure 1 Flow chart of literature search



Δ Search made on "benefit risk" OR "risk benefit" OR "benefit and risk" OR "risk and benefit", which will find terms including hyphen (- /) between benefit risk, this also apply for benefit harm.

ΔΔ Scopus search key words (box one) was searched for within author keywords, for Web of Science and PubMed the keywords (box one) was searched for within title/abstract/keywords and title/abstract, respectively.

3.3 The classification framework

The approached can be grouped into four major types – those that are embedded in **frameworks** for benefit risk decision making, those that are **metrics** for benefit risk-decision-making, **estimation techniques** for handling the evidence on benefits and risks, and **utility survey techniques** for eliciting values and preferences. Inevitably these distinctions sometimes become blurred, and these will be discussed in more detail later in the report. We briefly describe them here.

3.3.1 Frameworks

Frameworks for decision-making provide guidance for the whole process. We have sub-divided these into those which are **descriptive frameworks**, sometimes referred to as qualitative, or non-quantitative approaches, and those that provide **quantitative frameworks**. The latter, being more comprehensive, often contain elements of the other major categories.

3.3.2 Metrics

Metrics are measurements of risk benefit, and cover a spectrum, which have been subdivided into those that we have termed **threshold indices**, those that have been explicitly developed in a various health contexts, termed here **health indices,** and those that explicitly allow **trade-offs** of benefits and risks in general.

3.3.3 Estimation techniques

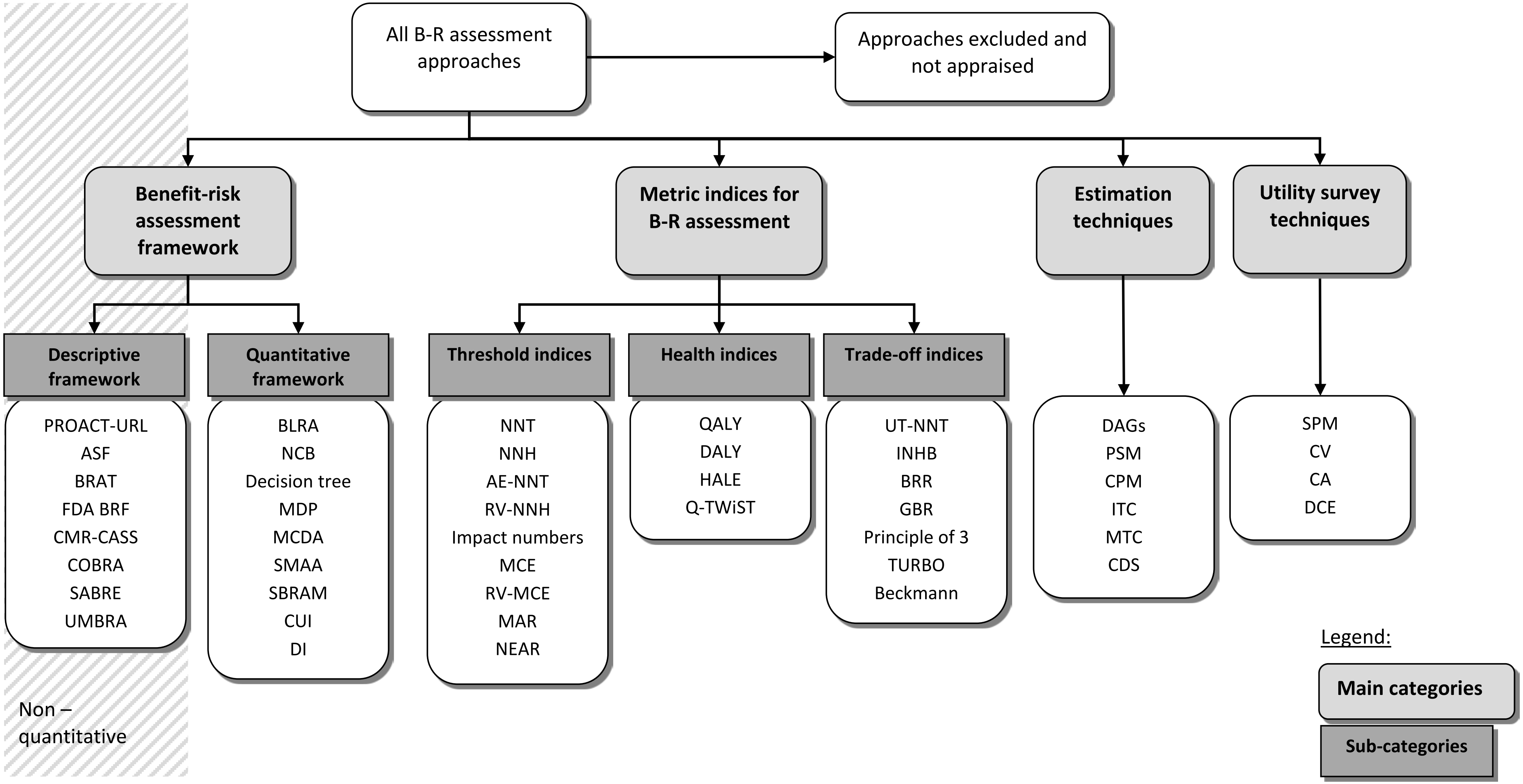
The **estimation techniques** range from the very simple, through to cutting-edge statistical methods for synthesis of complex data from multiple sources. These are typically generic to a much wider class of statistical tasks, but we review them here from the perspective of underpinning formal quantitative benefit-risk decision-making

3.3.4 Utility survey techniques

The **utility survey techniques** complement those approaches which need explicit elucidation of values and preferences of various outcomes from the relevant stakeholders.

3.3.5 Snapshot of classifications

Figure 2 Classifications of benefit-risk assessment approaches



3.4 Remarks

The classifications were performed in such way that the purpose of the different approaches can be clearly seen. There may be some overlaps between categories and the approaches that come underneath them, so we eventually classified the approaches to the more appropriate categories.

4 Benefit-risk assessment frameworks

4.1 Introduction

Many attempts have been made at developing *benefit-risk assessment frameworks*: some are meant to be general and some are more specific. *Benefit-risk assessment frameworks* aim to provide a structured way of dealing with decision problems from how to approach and define a problem, to how to quantify and communicate benefit-risk trade-offs.

This section introduces the frameworks as classified in Section 3.3 by firstly presenting the operational characteristics of the frameworks (Section 4.2). Then we describe and briefly appraise the descriptive frameworks in Section 4.3 and the quantitative frameworks in Section 4.4. Finally, a short conclusion and recommendations of candidate frameworks to be taken forward are made in Section 4.5.

The detailed appraisal of each approach in this section can be found in Appendices A.5 and A.6.

4.2 Operational characteristics of frameworks

This section summarises the operational characteristics of the frameworks by their sub-categories. These are broadly the parameters that each approach can accommodate, some main features, and the type of metrics obtained from the analysis, presented in Table 1. Descriptive frameworks do not formally satisfy any of the characteristics simply because they are general and non-quantitative but are included for completeness. The frameworks that produce weighted utilities are more complete approaches because they apply decision theory. These are followed by utility threshold because they incorporate value judgments based on actions and consequences but do not explicitly trade off benefits and risks; and weighted score as these are subjective judgments similar to utilities but are not based on specific action-consequence interactions.

Table 1 Operational characteristics of benefit risk assessment frameworks

	Descriptive frameworks								Quantitative frameworks							
	PrOACT-URL	ASF	BRAT	FDA BRF	CMR CASS	COBRA	SABRE	UMBRA	BLRA	NCB	Decision tree	MDP	MCDA	SMAA	CUI / DI	SBRAM
π	-	-	-	-	-	-	-	-	X	X	X	X	X	X	X	X
U	-	-	-	-	-	-	-	-	X	X	X	X	X	X	X	
S	-	-	-	-	-	-	-	-	X	X			X	X		X
w	-	-	-	-	-	-	-	-	X		O	X	X	X	X	X
I	-	-	-	-	-	-	-	-	X		X	X	X	X	X	
T	-	-	-	-	-	-	-	-	O		O	O	O	O	O	
ζ	-	-	-	-	-	-	-	-	X				X	X	X	
G	-	-	-	-	-	-	-	-			X	O	X	X	X	X
M	-	-	-	-	-	-	-	-	U_w	U_ρ	U_w	U_w	U_w	U_w	U_w	ψ_w

π = require probability, S = scoring involved, U = require utility, w = require weights, I = Integrate risk and benefit, T = integrate time trade-off, ζ = explicit sensitivity analysis required, G = graphical methods proposed, M = the resultant quantitative benefit-risk metric. X indicates relevant parameters; O indicates optional parameters. $U_\rho = \text{utility threshold}$, $U_w = \text{weighted utility}$, $\psi_w = \text{weighted score}$

4.3 Appraisal of descriptive frameworks

The descriptive frameworks are generic stepwise instructions to structure the thinking when facing decision problems. However, different frameworks may address different things and may be specific to certain type of problems. One of the earlier descriptive frameworks for structuring decision problems is the PrOACT-URL (Appendix A.5.1) [2]. PrOACT-URL simply refers to the eight steps of the framework which are Problems, Objectives, Alternatives, Consequences, Trade-offs, Uncertainty, Risk attitudes, and Linked decisions. **PrOACT-URL provides a framework addressing the necessary elements in dealing with decision problems** through these eight steps, and that other frameworks have been developed on the same principles [1]. This is in some way true because the first five elements – PrOACT – are very general and pertinent to any problem. The last three elements – URL – are more specific and address the state of uncertainties of the decisions.

Another simpler descriptive framework aimed at identifying the **sources of evidence and immediate parties involved in the health context** emerged a year later, which we refer as the Ashby and Smith framework (Appendix A.5.2) [3]. ASF does not explicitly address the trade-off between benefits and risks, but the core paper gave a worked example through the application of decision tree and mathematical manipulations of utilities and evidence data [3;4]. **The application of aspects included in ASF in addition to PrOACT–URL would increase transparency and strengthen the evidence side in a decision problem.**

The idea underpinning both PrOACT-URL and ASF can be seen through the later proposed Pharmaceutical Research and Manufacturers of America (PhRMA) Benefit Risk Action Team (BRAT) framework [5;6] (Appendix A.5.3). **BRAT provides guidelines on organising, understanding and summarising evidence of benefits and risks**. In dealing with decision making, **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**. This was consciously proposed to avoid synthesis of data into sophisticated statistical models which may not be easily understood. The transparency in **BRAT framework finds intuitive acceptance by those not familiar with complex statistical models**, thus making it an easy-to-implement tool to structure simple decision problems on daily basis. However, its recommendation to use odds ratios to summarise benefits and risks is somewhat controversial in comparative benefit-risk assessment.

We are aware of four other descriptive frameworks under development (Appendix A.5.4). 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 [7]. Another framework 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 [8]. The CMR CASS group tests the application of a similar framework [9] by omitting the assessment of numerical scores and weights [1]. 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 [10]. 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 “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 benefit-risk assessment process globally by establishing a consensus on a scientifically acceptable framework for decision-making [10].

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. In Table 2, the stakeholders’ perspectives are listed for each descriptive framework to illustrate the fitness for purpose of the relevant stakeholders when they are applied.

Table 2 Stakeholders’ perspective for descriptive frameworks

	Perspective for stakeholders
PrOACT-URL	pharmaceutical companies, healthcare providers, regulatory agencies
ASF	pharmaceutical companies, healthcare providers
BRAT	pharmaceutical companies, regulatory agencies
FDA BRF	regulatory agencies
CMR-CASS/COBRA	regulatory agencies
SABRE	regulatory agencies
UMBRA	pharmaceutical companies, healthcare providers, regulatory agencies

4.4 Appraisal of quantitative frameworks

Comprehensive quantitative frameworks are extensions of the descriptive frameworks where quantitative methods of trading off benefits and risks are made explicit. Sensitivity analysis and incorporation of value judgments play vital role in these frameworks. Although they may not be extremely different from each other, their capacity, focus and level of complexity in application may differ.

The benefit-less-risk analysis (BLRA) framework is one of the earlier frameworks for dealing with weighing benefits and risks in medicine based on a defined benefit endpoint which is usually known from pharmacology of drugs, and the more uncertain drug-induced adverse events [11]. It is a simple multi-criteria analysis framework which places quantitative measures of benefits and risks on the same scale for the purpose of trading off, in this case by taking the difference, using a proportionality constant derived from benefits and risks in the comparison group (Appendix A.6.1). **The unique feature of BLRA is that it explicitly considers the organisations of adverse events into body functions**. Despite its attractive proposals and simplicity, BLRA application in benefit-risk assessment has never really taken off.

In a similar way to how BLRA trades off benefits and risks, a Bayesian net clinical benefit (NCB) has been proposed as an evidence synthesis and benefit-risk assessment approach [12]. The application of NCB strengthens the use of evidence and value judgments (obtained through quality of life measures). However, the functional form for quantifying benefit-risk trade-off is somewhat vague with a general rule of taking the difference between the two measures (Appendix A.6.1).

The formal application of decision trees in medical decision making is fairly recent [13] but it has been applied in many other areas outside medicine much earlier. Under the decision tree framework, decision problems are 'branched' into options and consequences. The branches are characterised by utilities of stakeholders' value judgments, and the probabilities of the consequences occurring given a chosen route is taken (Appendix A.6.3). Decision tree helps to structure the problems clearly in a tree-like visual aid, and this idea has been incorporated in many other decision frameworks, either implicitly or explicitly.

Markov decision process (MDP) is a framework based on Markov system dynamics and decision tree. In MDP, the probability of being in the subsequent state only depends on the current state as defined by a set of transition probabilities and is dependent on taking certain actions at the current state for a given consequence [14]. Its use in decision making in medicine is somewhat limited due to the assumption that past history does not matter in future decisions (Appendix A.6.4).

A similar concept to decision tree, the value tree, has been seamlessly integrated into the principles of multi-criteria decision analysis (MCDA). This becomes obvious when almost all MCDA software contain the facility to set up problems as value trees. MCDA, to be precise, is an umbrella term for approaches based on decision theory that deal with multi-criteria decision problems but the one considered here is that described by Keeney and Raiffa [15]. Based on multi-attribute utility theory, MCDA enables each criterion to be judged preference-independently of the others [15] making assigning value judgments more straightforward. Although MCDA have been widely used in many areas outside medicine (particularly business decision making), MCDA application for medical decision making is still limited but has attracted massive interests since its introduction in the medical literature [16]. The strengths of MCDA lie in having **structured stepwise instructions in line with PrOACT-URL framework and the capability of assessing and integrating multiple benefits and risks criteria, as well as comparing different options**. Furthermore, MCDA is also the only approach that can **formally deal with multiple objectives at the same time**.

The standard MCDA approach for drug benefit-risk decision making [16] lacks the **ability to account for sampling variations** in the criteria measurements; and its validity can be threatened **in the case when preference information**

is missing or when consensus is not reached. The stochastic multi-criteria acceptability analysis (SMAA) is introduced in the drug benefit-risk decision-making as a way to overcome these limitations by carefully modelling them through simulations, and should be augmented by appropriate approach for evidence synthesis [17]. Other families of SMAA are available to deal with specific scenarios (Appendix A.6.6).

Another MCDA framework has been developed for assessing benefit-risk trade-off in drug development which we simply refer to the Sarac's benefit risk assessment method (SBRAM) [18]. It is similar to MCDA with added quantification of uncertainty through bootstrapping and concurs with the BRAT framework to present benefit-risk results un-integrated (Appendix A.6.7). However, the framework lacks the comprehensiveness of MCDA and does not add much more value than an MCDA model or the recent SMAA approach. Its specific tailoring to drug development also places it out of the scope of our charter.

Clinical utility index (CUI) and desirability index (DI) provide general framework in assessing benefit-risk trade-off of drugs in development when they are measured over a range of doses or time [19;20]. They are used to characterised the therapeutic index of drugs [21], which could then be used as the common metric for benefit-risk comparison. Their application is not model specific, therefore the resultant metric CUI or DI can be directly compared across different assessments.

Table 3 summarises our judgments about the different quantitative frameworks. Decision trees, CUI and DI rely on data, utilities and weightings but do not consider discriminative scoring. Other frameworks introduce scoring systems in the frameworks to differentiate the value judgments on the criteria further. SBRAM however has low level discriminative scoring system with only three levels. The other quantitative frameworks allow scoring to be made on continuous scale and therefore are highly discriminative. The complexity ranges from medium for BLRA and complex for the rests. Basic statistical knowledge such as ANOVA and t-tests are generally required in the application of BLRA although other aspects of the framework do require more extensive clinical expertise and judgments. The other frameworks were judged to be complex because they require greater understanding of utilities, complex statistical inference methods, and simulation methods. With the exception of SBRAM, all quantitative frameworks provide the means to compare more than two treatment options simultaneously, but only MCDA and SMAA can simultaneously deal with multiple objectives. NCB and SBRAM use summary data from the population, whilst BLRA, CUI and DI use individual level data in the analysis. Other frameworks are flexible as to which type of evidence data is used. Their fit for purpose for the application in benefit-risk assessment by different stakeholders is also listed.

Table 3 Quantitative frameworks comparison

	Discriminative scoring	Level of complexity	Number of options	Evidence data	Perspective for stakeholders
NCB	High	Complex	> 2	Population	pharmaceutical companies, healthcare providers, regulatory agencies
BLRA	High	Medium	≤ 2	Individual	same as above
Decision tree	N/A	Complex	> 2	Population or individual	same as above
MDP	N/A	Complex	> 2	Population or individual	same as above
MCDA	High	Complex	> 2	Population or individual	same as above
SMAA families	High	Complex	> 2	Population or individual	same as above
SBRAM	Low	Complex	≤ 2	Population	pharmaceutical companies
CUI / DI	N/A	Complex	> 2	Individual	Pharmaceutical companies and regulatory agencies

4.5 Conclusion and recommendations

In general, *descriptive frameworks* are only as good as good practice goes in terms of making decisions problems more transparent allowing for quantitative benefit-risk assessments to take place more naturally. They do not in their own rights perform any quantitative trade-offs of benefits and risks, therefore are useful in benefit-risk assessments only when used in combination with other quantitative approaches.

Approaches classified into *quantitative frameworks* provide the most useful methods to perform benefit-risk assessments. This is simply because they provide stepwise tasks to be followed usually from data gathering to making comparison between alternatives to making the decisions. It is important to point out that in many cases, these steps may still be too general and heavily dependent on the decision makers/modellers, but transparency is enforced.

Table 4 and Table 5 show the comparative overview and justifications of the recommendations of frameworks for further consideration and testing. The review team decides to recommend the following *benefit-risk frameworks* to be taken forward as candidate approaches in the case studies:

(1) PrOACT-URL frameworks. PrOACT-URL provides a framework addressing the necessary elements in dealing with decision problems. PrOACT-URL however does not clearly address the importance of identifying appropriate sources of evidence and immediate parties involved. The application of PrOACT–URL therefore requires improvement in these two aspects, for example by combining with elements of the Ashby and Smith framework (ASF) or other evidence synthesis approach. The extension of PrOACT-URL to drug benefit-risk decision analysis by the EMA

Benefit-Risk Methodology working group 2 may, in the future, provide better evolution of the frameworks in this specific domain.

(2) Benefit Risk Action Team (BRAT) framework. BRAT provides guidelines on organising, understanding and summarising evidence of benefits and risks into tabular outputs and graphical summaries. The framework proposes avoiding integration of benefits and risks evidence to make it more accessible and transparent to those not familiar with complex statistical models. The controversial aspect of BRAT is its proposal to use odds ratios as the basis for the decision on benefits and risks balance.

(3) Multi-Criteria Decision Analysis (MCDA). MCDA provides structured stepwise instructions in line with PrOACT-URL framework with the capability of assessing and integrating multiple benefits and risks criteria, as well as comparing different options. MCDA is also the only approach that can formally deal with multiple objectives simultaneously. Another appealing feature of MCDA is that specialist several software to perform the analysis are available.

(4) Stochastic Multi-criteria Acceptability Analysis (SMAA). The SMAA families of approaches are potentially serious contenders to the standard MCDA because of the ability to account for sampling variations arising from the type of experimental designs used as well as when there is missing information on preference values. However, the increased complexity of SMAA compared to MCDA may be the major barrier in real-life benefit-risk medical decision making applications. The application of SMAA in medical decision making should be explored further as there is a potential that the same SMAA model could be used to address different stakeholders.

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 judgments as well as the importance of sensitivity analysis •Proposes 'effects table' as snapshot of evidence	•Missing the importance of identifying appropriate sources of evidence and immediate parties involved •Has been extended in EMA benefit-risk Methodology working group 2 •Does not address communication	Yes	•Address the necessary elements in dealing with decision problems •Forms basis for other frameworks	•To structure decision problems using 8-step process •To allow transparency
BRAT	•Value tree model build-up •Does not integrate benefit and risk •Optional weights assignment to benefit and risk criteria •Summarise criteria as tables and forest/dot plots	•Summarises evidence and communicates them but does not assess B-R •Can be exhaustive	Yes	•Developed by PhRMA •Accessible to those not familiar with complex statistical models •Offers graphical presentation of results in the form of a forest plot	•To structure decision problems using 6-step process •To allow transparency •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 •Consider post-approval phase	•Aimed at small regulatory agencies	No	•Superseded by COBRA (see below)	•To support B-R decision-making in small regulatory agencies
FDA BRF	•Template to facilitate BR decision and communication within and outside of the US FDA	•Aiming for simplicity, transparency, and non-quantitative final arguments	No	•Still under development	•To provide decision-makers with the "big picture" of decision problems
COBRA	•Provides automated standard template for B-R evaluation •Semi-quantitative approach	•Aimed at small regulatory agencies •Successor of CMR-CASS	No	•Still under development	•To support B-R decision-making in small regulatory agencies
SABRE	•Unknown	•Details are not currently available	No	•Still under development	•To support B-R decision-making in the Southeast Asian countries
UMBRA	•Standardising elements across different B-R methodologies	•Development is driven mainly from the PhRMA BRAT framework	No	•Still under development	•To structure decision problems using 4-stage, 8-step process •To allow transparency •To aid B-R communication

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?		
				Reasons	Specific use
BLRA	• Organises adverse events into body functions for analysis • Sensitivity analysis involves varying proportionality constant	• Burdensome for limited evaluations and simple problems • More medical knowledge is required compared to others	No	• Very similar to MCDA	• To structure and analyse decision problems using 7-step process • To allow transparency • To integrate benefits and risks
NCB	• Can be expressed as NNT • Naturally allows evidence update as Bayesian models • Line graphs and distribution plots as visuals	• Bayesian model associated with long computing time • Functional form for quantifying benefit risk trade-off is not specific	No	• Could be difficult to apply since many steps require extensive statistical modelling expertise • Similar to MCDA	• To structure and analyse decision problems using 3-step process • To allow transparency • To integrate benefits and risks
Decision tree	• Represents the expected utility rule visually • 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 • Can be highly complex with large "trees"	• To structure and analyse decision problems using 5-step process • To allow transparency • To integrate benefits and risks • To investigate whether certain data is worth obtaining
MDP	• Similar to decision tree • Markov chain combined with decision tree	• Not all medical decision problems can be described by MDP's dynamic nature • The structure can be very complex with many criteria • 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 • Transition probabilities may be difficult to establish • Explicit steps are unclear • Similar to decision tree	• To structure and analyse decision problems • To allow transparency • To integrate benefits and risks
MCDA	• Explicit value judgments • High discriminative scoring system • Multiple sources of evidence can be defined as criteria	• Burdensome for limited evaluations and simple problems • Software: Hiview, V.I.S.A. Intelligent decision system, Logical Decisions, etc.	Yes	• Highly structured • Can deal with multiple objectives simultaneously • Combining multiple criteria is easy	• To structure and analyse decision problems using 8-step process of PrOACT-URL • To allow transparency • To integrate benefits and risks

Approach	Features	Comments	Has it been recommended that this be taken forward to the next stage?		
				Reasons	Specific use
	•Multiple objectives can be addressed simultaneously •Allows any data types •Value tree diagram, line graphs, bar graphs (including difference diagram), and area graphs (frontier plot) as visuals	•Hiview provides some useful visual representations •Does not account for uncertainties in data		•Several software to implement are available eliminating the need for mathematical knowledge of decision theory	
SMAA	•Features similar to MCDA •Uncertainties in evidence data are taken into account •Can deal with missing or partially-missing value preferences •Bar and line graphs as visuals	•Extends MCDA •Software: JSMAA •Requires extensive mathematical and computational knowledge •There are several specialist metrics/concepts in SMAA •There are variations of SMAA	Yes	•Complement MCDA •Provide better estimates than MCDA when evidence are unknown, uncertain, or when their distributions are skewed	•To structure and analyse decision problems •To allow transparency •To integrate benefits and risks •To account for uncertainties •To better reflect situations encountered in real-life decision problems
SBRAM	•Does not integrate benefit and risk •Incorporate value judgments implicit in the scoring in an objective manner •Clinical data evaluated and scored based on descriptive statistic •Low discriminative scoring system •Tornado-like diagram as visuals	•Only compares two options at a time •To evaluate multiple options additional scoring of data is required and analysis is compared visually •No software available for data analysis (scoring) •Underlying statistical analysis might be difficult to understand for layman •Developed for drug development	No	•Similar to but lacks comprehensiveness of MCDA •More specific to drug development •Requires greater knowledge of statistical inferences	•To structure and analyse decision problems using 8-step process •To allow transparency •To eliminate ambiguous absolute judgment in the final results
CUI/DI	•Not model specific, therefore the final metrics can be compared across different assessments •Data must be from controlled clinical trials with same indication •Does not require alternatives •Line graphs, surface and contour plots as visuals	•Transparency hampered from having a rather too general framework •More useful when utility index is expressed as a function of dose	No	•More specific to drug development •Requires data that are difficult to obtain •Limited applicability	•To structure and analyse decision problems using 4-step process •To allow transparency •To integrate benefits and risks •To establish the benefit-risk balance without a comparator

5 Metric indices for benefit-risk assessment

5.1 Introduction

Quantitative benefit-risk assessments require numerical representations of benefits and risks. This is achieved through the computations of various *metric indices* when performing benefit-risk assessments. The *metric indices* can be classified into three sub-categories: those that provide indices that are used as thresholds, those that characterise health outcomes an 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, attributable risks etc. but these are not reviewed here but may also be suitable to quantify benefits and risks for the purpose of decision making in medicine. Their descriptions can be found in many statistics textbooks [22] and also have been described as "quantitative framework for benefit-risk assessment" [23], and their concepts have been incorporated into many other metric indices described in this section as well as within other approaches in this report.

Section 5.2 compares the operational characteristics of the *metric indices* within and across the sub-categories, and Sections 5.3 to 5.5 provide overview comparisons of these indices. Section 5.6 concludes the essence of *metric indices* and proceeds to recommend several indices to be taken forward to the next stage of this project.

The detailed appraisals of *metric indices* are available in Appendices A.7, A.8 and A.9 at the end of this report.

5.2 Operational characteristics of metric indices

The operational characteristics of the metric indices are presented in Table 6 by their sub-categories. These are broadly the parameters that each approach can accommodate, some main features, and the type of metrics they are. Metrics that are utility-based have the capabilities to address preference from stakeholders, and those based on rates or probabilities rely only on empirical data. The metrics that are scores incorporate preference values through some ad-hoc scoring systems. Only the trade-off indices integrate benefits and risks in the sense that the two separate metrics are combined into a single metric for use in the final assessment. Some threshold indices and health utility indices may incorporate both benefits and risks in their derivation but do not explicitly trade off benefits and risks in the same way.

Table 6 Operational characteristics of metric indices

	Threshold indices							Health utility indices		Trade-off indices						
	NNT / NNH	AE-NNT / NEAR	RV-NNH	Impact numbers	MCE	RV-MCE	MAR	QALY / DALY / HALE	Q-TWiST	UT-NNT	INHB	BRR	GBR	Principle of three	TURBO	BM
π	X	X	X	X	X	X	X			X		X	X			
U			X			X	X	X	X	X	X					
S							X	X	X		X			X	X	
w								X	X		X					X
I		X			X	X	X	O	X	X	X	X	X	X	X	X
T								X	X		X					
ζ											X					
G							X								X	
M	Δ_ρ	Δ_ρ	U_ρ	Δ_ρ	Δ_ρ	U_ρ	U_ρ	U_w	U_w	U_ρ	U_w	Δ_ρ	Δ_w	ψ	ψ	ψ_w

π = require probability, S = scoring involved, U = require utility, w = require weights, I = Integrate risk and benefit, T = integrate time trade-off, ζ = explicit sensitivity analysis required, G = graphical methods proposed, M = the type of quantitative metric. X indicates relevant parameters; O indicates optional parameters. U_ρ = utility threshold, U_w = weighted utility, ψ_w = weighted score, ψ = score, Δ_ρ = rates threshold, Δ_w = weighted rates.

5.3 Appraisal of threshold metric indices

A group of general indices that are used in benefit-risk assessments to quantify benefits and risks are classified as the threshold indices. These indices are derived from mathematical and statistical manipulations of probabilities and/or utilities, and are generally used as thresholds (cut-points) when assessing benefit-risk balance or in deciding the best treatment options. These indices are either not designed to or are not formally used to trade-off benefits and risks.

A popular concept derived from the reciprocal of the difference between two probabilities of two different treatment success rates is the number needed to treat (NNT) (Appendix A.7.1). NNT is a metric quantifying the number of patients with a medical condition who need to be treated in order to ensure that one patient will be successfully treated [24]. NNT is a waiting time paradigm that at the population level, how long (in terms of number of patients) a healthcare provider needs to wait before seeing that the treatment delivered actually works, or at the individual level, how long (in terms of time) does it take to see a successful effect of a treatment. A parallel metric is the number needed to harm (NNH) or otherwise known as the threshold NNT (T-NNT) which accounts for treatment failure or adverse outcomes [25]. Benefits and risks thresholds described by NNT and NNH respectively are compared on their face values where NNT<NNH is favourable. Although confidence intervals can be easily calculated, only the point estimates are often used when making comparison, throwing away useful information on

uncertainty. **NNT and NNH are simple to calculate and to interpret** but their interpretation is threatened when the difference in probabilities is very small.

Many attempts have been made to improve NNT for use in drug benefit-risk assessment. One is the adverse event adjusted NNT (AE-NNT) where the original NNT is penalised when the successfully treated patients experience treatment-induced adverse events [26]. AE-NNT is derived from the probabilities in the relevant group of patients dubbed those who had “unqualified success” with the treatment. Therefore, AE-NNT can only be used correctly when individual level data on treatment in a group of patients are available (Appendix A.7.1).

With the classical NNT and AE-NNT, only one benefit or one risk can be accounted for, but another extension provides a method to combine more than one benefit and risk criteria as well as incorporating utility through relative value adjustment. The relative value adjusted number needed to harm (RV-NNH) is the proposed metric [27] since it is common that there are several adverse events to be considered for any treatment. Utilities are incorporated as the ratio of risk utility to benefit utility. RV-NNH is used as the threshold when comparing to the classical NNT, where NNT<RV-NNH is favourable.

A more recent addition to the NNT family is a group of threshold metrics known as the impact numbers (Appendix A.7.2). Impact numbers consist of a series of metric indices describing the number to be treated in a particular situation and population [28;29]. These are originated from the amalgamation of the NNT idea with the more established epidemiological measures such as relative risks, attributable risks, and attributable fractions. Impact number may have the **potential to make the interpretation of certain epidemiological concepts more accessible to the general audience.** Two particular impact numbers are most useful in comparison to the others: the population impact number of eliminating a risk factor over time *t* (PIN-ER-*t*) and the number of events prevented in the population (NEPP).

The epidemiological concepts of risk difference, as in NNT, is extended in the application of the minimum clinical efficacy (MCE) where the difference in benefits of two comparative treatments is penalised by the difference in their risks having considered the benefits and risks in the untreated population (Appendix A.7.4). MCE determines the minimal therapeutic benefit for a treatment to be worth considering [30]. A method to incorporate utilities through adjustment with relative value (RV-MCE) is also available and is similar to RV-NNH described above.

A similar metric to MCE, the maximum acceptable risk (MAR), has also been proposed (Appendix A.7.5). MAR compares two options based on several benefits criteria for a chosen risk [31]. It can be seen as a metric on the opposite end of MCE.

AE-NNT idea and the classical epidemiological measures relative risks and odds ratios are revisited and generalised to be used with expected probabilities when individual level data are not available. The net efficacy adjusted for risk (NEAR) approach circumvents the weakness of NNT when the risk difference is very small by considering ratios [28]. The metric indices are interpreted the same way as relative risks and odds ratios but care must be taken that interpretation is only made within the relevant context and medical problem.

Table 7 shows the brief comparison of threshold metric indices. Most of the threshold indices do not formally incorporate any form of scoring; therefore their discriminative ability is limited. Only MAR incorporates elements of scoring because it is closely linked with utility survey techniques (Appendix A.11). The complexities in the applications of these threshold indices in benefit-risk assessment are generally at simple level. The adjustment to incorporate utilities raises the level of complexity to medium due to the more difficult issues in obtaining and understanding utilities. NEAR has medium level complexity because it requires more statistical understanding of observed and expected frequencies as well as understanding of epidemiological concepts. None of the threshold indices can cope with comparing more than two options explicitly. Evidence data required to apply AE-NNT, RV-NNH, RV-MCE, and MAR are the individual level data (or marginal summary data) because specific individual benefit-risk trade-offs are addressed using these metrics. NNT, NNH, impact numbers, MCE, and NEAR requires only population level data because they describe benefit-risk trade-off as a whole rather than addressing individual-specific's circumstances. Obviously, the metrics requiring only population level data can also use individual level data but by doing so can jeopardise valuable information because data are collapsed. When individual level data are available, we strongly suggest that appropriate approaches that can deal with individual data are used so that valuable evidence is not wasted. Their usefulness to which stakeholders is also listed.

Table 7 Quantitative threshold metric indices comparison

	Discriminative scoring	**Level of complexity**	**Number of options**	**Evidence data**	**Perspective for stakeholders**
NNT and NNH	N/A	Simple	≤ 2	Population	Patients, healthcare providers, pharmaceutical companies
AE-NNT	N/A	Simple	≤ 2	Individual	same as above
RV-NNH	N/A	Medium	≤ 2	Individual	same as above
Impact numbers	N/A	Simple	≤ 2	Population	healthcare providers, pharmaceutical companies
MCE	N/A	Simple	≤ 2	Population	same as above
RV-MCE	N/A	Medium	≤ 2	Individual	patients, healthcare providers, pharmaceutical companies
MAR	High	Medium	≤ 2	Individual	same as above
NEAR	N/A	Medium	≤ 2	Population	same as above

5.4 Appraisal of health indices

A group of specialised indices that implicitly provides benefit-risk trade-off within their build are the health indices, which have been developed in specific areas of medicine. Their depiction of benefit-risk trade-off is not very transparent and is not very intuitively interpretable. **Their use in benefit-risk assessment however is widespread and no longer specific because of the generalisability of their quality of life concept**.

The most commonly used of these indices is the quality adjusted life years (QALY) [32] where a quantitative measure of the quality of life in certain time periods and health states is described. Relevant stakeholders provide utilities for their preference to spend time in various health states given treatment and medical conditions which result in QALY for a particular health state. The implicit trade-off is obtained when QALYs in different health states are summed up (Appendix A.8.1).

A parallel extension of QALY is the disability adjusted life years (DALY) which is an index quantifying number of years lost from treatment compared to the national life expectancy. The health adjusted life expectancy (HALE) is another extension of QALY to give an index for the population (Appendix A.8.1).

The **application of QALY in cancer epidemiology** is popularised by the quality adjusted time without symptoms and toxicity (Q-TWiST) where **three specific health states for undergoing cancer therapy** [33]. The time when a patient is subjected to toxicity from treatment, the time a patient is free of toxicity and disease, and time of relapse until time of death are considered. The combination of the QALYs from the three states is then the Q-TWiST index for a particular treatment option.

We found all health indices to be highly discriminative because the scoring methods consist of scales with many levels or continuous. Therefore the health indices can finely differentiate health states criteria. They also have the same medium level of complexity as they are very similar to QALY but with minor modifications. The application of QALY requires some knowledge of utility and awareness of the relevant statistical issues. Although health indices usually come with specific guidelines on their derivation algorithm, their real-life application may become complex when there is missing information in their derivation. Q-TWiST is moderately complex – a point lies between medium and complex – simply because the quality of life measurements require more attentive observation of the timeframe involved. Any health index is capable or comparing more than two treatment options if necessary as they are calculated separately and compared head to head simply because they are standardised metrics with the same unit. Evidence data required are at individual level because they are patient-specific but DALY requires a reference data from the population as comparison. Their relevant stakeholders are listed below.

Table 8 Health indices comparison

	Discriminative scoring	Level of complexity	Number of options	Evidence data	Perspective for stakeholders
QALY	High	Medium	> 2	Individual	patients, health care providers, pharmaceutical companies, regulatory agencies
DALY	High	Medium	> 2	Individual and population	same as above
HALE	High	Medium	> 2	Individual	same as above
Q-TWiST	High	Medium-complex	> 2	Individual	same as above

5.5 Appraisal of trade-off metric indices

Another set of metric indices take the idea further by proposing formal methods to trade off benefits and risks. Naturally, there is a fine line between metric indices in this group and those in Section 5.3. However, as opposed to the threshold metrics, quantitative trade-off indices integrate benefits and risks into a single metric index which represent the value of the trade-off.

An attempt at trading off benefits and risks for decision making linked with the NNT (Section 5.3) is the adjustment by utilities and the timing of adverse events (UT-NNT) [34]. The adjustment proposed was simple that patients' utilities and time lost or gained from treatment are multiplied with the probabilities (Appendix A.7.1). UT-NNT then took the NNT idea a bit further by taking the difference between the difference in benefits and difference in risks, before taking the reciprocal. UT-NNT idea of incorporating utilities and time as well as integrating benefit and risks into a single metric for drug benefit-risk assessment was ahead of its time, but its application in real-life decision-making never really took off and seems to have been superseded by other approaches.

A more coherent trade-off index based on common principles as UT-NNT is the incremental net health benefit (INHB) that trades off health outcome indices [35]. INHB simply estimates the net health benefits by subtracting incremental risks from incremental benefits. "Incremental" simply refers to the difference of benefits or risks between two treatment options (Appendix A.9.1). The application of INHB is **intuitive to the concept of benefit-risk trade off and would emphasise the use of health outcome indices such as QALYs in drug benefit-risk assessment.**

A parallel concept to taking differences of risks from benefits is by taking ratio of the two. Ratios provide a more intuitive trade-off metrics as they can be interpreted readily as the multiplicative effect of one relative to the other. One simple metric is the benefit-risk ratio (BRR) that divides benefits by risks, assuming equal weightings of benefits and risks (Appendix A.9.1). In its simplest form, the assumption of equal weightings is too strong and may not be valid for real-life decision problems.

The ideas of BRR and "risks difference" have been extended in a collection of trade-off metrics constructed around patients' outcomes in clinical trials, known as the global benefit risk (GBR) metrics [36]. The functional forms of GBR metrics are specific (Appendix A.9.3), and the collection consists of one risk difference metric and two risk ratio metrics. The general idea of GBR is to scale the individual components onto the same unit to allow quantitative trade-off using constants. The idea had later been refined into the BLRA framework [11] described in Section 4.4 and Appendix A.6.1.

Three approaches give ad-hoc trade-off metrics through set criteria and scoring of benefit and risk components. The first of the three is the principle of three [37] which considers three components of a drug benefit-risk decision problem that are the disease, effectiveness of treatments, and the adverse reactions from treatment. Each component is then scored by three criteria – the seriousness, the incidence, and the duration – on a scale of 0 to 3. The metric indices for benefits and risks are then derived from these scores (Appendix A.9.4).

Another similar metric index is the transparent score of the transparent uniform benefit risk overview (TURBO) approach [38]. TURBO is made up of two components – the benefit and the risk components. Each component is scored in two criteria: the primary criterion is scored from 1 to 5, and the secondary criterion is scored from 0 to 2, resulting in an index ranging from 1 to 7 for benefit and risk respectively. The "TURBO grid" then gives the trade-off value of the two components, however defined (Appendix A.9.5).

The Beckmann model [39] is another similar concept, but the "hierarchy of evidence" data available for each benefit and risk component is considered. The scoring system is not set in stone but suggestions include using the WHO [40] and the CIOMS III [41] classifications for adverse drug reactions. Beckmann model is thus a better concept than the principle of three and TURBO in terms of its flexibility and general applicability (Appendix A.9.6), but neither provides a robust metric for benefit-risk assessment in regulatory settings.

General comparisons of the trade-off metric indices are shown in Table 9. UT-NNT and BRR do not impose any scoring system in their applications. GBR has low discriminative scoring ability through its grouping of benefits and risks outcomes into five categories. Principle of three is also low at discriminating the criteria with only three levels. TURBO has medium level discriminative scoring by having maximum of five levels in one criterion. The Beckmann model's discriminative ability is somewhat undefined but can range from being low to being high depending on the choice of scoring scale being used. INHB inherits the high discriminative capability of health indices. The application of BRR is simple as only basic knowledge of probabilities is needed but may become complex when substantial aspects of benefit-risk assessment are taken into account. UT-NNT, INHB, GBR, and Beckmann model are of medium complexity due to their slightly more demanding technical requirements including the knowledge and understanding of utilities and awareness of suitable scoring systems. UT-NNT, INHB, and BRR are only capable of handling two options simultaneously as they explicitly compare two options directly. GBR, principle of three, TURBO, and Beckmann model can handle comparisons of more than two options at the same time because they provide trade-off indices for individual option which are then compared across different comparative options. Most trade-off indices require population level evidence data with the exception of GBR which requires individual level data. INHB's evidence requirements are inherited from the actual health index being used which could be either population data or individual data. Stakeholders' perspectives are also listed in Table 9 below.

Table 9 Quantitative trade-off metric indices comparison

	Discriminative scoring	**Level of complexity**	**Number of options**	**Evidence data**	**Perspective for stakeholders**
UT-NNT	N/A	Medium	≤ 2	Population	patients, physicians, healthcare providers
INHB	High	Medium	≤ 2	Population or individual	patients, physicians, healthcare providers, pharmaceutical companies, regulatory agencies
BRR	N/A	Simple	≤ 2	Population	patients, physicians, healthcare providers, pharmaceutical companies
GBR	Low	Medium	> 2	Individual	same as above
Principle of three	Low	Simple	> 2	Population	patients, physicians, healthcare providers
TURBO	Medium	Simple	> 2	Population	same as above
BM	Undefined	Medium	> 2	Population	same as above

5.6 Conclusion and recommendations

Metric indices for benefit-risk assessment can be regarded as the nucleus for decision-making. They are primarily what decision-makers will use to judge when making the final decisions. The choice of metric indices in the analysis, and the communication of benefit-risk assessment results to decision makers and stakeholders is therefore crucial. Otherwise, however good a decision model and evidence are, the decision will be harder if the chosen metric indices are not understandable or are unacceptable.

Metric indices cannot be used satisfactorily in drug benefit-risk decision-making on their own due to lack of transparency and variable subjective issues. Their use within a defined framework of choice (see a list of recommendations in Section 4.5), or in combination with appropriate evidence synthesis and estimation techniques can promote more objective application of these metric indices.

Table 10 – Table 12 show the comparative overview and justifications of the recommendations for metric indices. Our recommendations for *metric indices* to be taken forward into the next stage of the project are:

(5) Number Needed to Treat (NNT) and Number Needed to Harm (NNH). The popularity and its widespread use in clinical literature when describing the benefits or risks of a treatment are well established. These indices, heavily criticised, still provide an attractive feature for benefit-risk assessment – simplicity. They should not be used naively, but be supported by thorough evidence synthesis and suitable modelling of the evidence. We are not recommending the utility-adjusted variants of NNT because the reciprocals of expected utilities do not have the same meaning as the reciprocals of probabilities or rates.

(6) Impact numbers. These metric indices give a different perspective from NNT, and are useful in describing public health burden of a disease, and the potential impact of a treatment. Although other more established epidemiological measures are available and have already been used widely, impact numbers have an intuitive interpretation. Their application in benefit-risk assessment may contribute to making the interpretation more accessible to the general audience.

(7) Quality Adjusted Life Years (QALY). The QALY provides time trade-off with life quality in discrete health states which is absent from other more general trade-off indices. Its use is already established in many areas of medicine particularly in chronic diseases where time factor plays a major role in their assessment of benefits and risks. HALE may be another health index to be considered, but as it is just a summary of QALY in a particular group of people, we will not differentiate it further.

(8) Quality adjusted Time Without Symptoms and Toxicity (Q-TWiST). The discrete health states proposed in Q-TWiST are intuitive and very specific to cancer therapy. Therefore its usefulness is limited to cancer domain but nonetheless is a suitable metric to aid cancer patients to decide on the best acceptable treatment. We are only recommending the use of Q-TWiST within the cancer therapy domain.

(9) Incremental Net Health Benefit (INHB). The INHB builds on health indices like QALY. INHB incorporates time and utility which are desirable elements in benefit-risk assessments. The idea of penalising “incremental” benefits by “incremental” risks in INHB, directly follows the simple intuitive concept of benefit-risk trade-offs.

(10) Benefit-Risk Ratio (BRR). Ratios provide intuitive trade-off metrics as they can be interpreted readily as the multiplicative effect of one relative to the other. The recommendation here is only for the application of BRR when accompanied by high quality evidence data and appropriate statistical modelling, and is presented together with their absolute or baseline rates. The weighting of benefits and risks to be traded off should be carefully taken into consideration when using BRR as the metric for benefit-risk assessment because equal weights assumption may not always, and in most cases do not, hold.

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?		
				Reasons	Specific use
NNT and NNH	• Can only include one criteria • The reciprocal of absolute risk reduction • Benefit and risk are described separately as NNT or NNH • Implies equal weight between benefit and risk when direction comparing NNT to NNH	• Undefined with no treatment effect • CI's are problematic when absolute risk reduction includes zero (CI includes infinity) • Values of NNT for different conditions are not comparable • Timeframes must be considered carefully since they are used implicitly in the calculations	Yes	• Widespread use in clinical literature • Simple • Easy to understand	• To describe results in terms of number of people • 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 • May lead to implausible interpretations	• To incorporate utility and time factors into NNT analysis • Also see NNT
AE-NNT	• Extension of NNT based on marginal probabilities • Integrates one benefit with multiple risks	• Can only be used correctly when individual level data of treatment are available	No	• Similar to NNT • Individual level data may be difficult to obtain	• To integrate benefit and risk in NNT analysis • Also see NNT
RV-NNH	• Extension of NNH to include utilities • 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 • Incorporation of utilities changes the definition of the reciprocal	No	• Similar features to NNH • May lead to implausible interpretations	• To account for subjective judgments on criteria for NNT • To integrate multiple benefits and risks • Also see NNT
Impact numbers	• Similar to NNT and based on classical epidemiological metrics • Several impact numbers were proposed for different purposes • Provide population perspective	• Emphasise importance of justifying data sources • Not suitable for rare idiosyncratic reactions • Better interpretation to the	Yes	• Similar to NNT • Taken population of interest data into account • Relatively new concept with potential in benefit-risk	• To provide population perspective based on the number of people in the population of interest • To characterise benefit-risk

Approach	Features	Comments	Has it been recommended that this be taken forward to the next stage?		
				Reasons	Specific use
	• Does not integrate benefits and risks • Two of the metrics (NEPP and PIN-ER-t) do not require reciprocation	general audience • NEPP and PIN-ER-t do not suffer disadvantages of NNT		assessment particularly in epidemiology	balance in specific populations • Also see NNT
NEAR	• Avoids null point and "sign" problem of NNT • Presents results as relative risks or odds ratios • Compares one benefit and one risk • Tables and forest plots as visuals	• There is extension to deal with intention-to-treat and per protocol analyses • Uses expected frequencies hence does not need marginal probabilities	No	• Similar to NNT and AE-NNT • No clear advantage compared to other NNT-related metrics	• To integrate one benefit and one risk • 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 • Integrates one benefit and one risk	• Requires comparison of two active treatments • Statistical properties are not well-studied	No	• Similar to NNT • Statistical properties are not well-studied	• To assess the threshold at which efficacy can be established
RV-MCE	• Extension of MCE to include utilities • Integrates multiple benefits and risks	• Requires comparison of two active treatments • Statistical properties are not well-studied	No	• Similar to RV-NNH and MCE	• To account for subjective judgments on criteria in MCE analysis • Also see MCE
MAR	• Compares one risk to multiple benefits individually • Closely linked to utility survey techniques • Bar and antenna graphs as visuals	• Assumes benefit only occurs when risk does not (mutually exclusive events)	No	• Similar to SPM • Mutually exclusive events assumption does not normally apply	• To assess the threshold at which risk becomes no longer acceptable

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?		
				Reasons	Specific use
QALY	• Measure of life time with quality of life incorporated • Instruments used to derive quality of life are usually validated • Integrates benefit and risk and includes time dimension • Scatter plots as visuals	• Can be derived in a number of ways • The most appropriate health instrument for deriving QALYs is subjective in some areas • Validation for health instruments may not be in the population of interest	Yes	• Provides a measure of time trade-off with life quality • 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 • Conceptually opposite of QALY • Can act as a population measure	• See QALY	No	• Similar to QALY • 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 • Confined to survival endpoints only • Sensitivity analysis can be performed on choice of utility • Visualisation by stratified survival curve for one treatment only	• Developed for oncology • Easy to understand • Health states are defined	Yes	• QALY modified for cancer therapy	• To assess B-R of cancer therapy using defined health states

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?		
				Reasons	Specific use
INHB	• Commonly used with health indices • Implicitly assumes equal weights for benefits and risks • 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 • Easy to perform and understand	Yes	• Simple and intuitive • Uses established health indices such as QALYs	• To assess and integrate benefits and risks when described by health indices • Also see QALY
BRR	• Can only deal with one benefit and one risk • Assumes equal weighting of benefit and risks • Similar to NNT	• Not transparent for benefit-risk assessment when used in its simplest form • May be derived using thorough evidence synthesis and statistical modelling • Should only be used with high quality data • Must be presented together with their absolute or baseline rates	Yes	• Intuitive • Simple to calculate • Commonly used with other indices	• To assess and integrate benefit and risk • To characterise the equilibrium point when benefit equals risk
GBR	• Integrates benefit and risks • 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 • Collectively analysing criteria may result in loss of information • 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. • They are also not directly comparable to other measures	• To assess and integrate multiple benefits and risks

Approach	Features	Comments	Has it been recommended that this be taken forward to the next stage?		
				Reasons	Specific use
Principle of three	• Simple multi criteria model • Employs three criteria: “disease”, “effectiveness” and “ADRs” • Benefit and risk are not integrated • Tables as visuals	• Simple • Low discriminative scoring • Not suitable for complex situations • 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 • May be used for preliminary assessment
TURBO	• Simple multi-criteria decision making approach • Accommodates two benefit criteria and two risk criteria • Second criteria is regarded as correction factor to the first and scored on a shorter scale • 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 • Limited to two criteria	• To provide a snapshot of simple benefit-risk assessment • May be used for preliminary assessment
Beckmann Model	• Simple multi-criteria model • Quality of data contributes to the scores • Does not integrate benefit and risk	• Easy to perform • Does not take into account relative importance of benefit and risk criteria • Does not incorporate uncertainties	No	• Does not provide a robust metric for benefit risk assessment • Similar to principle of three	• To provide a snapshot of simple benefit and risk assessment • May be used for preliminary assessment

6 Estimation techniques for benefit-risk assessment

6.1 Introduction

Estimation techniques are not unique to benefit-risk assessment. They are general supporting tools for statistical modelling and parameter estimations. A wide variety of *estimation techniques* is available. The main role of these techniques in drug decision making is to bring together evidence of benefits and risks in order to quantify and communicate the trade-offs to decision makers. Because the quality of evidence data and data representativeness are vital in decision making [42;43], this section discusses how evidence is dealt with in benefit-risk assessments. This section is not devoted to addressing statistical modelling techniques in earnest but only to address those that are identified through benefit-risk literatures and are commonly used in this field.

Section 6.2 compares the operational characteristics of the *estimation techniques* to provide decision makers with a snapshot of the relevant parameters and capabilities for each technique. Brief appraisals of the different techniques follow in Section 6.3. Conclusion and recommendations to take forward suitable *estimation techniques* are made in Section 6.4.

Appendix A.10 appraises each *estimation technique* in more detail according to the set appraisal criteria.

6.2 Operational characteristics of estimation techniques

The operational characteristics of the estimation techniques are presented in Table 13. These are broadly the parameters that each approach can accommodate, and some main features. The resultant metrics (and many other parameters) are irrelevant here because these are estimation techniques therefore the results are heavily dependent on specific models.

Table 13 Operational characteristics of estimation techniques

	DAGs	PSM	CPM	ITC / MTC	CDS
π	X	X	X	X	O
U	X	O	O	O	O
S	O	O	O	O	O
w	O	O	O	O	O
I	O	O	O	O	O
T	O	O	O	O	O
ζ	O	O	O	O	O
G	X	O	X	O	O
M	-	-	-	-	-

π = require probability, S = scoring involved, U = require utility, w = require weights, I = Integrate risk and benefit, T = integrate time trade-off, ζ = explicit sensitivity analysis required, G = graphical methods proposed, M = the resultant quantitative benefit-risk metric. X indicates relevant parameters; O indicates optional parameters.

6.3 Appraisal of estimation techniques

In the realisation of a benefit-risk assessment model, the decision problems, the available evidence, the preference values and other assumptions are often brought together in a network. Networks (including value tree and decision tree) enable decision makers making visual connections between the relevant pieces of information.

The directed acyclic graphs (DAGs) can bring together pieces of information in connected networks that underpin a decision model. Through DAGs, the dependency and independency of different piece of information can be accurately established. DAGs require that the nodes corresponding to each piece of evidence to be acyclic to allow valid inferences to be made from the model. Acyclic means that a node cannot be dependent on itself, that is, there should not be circular references in the network. Conditional probabilities are placed on each node and then benefit-risk trade-offs can be estimated (Appendix A.10.1).

DAGs are sometimes (and often) regarded as the rudimentary structure for the application of other techniques of estimation, for example, for use with probabilistic simulation model (PSM), confidence profile method (CPM), and indirect and mixed treatment comparison (ITC and MTC). PSM is a stochastic estimation technique using probability distributions. This includes the more established Monte Carlo method of sampling from statistical distributions a large number of times. **PSM has the ability to propagate uncertainties in the simulated values from the uncertainties in the input values (evidence data and assumptions)** through the ‘DAGs-like’ network of connected nodes. The applications of PSM are highly reliant on high quality direct evidence data and account for uncertainties (Appendix A.10.2).

A more elaborate version of DAGs can be seen through the proposal of the confidence profile method (CPM) which carefully connects different pieces of information using "chains of evidence" [44]. CPM specifies probability distributions for single link chains for direct evidence and multiple link chains for linking together indirect evidence. The essence of CPM is in dealing with functional biases through careful parameterisation of the likelihood functions. It is a powerful and flexible technique for use in benefit-risk decision making but its existence is not well known among statisticians (Appendix A.10.3). However, its 'successor', the mixed treatment comparison (MTC) has gained much attention in recent years.

The mixed treatment comparison (MTC) can be thought of as a generalisation of the CPM. MTC concerns appropriately **combining different pieces of evidence** in order to warrant the results of an estimation or simulation model. MTC reduces to ITC in the case when evidence of direct comparison is unavailable. Benefit-risk assessments based solely on ITC may be necessary but are not enough to justify benefit-risk trade-off due to the lack of direct evidence data. Although MTC may be a general (meta-analytic) estimation technique, its popularity and the awareness of its users of the **issues surrounding evidence synthesis** make it an attractive approach for further consideration in drug benefit-risk decision-making.

The final *estimation technique* that we considered is the cross design synthesis (CDS) which proposes to combine randomised clinical trials evidence with evidence from clinical databases [45]. The application of CDS is intended to complement the weaknesses of one study design with another's strengths. CDS also specifically addresses the issues of bias in evidence synthesis through statistical adjustments, having conducted focussed assessment on each piece of evidence (Appendix A.10.5). The value CDS adds to benefit-risk assessment is the consideration on combining evidence from different study designs and the appropriate assessment of associated biases in doing so.

Table 14 summarises the appraisal on *estimation techniques*. The levels of complexity of these approaches are generally very demanding because of their focus in the synthesis of evidence. Users of these approaches require a wide range of knowledge from knowing the sources of evidence, including their strengths and weaknesses, awareness of available frameworks, appropriate metrics, appropriate statistical inference techniques, and arguably more importantly is the computational expertise for applying these techniques. DAGs may be of medium complexity to being very complex depending on the size of the network. All *estimation techniques* may address more than two options in the same model, could use population or individual level data in the analysis, and are relevant for the same stakeholders – which are listed in the table below.

Table 14 Comparison of estimation techniques

	Level of complexity	**Number of options**	**Evidence data**	**Perspective for stakeholders**
DAGs	Medium	> 2	Population or individual	pharmaceutical companies, healthcare providers, regulatory agencies
PSM	Complex	> 2	Population or individual	pharmaceutical companies, healthcare providers, regulatory agencies
CPM	Complex	> 2	Population or individual	pharmaceutical companies, healthcare providers, regulatory agencies
ITC and MTC	Complex	> 2	Population or individual	pharmaceutical companies, healthcare providers, regulatory agencies
CDS	Complex	> 2	Population or individual	pharmaceutical companies, healthcare providers, regulatory agencies

6.4 Conclusion and recommendations

The focus of *estimation techniques* in drug benefit-risk decision making is to ensure appropriate decision models are fitted in appropriate situations. Many aspects of statistical modelling and decision making can be addressed through these techniques, including statistical uncertainties, covariate adjustments, biases, heterogeneities, and quality of evidence. *Estimation techniques* can also be used to estimate personalised benefit-risk trade-offs based on individual's characteristics [46;47], which enhance the capability of these techniques for the purpose of decision making. However, for the *estimation techniques* to be applied in regulatory settings, good clinical practice guidelines must be adhered at all times – or used in combination with the recommended frameworks (Section 4.5) – to increase transparency in the justifications and decisions.

Recommendations of estimation techniques are not very straightforward because they are intertwined. It may be necessary to use some aspects from different techniques in order to reach the goal more satisfactorily. Table 15 shows the comparative overview and justifications of the recommendations for estimation techniques. We wish to recommend the following *estimation techniques* for further consideration in the next step of this project:

(11) Probabilistic Simulation Method (PSM). The abilities of PSM to deal with statistical adjustments and different kind of uncertainties under different assumptions are its most appealing features. The use of high quality evidence data for the application of PSM is desirable as the end results of the simulations are highly dependent on the input values and the assumptions in the underlying models.

(12) Mixed Treatment Comparison (MTC). MTC is recommended here as the method of choice because it can flexibly accommodate many important aspects of evidence synthesis in different circumstances. Issues of biases, as addressed in the confidence profile method should be considered in an MTC model as should other issues of combining different types of evidence addressed in cross design synthesis.

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?		
				Reasons	Specific use
DAGs	• Graphical method • Uses principles from Bayes Network • Network diagrams as visuals	• Graphical feature can help in communicating the characteristics of the underlying decision • Similar to MDP	No	• Similar to MDP	• To establish the relationship between evidence • To support the application of benefit-risk assessment • To assess benefit and risk
PSM	• Uses Monte-Carlo simulation or re-sampling form original data • Can be applied to any type of data • 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 • Flexibility	• To support the application of benefit-risk assessment • To assess benefit and risk • To deal with uncertainties
CPM	• Deals with multiple benefit-risk criteria • Deals with multiple sources of evidence • Evidence easily updated under the Bayesian framework	• Mathematically exhaustive • Requires extensive mathematical modelling expertise	No	• May be difficult to apply being mathematically exhaustive • Somewhat difficult for routine use • Similar to MTC	• To support the application of benefit-risk assessment • To assess benefit and risk • To deal with uncertainties and varieties of data
ITC	• A meta-analytic approach • Allows comparison of two options indirectly through common denominator when direct evidence is unavailable • Flexible	• Offer increased transparency in terms of clarifying the sources of evidence, bias and uncertainties • Otherwise incomparable options can be compared	No	• Superseded by MTC	• To support the application of benefit-risk assessment • To assess benefit and risk • To deal with uncertainties • To compare options where there is no direct evidence
MTC	• Generalisation of ITC • Include both direct and indirect evidence • Flexible to deal with complex structures	• Similar to ITC • Collapsed to ITC when there is no direct evidence • Requires statistical modelling expertise but fairly straightforward to understand • Computer codes to implement MTC are available	Yes	• Flexible to accommodate many aspects of evidence synthesis • Flexible to deal with complex structures • Computer codes to implement are available	• To support the application of benefit-risk assessment • To assess benefit and risk • To deal with uncertainties • To compare options where there is no direct evidence • To improve inference using both direct and indirect evidence

Approach	**Features**	**Comments**	**Has it been recommended that this be taken forward to the next stage?**		
				Reasons	**Specific use**
CDS	• A meta-analytic approach • Focusses on potential biases form study designs weaknesses • Focus on synthesising the evidence instead of comparing the outcomes. • 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 • Principles can be adopted using MTC	No	• Principles can be adopted using MTC	• To support the application of benefit-risk assessment • To assess benefit and risk • To deal with uncertainties and biases • To combine different sources of evidence

7 Utility survey techniques for preference elicitation in benefit-risk assessment

7.1 Introduction

Another important element in decision making is the preference values, whether scores or utilities. Through this elicitation, value judgments are incorporated. Value judgments from different stakeholders may vary and may affect the final benefit-risk assessment or decision. Thus preference values from appropriate stakeholders for a particular decision problem should be obtained in order to make the decision model count.

This section describes some *utility survey techniques* for the elicitation of preference values from relevant stakeholders. The approaches described here have descended from the same lineage. Section 7.2 describes the operational characteristics of the *utility survey techniques*. Section 7.3 describes and appraises these techniques in general, and Section 7.4 concludes their capacity to contribute in drug benefit-risk decision making, and carries on to make suitable recommendations of techniques to take forward for further consideration and testing.

The detailed appraisal of each *utility survey technique* is discussed in Appendix A.11.

7.2 Operational characteristics of utility survey techniques

Many parameters in Table 16 are included here for completeness although the differences are minimal. The operational characteristics for all *utility survey techniques* are the same. Their implementation requires utilities and scores from relevant stakeholders; and output either utilities or weighted utilities.

Table 16 Operational characteristics of utility survey techniques

	SPM	CV	CA	DCE
π	O	O	O	O
U	X	X	X	X
S	X	X	X	X
w			X	X
I	O	O	O	O
T	O	O	O	O
ζ				X
G				
M	U_E	U_E	U_w	U_w

π = require probability, S = scoring involved, U = require utility, w = require weights, I = Integrate risk and benefit, T = integrate time trade-off, ζ = explicit sensitivity analysis required, G = graphical methods proposed, M = the resultant quantitative benefit-risk metric. X indicates relevant parameters; O indicates optional parameters. U_E = expected utility, U_w = weighted utility.

7.3 Appraisal of utility survey techniques

The series of *utility survey techniques* appraised here are based on stated preferences, which are values or choices placed on hypothetical scenarios derived specifically for a particular decision problem. Incidentally, the umbrella approach that encases these techniques is known as the stated preference method (SPM). Following directly from the definition of "stated preference", SPM explores the response of relevant stakeholders for various hypothetical scenarios related to the problem to be solved (Appendix A.11.1). SPM can be divided into contingent valuation (CV) and conjoint analysis (CA).

Contingent valuation (CV) places monetary value on trade-offs using the willingness-to-pay concept. Stakeholders value the option and consequence; that is by stating their willingness to pay for a hypothetical treatment option given the hypothetical consequences. CV suffers from known bias for being over-sensitive if changes to the hypothetical scenario directly affect the stakeholders, and for being under-sensitive if they do not (Appendix A.11.2).

Conjoint analysis (CA) breaks decision problems into smaller pieces to be evaluated separately. Stakeholders then place their preference values on the smaller pieces of the hypothetical scenarios, where preference independence assumption applies. The utilities elicited from the hypothetical scenarios are then combined to arrive at the overall utilities (Appendix A.11.3).

The discrete choice experiment (DCE) takes CA further by providing a structured framework to eliciting utilities from relevant stakeholders. DCE provides guidelines on defining important characteristics of a decision problem; defining and assigning scoring systems; systematically formulating orthogonal configurations of criteria (or attributes); and collating the orthogonal configurations of criteria into choice sets to create the hypothetical scenarios. The hypothetical scenarios generated are used to elicit utilities from relevant stakeholders. The utilities are then combined and the expected utilities are estimated (Appendix A.11.3).

Table 17 shows comparative overview of the *utility survey techniques*. The flexibility of the techniques to incorporate suitable scoring systems allows them to be highly discriminative of the options against the criteria. SPM in its simplest from, and CV are of medium complexity because they require some knowledge of survey techniques and basic statistical concepts. CA and DCE are complex since they require extensive knowledge of survey techniques and statistical experimental designs as well as regression techniques. All *utility survey techniques* can handle more than two options simultaneously, and require individual level data to implement. Individual level data here can refer to individual person or a group of people representing relevant stakeholders.

Table 17 Comparison of utility survey techniques

	Discriminative scoring	**Level of complexity**	**Number of options**	**Evidence data**	**Perspective for stakeholders**
SPM	High	Medium	> 2	Individual	Any
CV	High	Medium	> 2	Individual	Any
CA	High	Complex	> 2	Individual	Any
DCE	High	Complex	> 2	Individual	Any

7.4 Conclusion and recommendations

Although this report was initially intended to appraise benefit-risk assessment approaches, we have come to recognise that *utility survey techniques* also play crucial role in decision-making, and therefore have included them here. Their applications can help to increase transparency in drug benefit-risk assessment as they provide robust and justifiable value judgments. By systematically eliciting utilities, means that the process can be easily replicated or even reusable for different decision problems.

Table 18 shows the comparative overview and justifications of the recommendations for utility survey techniques. Based on our appraisal, the obvious recommendation of a *utility survey technique* to take forward is the:

(13) Discrete Choice Experiment (DCE). The utility survey techniques from the stated preference methods family are suitable for the purpose of eliciting utilities but we recommend DCE as an approach to elicit utilities. This is simply because these approaches are similar and roughly based on the same principles but DCE provides the most comprehensive instructions on the steps required from designing the elicitation process to the methods of combining the utilities obtained to be used in a benefit-risk assessment model. The use of DCE in drug benefit-risk decision making is less mature than other recommended approaches but its potential needs further exploration. Conducting a DCE is resource-consuming. Therefore alternatives would be required. A structured elicitation is encouraged because ultimately obtaining appropriate value judgments from relevant stakeholders is crucial to the validity of any specific benefit-risk decision analysis.

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?		
				Reasons	Specific use
SPM	• Benefit and risk described through a hypothetical scenario • Accommodates multiple benefits and risks which has the potential to vary over time • Can collect large amounts of data with moderate cost • 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 • Similar to DCE	• To elicit preference values through a hypothetical scenario
CV	• Places monetary value on trade-offs using the willingness to pay concept • Assumes benefits and risks in medicine behave like market goods • Similar to SPM	• Known bias for being to over-sensitive if stakeholder directly affected and opposite if not • Monetary valuations differ between people • 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 • Treatments may be free • Similar to DCE	• To elicit preference values through a hypothetical scenario • 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 • Based on random utility theory and statistical experimental design • Hypothetical scenario is broken down to a specific number of attributes before evaluated • Anything can be defined as an attribute including time • Minimises bias in response	• Requires statistical expertise in experimental designs • Takes time • There are some debates on internal validity, consistency and test-retest reliability • Can be used to investigate how specific attributes may be viewed differently by different stakeholders	Yes	• Well-structured approach • Most comprehensive of the utility survey techniques reviewed • Provides transparency in the elicited preference values	• To elicit preference values through hypothetical scenarios • To make robust inference on preference values • To collect preference values for use with other benefit-risk approaches

8 Discussion

8.1 Foreword

This report has reviewed quantitative benefit-risk assessment approaches for use in drug regulation. Some of these approaches have been around for over three decades but have only been excavated in recent years due to the arising interests in and the need for formal quantitative benefit-risk decision-making. Many authors have attempted to refine and revive these approaches for the use in the current climate which, has resulted in many variations of generally a similar concept.

Despite the multitude of available approaches and worked examples, their applications in real-life benefit-risk decision making are still limited. Their complexity in applications and the lack of good worked examples may be the limiting factors. The dearth of having direct comparison of the performance and their effects on the decisions for the same problems may also contribute to this circumstance. Currently, there is no consensus on which approach should be preferred in which situation.

The mathematical basis and principles of the approaches are discussed in Section 8.2. The discussion on classification strategy adopted here then follows in Section 8.3. We summarise the discussion on the evaluations of the approaches in Section 8.4, and discuss in Section 8.5 their inter-relationship and the necessity to use more than one approach in combination. The issues of appropriate use of evidence cropped up many times and we reflect on this topic in Section 8.6. The issues on stakeholders' perspective are discussed in Section 8.7; and the issues on implementation of benefit-risk approaches for real life decisions and their communications to the general public are discussed in Section 8.8.

8.2 Mathematical basis

In terms of the underlying mathematical principles, there are three distinct groups: no formal mathematical basis, probability theory, and decision theory. Approaches that have no formal mathematical basis are practically general instructions for good practices when performing benefit-risk assessments. However, some of these guidelines may have been developed based on the ideas from decision theory for example actions have consequences therefore these elements need to be addressed. In more formal settings, these resemble, for example, the standard operating procedures or the CONSORT statements [48;49] in clinical trials, and the STROBE statements [50] in epidemiological studies.

Approaches that are based on probability theory are generally simple and easily understandable, but often missing the much required element in decision making that is the incorporation of utilities and preference values in the inference. Some authors have attempted to rectify this issue by introducing the missing element into the probability-based approaches but this has been criticised by decision theorists for violating the theory of decision because the interpretations of the outcomes are still within the domain of probability theory, thus conflict with the decision theory. On the other hand, these approaches offer straightforward interpretation to many people as the concept of probabilities (or fractions or rates) lies within their familiar territories.

The alien concept (to general stakeholders) of utilities forms a cornerstone of the approaches based on decision theory. Utilities are important to characterise the "true" benefit-risk balance specific to stakeholders facing a decision problem based on the stakeholders' risk attitudes. This is analogous to the recent development of the research in personalised medicines where treatment decisions are to be tailored to individual patient's characteristics [46;47]. Although these approaches have all the required elements for a comprehensive and well-informed decision, they are generally more resource consuming (time, workforce and money) when compared to the simpler probability-based approaches.

8.3 Classifications

The classifications strategy that we have adopted provides a structured outlook on some of the approaches that are currently available and have been used for the assessment of benefit-risk trade-offs in medicine. The classifications have gone through several refinements in terms of the categories as well as which approach goes under which category. The working definitions for these classifications are listed in Appendix A.3. There might still be some overlaps but this was necessary to avoid having too many unmanageable categories. Wherever appropriate, we have addressed these in the text.

The most important take-home message from this classification exercise is that the approaches that come under a category are only useful and should be used within the context of that category (see Section 8.4).

8.4 Evaluations

It has been anticipated from the beginning that no one approach would satisfy all the appraisal criteria set in this review. This is because the criteria are diverse and it is necessary to be so in order to address the diverse interests of this project. Whilst this review provides “high level” descriptions and evaluations, it should help to give overviews of the many approaches that are available in medical decision making as well as to clarify their roles and domains of applications. References to their more detailed descriptions should be followed to learn more about the approaches.

Follow on from our comment that approaches within each category should be used only within their context (Section 8.3); we argue that *descriptive frameworks* do not perform benefit-risk assessment but only framing the decision problems that would ensure transparency. On the other hand, *metric indices* do not properly frame the decision problems but provide quantitative measurements for the 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 considerations but do not allow trade-offs. The *quantitative frameworks* do both in terms of framing the decision problems and then progress to perform benefit-risk trade-offs.

The more specific considerations are the evidence side of decisions problems. In more complex situations especially, simple benefit-risk assessment models may impose implausible assumptions on the evidence thus become inappropriate. In these cases, complex evidence synthesis and simulation models may be more appropriate which could be achieved through the applications of the *estimation techniques*. Some data vital to the applications of these approaches come in the form of preference values which require formal elicitations from relevant stakeholders; and can be obtained through the *utility survey techniques* we have described.

These also apply to the interpretation of their results.

8.5 Inter-relation and combination of approaches

The application of the approaches within and between groups is not necessarily exclusive; and most often there is a need for several approaches to be used in combination or in parallel. The simple reason for this is, many approaches are related to one another and many have been derived based on the idea of another, for example the NNT family of approaches. We have cross-referred throughout the report – particularly in the extensive Appendices – these similarities to demonstrate this idea and hopefully lay on the table the reasoning as to why we have pursued with some recommended approaches but not the others.

A combination of approaches does not necessarily mean that the different approaches are used in whole and 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 to estimate benefit-risk ratio through probabilistic simulations.

8.6 Reflection on evidence requirement for decision making

The key to good decision making lies in the representativeness of the data used in a decision model. Although many benefit-risk trade-offs are often based on data from clinical trials, the use of observational data has grown in recent years with the availability of many public clinical databases such as the General Practices Research Database (http://www.gprd.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.

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. Furthermore, the true benefit-risk balance emerges over time, and not as a “snapshot” as seen at the time of marketing authorisation applications. Similarly for risk endpoints, some risks or adverse events occur immediately and others might take longer to surface. This makes the timing of benefits and risks more difficult because it may take longer to observe the benefits or risks of a treatment. In the worst case scenario, benefits or risks may not be observed at all within the planned assessment period. These issues have been known to be the limitations of clinical trials. At a technical level, Cross Design Synthesis addresses these issues specifically and attempted to compensate the weaknesses of clinical trials’ evidence with evidence from clinical databases, and vice versa.

Although observational data from clinical database could in principle provide the desired measure of “effectiveness” of a treatment, they have larger uncertainty related to the precision, bias and confounding. The consistency of records is also questionable but the implementation of coding systems such as the Read clinical classification codes and the Medical Dictionary for Regulatory Activities (MedDRA) (http://www.meddramsso.com/) does help, thus should be favoured over free text fields for analyses where possible. There is also the limitation in the type of benefit endpoints captured in such databases. More serious events like the (non-)occurrence of myocardial infarctions and deaths are recorded well, but less serious ones may not be recorded at all, for example recovery from certain diseases or infections. There may be more evidence of risks or adverse events available than the evidence of benefits from clinical databases because medical complaints are often recorded. However, the measure of severity of adverse events is not generally available. Furthermore, there may be some discrepancy between the timing of a recorded event and the timing of the actual event because of the delay in patients reporting the event to physicians. Compromise and reasonable assumptions should be made in decision models to account for such discrepancy. More specifically to the clinical databases mentioned at the beginning of this section, the evidence available is restricted to treatments and complaints through the general practices.

In real life decision-making, the circumstances are more complex than just identifying and using suitable evidence. The process of identifying evidence itself is also important and should be made transparent. Justifications for strategy to find sources of evidence and evidence selection should also be documented to minimise biases. Future uncertainties and likely scenarios, for example as required by the EU Risk Management Plan [51] are to be taken into account 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.

8.7 Stakeholders' perspectives

The perspectives of different decision makers may lead to different decisions being made, thus are to be made explicit at the beginning of a benefit-risk assessment exercise. Specifying the perspectives should involve the considerations of:

- 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.

Ideally, the perspectives of all stakeholders involved should be taken into account but it is not always possible or amenable. For example, although patients' involvement in regulatory decision-making is currently seen as an important way forward in this field, the regulatory agencies may not be prepared to value patients' preference with greater weights than their own when making regulatory decisions. This is because the decision-makers such as the regulatory agencies have specific questions to consider, and therefore certain issues may not be as relevant.

8.8 Implementation and communications

Through this review, we have demonstrated that visual representations already find its role in benefit-risk decision-making. Although many aspects of communications and disseminations revolve around analysis results, they are as important in justifying input values, assumptions, and to encourage transparent and robust benefit-risk assessments.

Communication of the results of benefit-risk assessments is vital, and should be done consistently. Difference in concepts should be addressed, and their cross-interpretabilities and translations should also be acknowledged to avoid confusions and misinterpretations. The availability of specialised software for the implementation of specific approach, and to generate 'standard' sets of results and visual outputs is desirable. In many cases, specialised software is unavailable but most of the times this is because the implementation is straightforward and can be easily performed in various statistical packages – either come as standard syntax commands or as add-ons, otherwise could be programmed in.

Many available software already incorporate visual representations as major components or interface in the applications of the approaches. This is demonstrated, for example, by various MCDA decision support software such as the Web-HIPRE (http://www.hipre.hut.fi/), IDS (http://www.e-ids.co.uk/), and Hiview 3 (http://www.catalyze.co.uk/products/hiview). They not only have good visual representations of benefit-risk balance of each option and visual interfaces for sensitivity analysis, they also use visual elements to set up decision problems and for inputting evidence data and preference values, as well as implementing scoring system. More detailed review of MCDA software is also available comparing their functionality and interfaces as well as their suitability for different decision making processes [52].

The graphical structure of decision problems is central to the implementation of DAGs, and as a result, several software packages have been developed to support inferences based on DAGs. For example, in the application of Bayesian Networks and Influence Diagrams, GeNIe software is freely available from http://genie.sis.pitt.edu/. More general software packages to perform Bayesian modelling such as WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml) and the R statistical software package JAGs (http://www-ice.iarc.fr/~martyn/software/jags/) are also freely available. The latter two however require the users to be more technically proficient in statistical modelling.

We have illustrated in the Appendices some of the visualisation techniques that have been used in communicating the ideas of benefit-risk trade-offs. More work is needed to investigate whether the proposed visualisation techniques are good or useful for conveying forward the precise meanings of the ideas, and their vulnerability to misinterpretation – deliberately or otherwise. The dissemination of our findings through the right channels is important and is to be explored further. Part II of PROTECT WP5 review will specifically address communications issues through their visual representations as well as addressing the importance of dissemination so that it reaches wider audience and receives fair jurisdictions.

9 Conclusion

9.1 Foreword

This section first collates the general methodology recommendations by category (Section 9.2) as made in their respective sections. The limitations of this review are addressed in Section 9.3, which then followed by some future plans of this project in Section 9.4. We make concluding remarks on this review in Section 9.5.

9.2 General methodology recommendations

The recommendation of methodologies to be taken forward for further consideration and testing is made according to the categories and therefore to be applied in the contexts of their categories.

In order to assess aspects of benefits and risks for decision-making descriptively as well as to set up a decision problem, the recommended methodologies are the *descriptive frameworks*. Of the *descriptive frameworks* reviewed here, we are particularly interested to take forward two:

(1) PrOACT-URL frameworks. PrOACT-URL provides a framework addressing the necessary elements in dealing with decision problems. PrOACT-URL however does not clearly address the importance of identifying appropriate sources of evidence and immediate parties involved. The application of PrOACT–URL therefore requires improvement in these two aspects, for example by combining with elements of the Ashby and Smith framework (ASF) or other evidence synthesis approach. The extension of PrOACT-URL to drug benefit-risk decision analysis by the EMA Benefit-Risk Methodology working group 2 may, in the future, provide better evolution of the frameworks in this specific domain.

(2) Benefit Risk Action Team (BRAT) framework. BRAT provides guidelines on organising, understanding and summarising evidence of benefits and risks into tabular outputs and graphical summaries. The framework proposes avoiding integration of benefits and risks evidence to make it more accessible and transparent to those not familiar with complex statistical models. The controversial aspect of BRAT is its proposal to use odds ratios as the basis for the decision on benefits and risks balance.

To assess benefit-risk balance as a whole within a defined framework, the *quantitative frameworks* can be used. These frameworks generally take the guidelines a step forward by proposing specific methods to quantify benefit-risk trade-offs having considered the important aspects of a decision problem. We would explore further the applications of:

(3) Multi-Criteria Decision Analysis (MCDA). MCDA provides structured stepwise instructions in line with PrOACT-URL framework with the capability of assessing and integrating multiple benefits and risks criteria, as well as comparing different options. MCDA is also the only approach that can formally deal with multiple objectives simultaneously. Another appealing feature of MCDA is that specialist several software to perform the analysis are available.

(4) Stochastic Multi-criteria Acceptability Analysis (SMAA). The SMAA families of approaches are potentially serious contenders to the standard MCDA because of the ability to account for sampling variations arising from the type of experimental designs used as well as when there is missing information on preference values. However, the increased complexity of SMAA compared to MCDA may be the major barrier in real-life benefit-risk medical decision making applications. The application of SMAA in medical decision making should be explored further as there is a potential that the same SMAA model could be used to address different stakeholders.

The *metric indices* can be used to quantify benefits and risks, whether integrated as a single metric or separately. Their use in benefit-risk assessments is to summarise the evidence numerically and may therefore help in communicating the results to more general audience. In simple cases, where benefits or risks are rates (or probabilities), and the interest is mainly to compare these rates, then simple *threshold indices* are widely used to provide crude cut-points as guide to decision-making. The two that should be explored further are:

(5) Number Needed to Treat (NNT) and Number Needed to Harm (NNH). The popularity and its widespread use in clinical literature when describing the benefits or risks of a treatment are well established. These indices, heavily criticised, still provide an attractive feature for benefit-risk assessment – simplicity. They should not be used naively, but be supported by thorough evidence synthesis and suitable modelling of the evidence. We are not recommending the utility-adjusted variants of NNT because the reciprocals of expected utilities do not have the same meaning as the reciprocals of probabilities or rates.

(6) Impact numbers. These metric indices give a different perspective from NNT, and are useful in describing public health burden of a disease, and the potential impact of a treatment. Although other more established epidemiological measures are available and have already been used widely, impact numbers have an intuitive interpretation. Their application in benefit-risk assessment may contribute to making the interpretation more accessible to the general audience.

The *health indices* are commonly used to quantify the trade-offs of benefits and risks in medicine, but many are intertwined and most are specific. They are all based on the same idea of trading off preference against time. They are widely acceptable since they intuitively place different benefit and risk criteria onto the same unit to allow straightforward comparison. The *health indices* should be considered further, and where appropriate, we recommend:

(7) Quality Adjusted Life Years (QALY). The QALY provides time trade-off with life quality in discrete health states which is absent from other more general trade-off indices. Its use is already established in many areas of medicine particularly in chronic diseases where time factor plays a major role in their assessment of benefits and risks. HALE may be another health index to be considered, but as it is just a summary of QALY in a particular group of people, we will not differentiate it further.

(8) Quality adjusted Time Without Symptoms and Toxicity (Q-TWiST). The discrete health states proposed in Q-TWiST are intuitive and very specific to cancer therapy. Therefore its usefulness is limited to cancer domain but nonetheless is a suitable metric to aid cancer patients to decide on the best acceptable treatment. We are only recommending the use of Q-TWiST within the cancer therapy domain.

The *trade-off indices* can be used to quantify benefits and risks into a single metric that describes the balance between them. The recommended *trade-off indices* to be taken forward are:

(9) Incremental Net Health Benefit (INHB). The INHB builds on health indices like QALY. INHB incorporates time and utility which are desirable elements in benefit-risk assessments. The idea of penalising "incremental" benefits by "incremental" risks in INHB, directly follows the simple intuitive concept of benefit-risk trade-offs.

(10) Benefit-Risk Ratio (BRR). Ratios provide intuitive trade-off metrics as they can be interpreted readily as the multiplicative effect of one relative to the other. The recommendation here is only for the application of BRR when accompanied by high quality evidence data and appropriate statistical modelling, and is presented together with their absolute or baseline rates. The weighting of benefits and risks to be traded off should be carefully taken into consideration when using BRR as the metric for benefit-risk assessment because equal weights assumption may not always, and in most cases do not, hold.

In many situations, thorough evidence synthesis and complex benefit-risk modelling are required to address the heightened complexity of a decision problem, particularly to address the uncertainties. The *estimation techniques* can readily accommodate these issues. In general, they are very similar. For simpler simulations, we recommend:

(11) Probabilistic Simulation Method (PSM). The abilities of PSM to deal with statistical adjustments and different kind of uncertainties under different assumptions are its most appealing features. The use of high quality evidence data for the application of PSM is desirable as the end results of the simulations are highly dependent on the input values and the assumptions in the underlying models.

More careful evidence synthesis is required when little direct evidence is available and the use of additional indirect evidence is required. To properly assess benefits and risks trade-offs from the mixture of evidence, we recommend:

(12) Mixed Treatment Comparison (MTC). MTC is recommended here as the method of choice because it can flexibly accommodate many important aspects of evidence synthesis in different circumstances. Issues of biases, as addressed in the confidence profile method should be considered in an MTC model as should other issues of combining different types of evidence addressed in cross design synthesis.

The last recommendation is for the purpose of eliciting stakeholders' preference values to be incorporated into benefit-risk assessments through the application of the *utility survey techniques*. Stakeholders' preference values form their utilities towards certain treatment-outcome (option-consequence) pairs. Their role is vital in decision-making but they are less mature in application compared to the other approaches. Nevertheless, their importance has started to reel in but more work is still required to ensure more routine applications in benefit-risk decision-making. The *utility survey technique* that we found to be the most comprehensive and should be taken forward is:

(13) Discrete Choice Experiment (DCE). The utility survey techniques from the stated preference methods family are suitable for the purpose of eliciting utilities but we recommend DCE as an approach to elicit utilities. This is simply because these approaches are similar and roughly based on the same principles but DCE provides the most comprehensive instructions on the steps required from designing the elicitation process to the methods of combining the utilities obtained to be used in a benefit-risk assessment model. The use of DCE in drug benefit-risk decision making is less mature than other recommended approaches but its potential needs further exploration. Conducting a DCE is resource-consuming. Therefore alternatives would be required. A structured elicitation is encouraged because ultimately obtaining appropriate value judgments from relevant stakeholders is crucial to the validity of any specific benefit-risk decision analysis.

These recommended approaches are neither meant to be exhaustive nor meant to be restrictive. Other approaches we reviewed here may also be suitable for benefit-risk decision-making so long as the decision makers are willing to accept the underlying assumptions that are associated with them. However, we believe that the use of the recommended approaches whether individual or in combination would be sufficient to address the benefit-risk trade-off in medicine for various stakeholders.

9.3 Limitations of this review

We have performed this review based on approaches that were reviewed elsewhere. We did not attempt literature search for individual approaches. As a result, newer approaches that we were not aware of or unpopular approaches may have been missed. However, it is not necessary at this stage to perform comprehensive literature search due to the variety of approaches included in this review.

This review did not include disease-specific benefit-risk approaches or other model-based approaches. The reason for excluding disease-specific approaches is because they are too specific to be used in different contexts. There are many model-based approaches in benefit-risk literature but because model-based approaches are often combinations of other approaches, they are excluded. Both disease-specific and model-based approaches may contain some useful considerations; therefore future benefit-risk research may gain some insights from learning from these approaches. On the other hand, considerations from these approaches often arise automatically when setting up decision problems, and eventually can be dealt with appropriately.

We also acknowledge that some authors have worked extensively with some methods, and therefore their descriptions and evaluations may contain greater details, but we have tried to make them as consistent as possible. This may also lead to favouring of certain approaches over another and therefore our recommendations may be biased towards these. We took into account both sides of arguments in many situations, but some may still be unresolved.

Many decision makers simply favour simplicity over complexity when performing benefit-risk assessments but without robust experimental evidence of the superiority of any approach, we are unable to make our recommendations based on this criterion. Superiority of approaches is very difficult to test because it is difficult to tell which decision is the correct decision. In the end, if a simple approach results in different decision to a more complex one, it may just be down to good faith that a more complex model addressing all the appropriate aspects of benefits and risks would be the approach to use to arrive at the “correct” decision.

9.4 Future directions

In preparing this report, we learned various aspects many previous authors and guardians of the approaches have been striving to advocate repeatedly. Five aspects come to light as being the most important:

1) Transparency
   The decision process needs to be clear.
2) Evidence
   The evidence should be of high quality and representative of the decision problems.
3) Uncertainty and bias
   These should be addressed and quantified appropriately
4) Integration
   Benefits and risks should be assessed together
5) Communications
   The communication of the benefit-risk trade-offs needs to be concise, easily interpretable and not misleading.

Aspects (1) to (4) have been addressed in this report, as well as in many other literatures. We will next proceed to apply recommended approaches in wave 1 case studies, highlighting and carefully consider these aspects, to assess the practicality in applying them to real life decision problems when difficult decisions are to be made. The feedback from wave 1 case studies task forces will be collated, where necessary we will review the recommendations, and then finalise the methodology recommendations based on the practical experience.

The last aspect on communicating the benefit-risk trade-offs forward is still unresolved but many visual representations have been proposed and used. We have demonstrated some of the visual representations related to some approaches to support decision-making (see the Appendices). These visual representations will form the basis for Part II of this review where we will review suitable visual techniques to represent benefit-risk trade-offs, as well as the means for communicating and disseminating them. The review on visual representation will proceed to make recommendations of the type of visuals most useful to concisely describe the concepts and outcomes of a benefit-risk approach.

Wave 1 case studies and Part II of the review on visual representations and communications will take place in parallel. Their findings may contribute to each other's work. The outcomes from the wave 1 case studies and Part II of the review will then be carried into wave 2 case studies involving more complex decision problems, and subsequently the final consensus will be applied to wave 3 case studies to assist future work of PROTECT Work Package 6 on validation of methodologies for benefit-risk assessments.

9.5 Concluding remarks

We attempted to cover as completely as possible the plethora of approaches currently available in the literature that has been used or described as benefit-risk assessment approaches. Although, the list of reviewed approaches is long, it is by no means complete but should cover the basic grounds and the mainstream approaches.

The classifications strategy adopted to categorise the approaches should clarify their stance in benefit-risk assessments and how they are to be used for decision-making. We hope that our effort will help to make choosing suitable approaches from many available more straightforward, and thus help making better decisions.

10 Acknowledgement

We would like to thank Mrs Jane Okwesa (Imperial College London) for proofreading this review for publication.

11 References

(1) Phillips LD, Fasolo B, Zafiropoulos N, Beyer A. Benefit-risk methodology project work package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. London: European Medicines Agency; 2010 Aug 31. Report No.: EMA/549682/2010.

(2) Hammond JS, Keeney RL, Raiffa H. Smart Choices: A Practical Guide to Making Better Decisions. Boston, MA: Harvard Business School Press; 1999.

(3) Ashby D, Smith AF. Evidence-based medicine as Bayesian decision-making. Stat Med 2000 Dec 15;19(23):3291-305

(4) Spiegelhalter D, Abrams K, Myles JP. An Overview of the Bayesian Approach: Decision Making. Bayesian Approaches to Clinical Trials and Health-Care Evaluation.Chicester, West Sussex: John Wiley & Sons Ltd.; 2004. p. 49-120.

(5) Coplan PM, Noel RA, Levitan BS, Ferguson J, Mussen F. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. Clin Pharmacol Ther 2011 Feb;89(2):312-5

(6) Levitan BS, Andrews EB, Gilsenan A, Ferguson J, Noel RA, Coplan PM, et al. Application of the BRAT framework to case studies: observations and insights. Clin Pharmacol Ther 2011 Feb;89(2):217-24

(7) Jenkins J. A United States Regulator's Perspective on Risk-Benefit Considerations. Rockville, MD: Shady Grove Conference Center; 2010.

(8) Walker S, McAuslane N, Liberti L, Salek S. Measuring benefit and balancing risk: strategies for the benefit-risk assessment of new medicines in a risk-averse environment. Clin Pharmacol Ther 2009 Mar;85(3):241-6

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

(10) CIRS Workshop Synopsis on Benefit-Risk. 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_June_2012_Workshop_Synopsis.pdf . 20-6-2012. Washington, DC, Centre for Innovation in Regulatory Science.

(11) Chuang-Stein C. A New Proposal for Benefit-Less-Risk Analysis in Clinical Trials. Controlled Clinical Trials 1994;15:30-43

(12) Sutton AJ, Cooper NJ, Abrams KR, Lambert PC, Jones DR. A Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. J Clin Epidemiol 2005 Jan;58(1):26-40

(13) Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, et al. Decision making in health and medicine: Integrating evidence and values. Cambridge: Cambridge University Press; 2001.

(14) Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. Med Decis Making 1993 Oct;13(4):322-38

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

(16) Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines - part 1: the development of a new model using multi-criteria decision analysis. Pharmacoepidemiol Drug Saf 2007 Jul;16 Suppl 1:S2-S15

(17) Tervonen T, van Valkenhoef G, Buskens E, Hillege HL, Postmus D. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. Statist Med 2011;30(12):1419-28

(18) Sarac SB, Rasmussen CH, Rasmussen MA, Hallgreen CE, Soeborg T, Colding-Jorgensen M, et al. Balancing benefits and risks: Data-driven clinical benefit-risk assessment. 2011.

(19) Renard D, Wu K, Wada R, Flesch G. Using desirability indices for decision making in drug development. Population Approach Group Europe (PAGE); 2009 Jun 23; St. Petersburg, Russia.: Novartis Pharma AG, Basel, Switzerland; 2009.

(20) Ouellet D. Benefit-risk assessment: the use of clinical utility index. Expert Opin Drug Saf 2010 Mar;9(2):289-300

(21) Rowland M, Tozer TN. Therapeutic response and toxicity. Clinical pharmacokinetics: concepts and applications. 1 ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 1980.

(22) Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2 ed. Blackwell; 2003.

(23) Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. Value Health 2010 Aug;13(5):657-66

(24) Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988 Jun 30;318(26):1728-33

(25) Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA 1995 Dec 13;274(22):1800-4

(26) Schulzer M, Mancini GB. 'Unqualified success' and 'unmitigated failure': number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events. Int J Epidemiol 1996 Aug;25(4):704-12

(27) Guyatt GH, Sinclair J, Cook DJ, Glasziou P. Users' guides to the medical literature: XVI. How to use a treatment recommendation. Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group. JAMA 1999 May 19;281(19):1836-43

(28) Boada JN, Boada C, Garcia-Saiz M, Garcia M, Fernandez E, Gomez E. Net efficacy adjusted for risk (NEAR): a simple procedure for measuring risk:benefit balance. PLoS One 2008;3(10):e3580

(29) Boada J, Boada C, Garcia MM, Rodriguez C, Garcia M, Fernandez E. Net efficacy adjusted for risk: Further developments. Expert Opinion on Drug Safety 2009;8(6):649-54

(30) Holden WL, Juhaeri J, Dai W. Benefit-risk analysis: examples using quantitative methods. Pharmacoepidemiol Drug Saf 2003 Dec;12(8):693-7

(31) Johnson FR, Ozdemir S, Mansfield C, Hass S, Siegel CA, Sands BE. Are adult patients more tolerant of treatment risks than parents of juvenile patients? Risk Anal 2009 Jan;29(1):121-36

(32) Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health 2009 Mar;12 Suppl 1:S5-S9

(33) Gelber RD, Cole BF, Gelber S, Aron G. Comparing Treatments Using Quality-Adjusted Survival: The Q-Twist Method. The American Statistician 1995 May 1;49(2):161-9

(34) Riegelman R, Schroth WS. Adjusting the number needed to treat: incorporating adjustments for the utility and timing of benefits and harms. Med Decis Making 1993 Jul;13(3):247-52

(35) Garrison LP, Towse A, Bresnahan BW. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. Health Aff (Millwood) 2007 May;26(3):684-95

(36) Chuang-Stein C, Mohberg NR, Sinkula MS. Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. Stat Med 1991 Sep;10(9):1349-59

(37) Edwards R, Wiholm BE, Martinez C. Concepts in risk-benefit assessment. A simple merit analysis of a medicine? Drug Saf 1996 Jul;15(1):1-7

(38) CIOMS Working Group IV. Benefit-Risk Balance for Marketed Drugs. Evaluating Safety Signals. 1998. Geneva, Council for International Organizations of Medical Sciences.

(39) Beckmann J. Basic aspects of risk-benefit analysis. Semin Thromb Hemost 1999;25(1):89-95

(40) WHO-UMC. WHO Adverse Reaction Terminology - Critical Term List. http://www.umc-products.com/DynPage.aspx?id=73589&mn1=1107&mn2=1664 . 1996. WHO Collaborating Centre for International Drug Monitoring.

(41) CIOMS Working Group III. Guidelines for preparing core clinical-safety information on drugs. 1995. Geneva, Council for International Organizations of Medical Sciences.

(42) Jarvinen TL, Sievanen H, Kannus P, Jokihaara J, Khan KM. The true cost of pharmacological disease prevention. BMJ 2011;342:d2175

(43) van Staa TP, Leufkens HG, Zhang B, Smeeth L. A comparison of cost effectiveness using data from randomized trials or actual clinical practice: selective cox-2 inhibitors as an example. PLoS Med 2009 Dec;6(12):e1000194

(44) Eddy DM. The confidence profile method: a Bayesian method for assessing health technologies. Oper Res 1989 Mar;37(2):210-28

(45) Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. Int J Technol Assess Health Care 1993;9(3):440-9

(46) van Staa TP, Smeeth L, Persson I, Parkinson J, Leufkens HG. What is the harm-benefit ratio of Cox-2 inhibitors? Int J Epidemiol 2008 Apr;37(2):405-13

(47) van Staa TP, Cooper C, Barlow D, Leufkens HG. Individualizing the risks and benefits of postmenopausal hormone therapy. Menopause 2008 Mar;15(2):374-81

(48) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332

(49) Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001 Apr 14;357(9263):1191-4

(50) Elm Ev, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-8

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

(52) French S, Xu DL. Comparison study of multi-attribute decision analytic software. J Multi-Crit Decis Anal 2005;13(2-3):65-80

(53) Holden WL, Juhaeri J, Dai W. Benefit-risk analysis: a proposal using quantitative methods. Pharmacoepidemiol Drug Saf 2003 Oct;12(7):611-6

(54) Hammond JS, Keeney RL, Raiffa H. Smart choices. A practical guide to making better life decisions. New York: Broadway Books; 2002.

(55) Mt-Isa S, Tzoulaki I, Callreus T, Micaleff A, Ashby D. Weighing benefit-risk of medicines: concepts and approaches. Drug Discovery Today: Technologies 2011 Apr;In Press, Corrected Proof

(56) Frey P. 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.

(57) Raiffa H. Decision analysis: Introductory lectures on choices under uncertainty. McGraw Hill; 1968.

(58) Clemen RT, Reilly T. Making hard decisions with decision tools suite: Update. Brooks/Cole; 2005.

(59) Goodwin P, Wright G. Decision Analysis for Management Judgment. Chicester: John Wiley & Sons; 2009.

(60) Kirkwood CW. Strategic Decision Making: Multiobjective decision analysis with spreadsheets. Belmont, CA: Duxbury Press; 1997.

(61) Spiegelhalter D, Abrams K, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chicester, West Sussex: John Wiley & Sons Ltd.; 2004.

(62) Phillips LD, Fasolo B, Zafiropoulos N, Beyer A. Is quantitative benefit-risk modelling of drugs desirable or possible? Drug Discovery Today: Technologies 2011;8(1):e3-e10

(63) Thompson JP, Noyes K, Dorsey ER, Schwid SR, Holloway RG. Quantitative risk-benefit analysis of natalizumab. Neurology 2008 Jul 29;71(5):357-64

(64) Belton V, Stewart TJ. Multiple Criteria Decision Analysis: An Integrated Approach. United Kingdom: Springer; 2002.

(65) Mussen F, Salek S, Walker S. Benefit-Risk Appraisal of Medicines. John Wiley & Sons, Ltd.; 2009.

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

(67) Mussen F, Salek S, Walker S. Development and Application of a Benefit-Risk Assessment Model Based on Multi-Criteria Decision Analysis. Benefit-Risk Appraisal of Medicines. John Wiley & Sons, Ltd.; 2009. p. 111-49.

(68) Lim R. Modernizing drug regulatory decisions to manage benefits, risks and uncertainties: Health Canada Perspective. Rockville, MD: Shady Grove Conference Center; 2010.

(69) Lahdelma R, Hokkanen J, Salminen P. SMAA - Stochastic multiobjective acceptability analysis. European Journal of Operational Research 1998 Apr 1;106(1):137-43

(70) Tervonen T, Figueira JR. A survey on stochastic multicriteria acceptability analysis methods. Journal of Multi-Criteria Decision Analysis 2008;15(1-2):1-14

(71) Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

(72) Harrington J. The desirability function. Industrial Quality Control 1965;21(10):494-8

(73) Ouellet D, Werth J, Parekh N, Feltner D, McCarthy B, Lalonde RL. The use of a clinical utility index to compare insomnia compounds: a quantitative basis for benefit-risk assessment. Clin Pharmacol Ther 2009 Mar;85(3):277-82

(74) Trautmann H, Weihs C. On the distribution of the desirability index using Harrington's desirability function. Metrika 2006;63:207-13

(75) Altman DG. Confidence intervals for the number needed to treat. BMJ 1998 Nov 7;317(7168):1309-12

(76) Grieve R, Hutton J, Green C. Selecting methods for the prediction of future events in cost-effectiveness models: a decision-framework and example from the cardiovascular field. Health Policy 2003 Jun;64(3):311-24

(77) Attia J, Page J, Heller RF, Dobson AJ. Impact numbers in health policy decisions. J Epidemiol Community Health 2002 Aug;56(8):600-5

(78) Heller RF, Buchan I, Edwards R, Lyratzopoulos G, McElduff P, Leger SS. Communicating risks at the population level: application of population impact numbers. BMJ 2003 Nov 15;327(7424):1162-5

(79) Heller RF, Edwards R, McElduff P. Implementing guidelines in primary care: can population impact measures help? BMC Public Health 2003 Jan 23;3:7

(80) Heller RF, Dobson AJ, Attia J, Page J. Impact numbers: measures of risk factor impact on the whole population from case-control and cohort studies. J Epidemiol Community Health 2002 Aug;56(8):606-10

(81) Pliskin JS, Shepard DS, Milton CW. Utility Functions for Life Years and Health Status. Operations Research 1980 Jan 1;28(1):206-24

(82) Goldhirsch A, Gelber RD, Simes RJ, Glasziou P, Coates AS. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. J Clin Oncol 1989 Jan;7(1):36-44

(83) Gelber RD, Goldhirsch A, Cole BF, Wieand HS, Schroeder G, Krook JE. A quality-adjusted time without symptoms or toxicity (Q-TWiST) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. J Natl Cancer Inst 1996 Aug 7;88(15):1039-45

(84) Gelber RD, Lenderking WR, Cotton DJ, Cole BF, Fischl MA, Goldhirsch A, et al. Quality-of-Life Evaluation in a Clinical Trial of Zidovudine Therapy in Patients with Mildly Symptomatic HIV Infection. Annals of Internal Medicine 1992 Jun 15;116(12 Part 1):961-6

(85) Schwartz CE, Cole BF, Gelber RD. Measuring Patient-Centered Outcomes in Neurologic Disease: Extending the Q-TWiST Method. Arch Neurol 1995 Aug 1;52(8):754-62

(86) Lynd LD, Najafzadeh M, Colley L, Byrne MF, Willan AR, Sculpher MJ, et al. 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 2010 Jun;13(4):411-7

(87) Lynd LD, Marra CA, Najafzadeh M, Sadatsafavi M. 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. Pharmacoepidemiol Drug Saf 2010 Nov;19(11):1172-80

(88) Minelli C, Abrams KR, Sutton AJ, Cooper NJ. Benefits and harms associated with hormone replacement therapy: clinical decision analysis. BMJ 2004 Feb 14;328(7436):371

(89) Chuang-Stein C, Entsuah R, Pritchett Y. Measures for conducting comparative benefit: Risk assessment. Drug Information Journal 2008;42(3):223-33

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

(91) Payne JT, Loken MK. A survey of the benefits and risks in the practice of radiology. CRC Critical Reviews in Clinical Radiology and Nuclear Medicine 1975;6(3):425-39

(92) Mussen F, Salek S, Walker S. Review of the Current Benefit-Risk Assessment Models. Benefit-Risk Appraisal of Medicines. 1 ed. John Wiley & Sons, Ltd.; 2009. p. 63-97.

(93) Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, CA: Morgan Kaufmann Publishers, Inc.; 1988.

(94) Darwiche A. Bayesian Networks. http://reasoning.cs.ucla.edu/ . 2009.

(95) Jensen FV. An introduction to Bayesian Networks. 1 ed. Secaucus, NJ: Springer-Verlag New York, Inc.; 1996.

(96) Jensen FV, Nielsen TD. Bayesian Networks and Decision Graphs. New York: Springer; 2007.

(97) Druzdzel MJ, van der Gaag LC. Building probabilistic networks: where do the numbers come from? - a guide to the literature. IEEE Transactions on Knowledge and Data Engineering 2000;12:481-6

(98) Lunn DJ, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. Statist Med 2009;28:3049-67

(99) Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics 2004 Jun;60(2):418-26

(100) DuMouchel WH, Harris JE. Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species. Journal of the American Statistical Association 1983 Jun 1;78(382):293-315

(101) Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002 Aug 30;21(16):2313-24

(102) Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet 2007 Jan 20;369(9557):201-7

(103) Carlin BP, Louis T. Bayes and Empirical Bayes methods for data analysis. 2 ed. Chapman & Hall/CRC; 2000.

(104) Edwards D. Introduction to Graphical Modelling. 2 ed. New York: Springer-Verlag; 2000.

(105) Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. 1 ed. MIT Press; 1993.

(106) Lynd LD, O'Brien BJ. Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis. J Clin Epidemiol 2004 Aug;57(8):795-803

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

(108) Eddy DM, Hasselblad V, McGivney W, Hendee W. The Value of Mammography Screening in Women Under Age 50 Years. JAMA: The Journal of the American Medical Association 1988 Mar 11;259(10):1512-9

(109) Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004 Oct 30;23(20):3105-24

(110) Nixon RM, Bansback N, Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. Stat Med 2007 Mar 15;26(6):1237-54

(111) Deal L, Gold BD, Gremse DA, Winter HS, Peters SB, Fraga PD, et al. Age-specific questionnaires distinguish GERD symptom frequency and severity in infants and young children: development and initial validation. J Pediatr Gastroenterol Nutr 2005 Aug;41(2):178-85

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

(113) Sacristán JA, Soto J, Galende I, Hylan TR. A review of methodologies for assessing drug effectiveness and a new proposal: randomized database studies. Clinical Therapeutics 1997;19(6):1510-7

(114) Ryan M, Gerard K, Amaya-Amaya M. Using Discrete Choice Experiments to Value Health and Health Care. Dordrecht, The Netherlands: Springer; 2008.

(115) Pearce D, Özdemiroglu E, Bateman I, Carson RT, Day B, Hanemann M, et al. Economic valuation with stated preference techniques. Summary guide. Norwich: Queen's Printer and Controller for Her Majesty's Stationery Office; 2002. Report No.: 01 SCSG 1158.

(116) Smith RD. Construction of the contingent valuation market in health care:a critical assessment. Health Econ 2003;12(8):609-28

(117) Mitchell RC, Carson RT. Using surveys to value public goods: the contingent valuation method. Washington, D.C.: The Johns Hopkins University Press; 2005.

(118) Havet N, Morelle M, Remonnay R, Carrere MO. 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 2011 Jun 10;1-12

(119) Louviere JJ, Flynn TN, Carson RT. Discrete choice experiments are not conjoint analysis. Journal of Choice Modelling 3[3], 57-72. 2010.

(120) Louviere JJ, Hensher DA, Swait J. Stated choice methods: analysis and application. Cambridge: Cambridge University Press; 2000.

(121) Drummond MF, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the Economic Evaluation of Health Care Programmes. 3 ed. Oxford: Oxford University Press; 2005.

(122) Ryan M, Hughes J. Using Conjoint Analysis to Assess Women's Preferences for Miscarriage Management. Health Econ 1997;6(3):261-73

(123) Ryan M, Bate A, Eastmond CJ, Ludbrook A. Use of discrete choice experiments to elicit preferences. Quality in Health Care 2001 Sep;10:I55-I60

(124) Townsend J, Buxton M, Harper G. Prioritisation of health technology assessment. The PATHS model: methods and case studies. Health Technol Assess 2003;7(20):iii, 1-iii,82

Appendices

A.1 PROTECT WP5 Charter: Benefit-Risk Analysis

The overall objective of WP5 is to assess the relevance of various methodologies for B-R assessment of medicinal products including the provision of usable data and information, the underpinning modelling and the presentation of the results, with a particular emphasis on visualisation methods. The overall plan to achieve this task is outlined below.

Consideration will be given to:

- Submission and post –approval periods, while recognising the relevance of pre - approval B-R assessment
- individual and population- based decision making
- the perspectives of patients, physicians, regulators and other stakeholders such as societal views needed for Health Technology Assessment (HTA)
- possible interdependencies with other PROTECT Work Packages as well as other relevant external initiatives.

Review of methodologies, technologies and representation

The first step will be to review methodologies, technologies and representation that have been proposed or used in analogous contexts for benefit and risk analyses. It will cover:

- analysis and integration of the different sources of evidence of benefits and risks
- elicitation of preference values and uncertainties
- the integration of evidence , preference values and uncertainty
- extensions needed for head to head comparisons of medicines, as needed for HTA
- benefit risk modelling techniques
- visual methods of presentation

Selection of candidate methodologies

We will then decide which methodologies should be taken forward as candidate methodologies for this work package. This task will be achieved by establishing criteria that will include ability to address decision-making for medicines for different stakeholders.

Establishment of criteria and process for selection of case-studies

During the conduct of the aforementioned reviews, a separate working group will establish criteria and processes for selection of case-studies. This work will include ‘initial’, ‘development’ and ‘validation’ case-studies. Case studies will include those that can address patient’s and HTA perspectives in addition to the regulators and prescriber’s one.

Choice and implementation of case studies

Once the criteria and parameters have been established, we will undertake case-studies based on public availability of relevant usable data on both benefits and harms. We will;

- work through the case-studies using relevant methodologies, including the consideration of stakeholder preferences, which may be illustrative at this stage
- compare the outputs of the competing methodologies on the basis of ability to deal with necessary levels of complexity, interpretability and accessibility of outputs and potential for further development for use in benefit-risk decision-making for medicines
- apply methodologies to case-studies, including those that incorporate benefit-risk synthesis over time
- include in outputs, recommendations for methodologies for adoption

Visualisation

We will review and test available technologies as part of the case studies. We will provide recommendations for available software, specially developed software or add-ons for graphical and other visual displays.

Communication

One or more publishable reviews will be written and then case studies will be published in the appropriate literature. In addition, presentations will be made at conferences and within relevant organisations.

A.2 Performance of literature search algorithm

Prior to starting the review we were aware of key research papers that had been judged to satisfy the inclusion criteria. For all three database searches (Web of Science, PubMed, and Scopus), we identified all the key review papers [23;35;53] which gives some validation of the suggested search strategy. The literature search criteria were also ran on CRD databases (including DARE, NHS EED, HTA) but only returned 11 hits, but none were relevant for the scope of our review.

A.3 Criteria for methodology appraisal

We appraised the approaches in four dimensions: within each dimension, specific criteria as listed below were addressed and discussed where applicable:

(1) Fundamental principle
- (a) 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
- (b) 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.
- (c) 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.
- (d) 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.
- (e) 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.
- (f) Does the approach appropriately incorporate value judgements, either explicitly or implicitly? Stakeholders' involvement in providing preference value is needed to satisfy this criterion.
- (g) 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.

(2) Features of respective approaches
- (a) 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.
- (b) 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.
- (c) 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.
- (d) 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.
- (e) Can the method incorporate time dimension? We will describe how time variables are dealt with.

- (f) 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.
- (g) 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.

(3) Visual representation of model

- (a) Does the model propose potential visualisations of the results? We will describe the proposed visualisation techniques and what are they intended to represent.

(4) Assessability and accessibility

- (a) 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.
- (b) 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.
- (c) 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.
- (d) In what respect the use of the approach can lead to make better decision-making?

A.4 Working definitions for the classifications of benefit-risk approaches

The working definitions that are later used in the classifications of the approaches are described in (1) to (4) below:

(1) Benefit-risk assessment frameworks

(a) Descriptive frameworks (non-quantitative)

These are guidelines to conducting benefit-risk assessment which consist of step-by-step guide to follow for good decision-making practice and to increase transparency. Descriptive frameworks are usually general, and most of the times reiterate common sense.

(b) Quantitative frameworks

These approaches provide descriptive step-by-step guidelines for the purpose of good decision-making practice and to increase transparency in combination with quantitative methods of trading risks and benefits based on grounded mathematical principle. They depend on the availability of consistent data that allow quantifications and comparisons of different options (treatments) to be made.

(2) Metric indices for benefit-risk assessment

(a) Quantitative threshold indices

This is a group of general indices that are used in benefit-risk assessments to quantify benefits and risks. These indices are derived from mathematical and statistical manipulations of probabilities and/or utilities, and are generally used as thresholds (cut-points) when assessing benefit-risk balance or in deciding best treatment options. These indices are either not designed to or are not formally used to trade-off benefits and risks.

(b) Quantitative trade-off indices

Another set of metric indices take the idea further by proposing a formal method to trade off benefits and risks. Naturally, there is a fine line between metric indices in this group with those in (2)(a). However, as opposed to the threshold metrics in (2)(a), quantitative trade-off indices integrate benefits and risks into a single metric index which represent the value of the trade-off.

(c) Quantitative health utility indices

A particular group of specialist trade-off metric indices are the health utility indices. Health utility indices are established indices which are specific to health-related outcomes and usually have been validated internally and externally for use under certain conditions.

(3) Estimation techniques for benefit-risk modelling

Estimation techniques deal with how benefit-risk parameters/inputs are processed or analysed. They bring together the evidence of benefits and risks into a benefit-risk model, where commonly trade-offs are estimated. The inferences on benefit-risk trade-off are made on the resultant metrics having considered the evidence, data and assumptions. The choice of metrics to be used is not fixed therefore any sensible one from (2) above can be used. They may also be applied within benefit-risk frameworks defined in (1).

(4) Utility survey techniques for preference elicitation

These are methods that focus on the design and conduct of data (for example utilities and preference values) elicitation or collection process. They do not formally provide any specific way on how these data are to be used in benefit-risk assessment.

A.5 Descriptive frameworks

Authors: Nan Wang, Lawrence D. Phillips, and Shahrul Mt-Isa

A.5.1 PrOACT-URL

Description

PrOACT-URL is a generic framework which provides a generic problem structure to be considered when facing a decision problem [2]. The acronym PrOACT-URL represents the steps of this framework: (1) determine the decision context and frame the ***Pr****oblems*; (2) establish ***O****bjectives* and identify criteria; (3) identify options and ***A****lternatives*; (4) evaluate the expected ***C****onsequences* of the options for each criterion; (5) assess the ***T****rade-offs* of benefit and risk; (6) report the ***U****ncertainty* in benefit and risk, and assess the impact of uncertainty on B-R balance; (7) judge the relative importance and the ***R****isk* attitude of the decision maker and assess how this affect the B-R balance; and (8) consider the decision's consistency with other ***L****inked decisions*, both in the past and its impact on future decisions.

We suggest reading Hammond (2002) [54] as an introductory and core reference, and Hunink (2001) [13] for worked example.

Evaluation

This is a descriptive framework covering the important aspects of when structuring a decision problem. Because the framework is very generic, and in other words as described previously, a generic framework for good practice in framing decision problems. Phillips (2010) used PrOACT-URL as the basis of determining whether a benefit-risk approach is comprehensive enough for decision-making [1]. PrOACT-URL strongly emphasise uncertainties in input values and value judgments as well as the importance of sensitivity analyses. PROTECT WP5 work stream D uses this framework to identify and generate the requirements for benefit-risk assessment when preparing case studies. In the light of this usefulness, PrOACT-URL is a suitable candidate to be taken forward as a benefit-risk methodology recommendation from this review.

We present an improvised framework based on PROACT-URL (Table 19). The framework extends the work begun in the WP2 report from the EMA's Benefit-Risk Project, and as such could be useful as a 12-step protocol for modelling benefit-risk of medicinal products, and has been proposed in PROTECT WP5 as means to prepare case studies.

The first two columns are generic. The third column has been completed to show what data sources are indicated, here with reference to rimonabant (Acomplia). The key document for measurable data is the EPAR, which can be accessed from the EMA public website by following the Human Medicines link. However, incorporating the judgements necessary to make the benefit-risk balance explicit requires information from stakeholders, key players and decision makers: drug developers, regulators, health technology assessors, prescribers, patients. Here, the types of judgements are indicated.

Table 19 PrOACT-URL (adapted for decision-making for drug benefit-risk assessment)

STEP	DESCRIPTION	INFORMATION SOURCES
PrOBLEM 1. Determine the nature of the problem and its context. 2. Frame the problem.	1a. The medicinal product (e.g., new or marketed chemical or biological entity, device, generic). 1b. Indication(s) for use. 1c. The therapeutic area and disease epidemiology 1d. The unmet medical need, severity of condition, affected population, patients' and physicians' concerns, time frame for health outcomes. 1e. The decision problem (what is to be decided and by whom, e.g., industry, regulator, prescriber, patient) 2a. Whether this is mainly a problem of uncertainty, or of multiple conflicting objectives, or some combination of the two, or something else (e.g., health states' time progression). 2b. The factors to be considered in solving the problem (e.g., study design, sources and adequacy of data, disease epidemiology, presence of alternative treatments).	1. "Acomplia: EPAR-Scientific Discussion" (on EMA website, under Acomplia, Assessment history, Initial Marketing authorisation documents). See 1. Introduction Indication for weight loss, approved June 2006. Withdrawn January 2009. See "Assessment Report for Acomplia, Procedure No. EMEA/H/C/000666/A20/0012" Report No. EMEA/65105/2009 (on website, under Acomplia, Assessment History, Changes since initial authorisation of medicine) 2a. Usually it is a mixture of favourable effect size, unfavourable effect seriousness and their uncertainties, but this may be more revealed in the EPAR than explicitly spelled out. 2b. Ideally, only factors that make a difference to a decision need be included.
OBJECTIVES 3. Establish objectives that indicate the overall purposes to be achieved. 4. Identify criteria for a) favourable effects b) unfavourable effects	3. The aim (e.g., to evaluate the benefit-risk balance, to determine what additional information is required, to assess change in the benefit-risk balance, to recommend restrictions). 4. A full set of criteria covering the favourable and unfavourable effects (e.g., endpoints, relevant health states, clinical outcomes). An operational definition for each criterion along with a measurement scale with two points defined to encompass the range of performance of the alternatives (not just reported measures of central tendency,	3. EPAR: 1. Introduction 4a. EPAR: 4. Clinical aspects especially Table 6 & Figure 1 4b. EPAR: Table 15 & associated text Establishing two points on each measurement criterion facilitates scaling of the alternatives. Usually, data are reported only for the alternatives considered, but quantitative modelling requires definitions of two points on each measurement scale: e.g., lowest and highest practically-

	but also confidence intervals). Considerations of the clinical relevance of the criteria—some are of more concern to decision makers than others.	realisable measures. Quantitative weights assigned to the scales are based on considerations of relevance, which may not be documented, in which case the relevant stakeholders or key players can provide the information.
ALTERNATIVES 5. Identify the options to be evaluated against the criteria.	5a. Pre-approval: dosage, timing of treatment, drug vs. placebo and/or active comparator; the decision or recommendation required (e.g., approve/disapprove, restrict, withdraw). 5b. Post-approval: do nothing, limit duration, restrict indication, suspend.	5. As above, Step 1. Provide a clear definition of each option.
CONSEQUENCES 6. Describe how the alternatives perform for each of the criteria, i.e., the magnitudes of all effects, and their desirability or severity, and the incidence of all effects.	6. The consequences separately for each alternative on each criterion (e.g., efficacy and safety effects that are clinically relevant, positive and negative health outcomes), summarised in an 'Effects Table' with alternatives in columns and criteria in rows. Qualitative and quantitative descriptions of the effects in each cell, including statistical summaries with confidence intervals, and references to source data, graphs and plots.	6. As above for Steps 3 and 4. It is rare to see all this information in one place. Usually, it is necessary to search for the information. If more than one study is reported, are decisions to be based on a single 'best' study or on combined data? Is a meta-analysis available? Can the effects table be populated with the results from several studies? Head-to-head comparisons are not necessarily needed for quantitative modelling. Report missing data. A quantitative model will require judgements of value functions, which express the clinical relevance of the data.
TRADE-OFFS 7. Assess the balance between favourable and unfavourable effects.	7. The judgement about the benefit-risk balance, and the rationale for the judgement.	7. EPAR: 6. Overall conclusions, benefit/risk assessment and recommendations. A quantitative model will also require judgements of weights associated with the criteria.
At this point, only issues concerning the favourable and unfavourable effects, and their balance, have been considered. The next three steps are relevant in considering how the benefit-risk balance is affected by taking account of uncertainties.		
UNCERTAINTY 8. Report the uncertainty associated with the favourable and unfavourable effects.	8. The basis for and extent of uncertainty in addition to statistical probabilities (e.g., possible biases in the data,	8. EPAR: Tables 6 & Figure 1, and Table 15 for confidence intervals. Also "Discussion on clinical safety" under Table 16,

9. Consider how the balance between favourable and unfavourable effects is affected by uncertainty.	soundness and representativeness of the clinical trials, potential for unobserved adverse effects) 9. The extent to which the benefit-risk balance in step 7 is reduced by considering all sources of uncertainty, to provide a benefit-risk balance, and the reasons for the reduction.	and Table 17, and Overall conclusions (as at Step 7, above). Incidence data, reported at step 6 in the effects table, provide information relevant to the probabilities of realising the effects. 9. No data for this. EPARs from late 2010, should provide information on this step. Judgement plays a key role in this step. A quantitative model will explore in sensitivity analyses and scenario analyses (or by explicitly incorporating probability distributions in the model) the effects on the overall benefit-risk balance of all sources of uncertainty.
RISK TOLERANCE 10. Judge the relative importance of the decision maker's risk attitude for this product. 11. Report how this affected the balance reported in step 9.	10. Any considerations that could or should affect the decision maker's attitude toward risk for this product (e.g., orphan drug status, special population, unmet medical need, risk management plan). 11. The basis for the decision maker's decision as to how tolerable the benefit-risk balance is judged to be (taking into account stakeholders' views of risk?).	10. No data for this Some idea of the risk tolerance can be inferred from any report of step 9—how the favourable-unfavourable effects balance was affected by uncertainty. Another key role for judgement. 11. See Step 1 document about the decision to withdraw Acomplia. In this case, it was new data obtained post-approval, as reported in sketchy form by EMEA/H/C/000666/A20/0012.
LINKED DECISIONS 12. Consider the consistency of this decision with similar past decisions, and assess whether taking this decision could impact future decisions.	12. How this decision, and the value judgements and data on which it is based, might set a precedent or make similar decisions in the future easier or more difficult.	12. See EPAR Conclusions. As all decisions are based not only on evidence, but also interpretations of that evidence that invoke value judgements and beliefs about uncertainty, decision makers may wish to reflect on whether those judgements and beliefs are consistent across similar past decisions, allow future changes and can be defended.

Revision 5 (23 June 2011). Lawrence D. Phillips & Work Group members

Disclaimer: Table 19 is an adaptation of the original PrOACT-URL for PROTECT and is a working document. Please forward suggested additions and improvements derived from its use to lawrence.phillips@ema.europa.eu.

A.5.2 Ashby and Smith Framework (ASF)

Description

Ashby and Smith framework [3] provides analysts or decision makers a procedure to follow when facing a medical decision problem. The framework is based on evidence and reiterates decision making as Bayesian in nature. ASF is similar in spirit to PrOACT-URL where the focus is to structure and clarify decision problems. This involves breaking down a problem into elements in order to aid future benefit-risk assessment. The elements to be defined are: (1) the decision-maker: who is making the decision and for whom the decision is to be made; (2) the possible actions: these are the available options in a particular decision problem; (3) the uncertain consequences: all possible outcomes (as exclusive events) by each action; (4) the sources of evidence: available data to support any inference about action-consequence pairs are to be specified; (5) the utility assessments: preference values that are required in the benefit-risk assessment are to be made explicit. In the paper [3] it was shown that option with maximum expected utility is favourable.

We suggest reading Ashby (2000) [3] as an introductory and core reference, and Mt-Isa (2011) [55] for worked example.

Evaluation

This framework is both a descriptive framework and a modelling proposal. It combines the objective aspect (chances of consequences) and the subjective aspect (preferences of consequences) of the problem in a clear and explicit way based on available evidence. This framework is suitable for any decision makers, and especially suitable for physicians and patients in treatment selection as the action-consequences arguments address directly to their concerns. Table 20 provides the steps and description for ASF.

Table 20 Ashby and Smith Framework

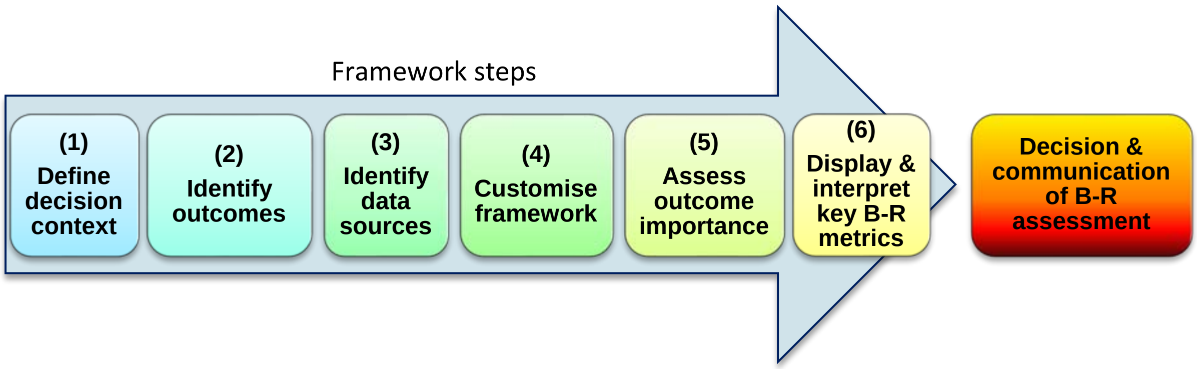
STEP	DESCRIPTION
1. Decision-maker	Identify who is making the decision and for whom the decision is to be made
2. Possible actions	List down the options or alternatives available
3. Uncertain consequences	List down all uncertain consequences following each action, for example the any benefit or risks that may be observed
4. Sources of evidence	Identify and obtain suitable sources of evidence regarding the decision problems
5. Utility assessments	The type of preference values required and where they should come from are to be made explicit

A.5.3 PhRMA BRAT framework

Description

The framework is developed by Pharmaceutical Research and Manufacturers of America (PhRMA) benefit-risk action team (BRAT), which aims to guide decision-makers in selecting, organizing, understanding and summarising the evidence relevant to benefit-risk decisions [5;6]. The key steps are illustrated in Figure 3.

Figure 3 Steps in the BRAT benefit-risk assessment framework (reproduced from [55])


Framework steps
(1) Define decision context
(2) Identify outcomes
(3) Identify data sources
(4) Customise framework
(5) Assess outcome importance
(6) Display & interpret key B-R metrics
Decision & communication of B-R assessment

Adapted by permission from Macmillan Publishers Ltd: Clinical Pharmacology & Therapeutics [5], copyright (2011)

We suggest reading Coplan (2011) [5] as an introductory and core reference, and Levitan (2011) [6] for worked example.

Evaluation

The BRAT framework emphasises the value tree (criteria tree) build-up, data selection and data preparation. It focuses on the comparison of new drug and a comparator using a key benefit-risk summary table and a forest plot as illustrated by an example shown in Figure 4.

Benefits and risks are not integrated in this framework, but are assessed separately. This was a conscious proposal to avoid synthesising data into complex statistical models which may not be easily understood by readers. It is more of a framework for pharmaceutical companies to collect all available and relevant pieces of evidence of a new drug for the communications with regulatory authorities. The use of such framework can increase the transparency, the predictability and the consistency with which benefit-risk assessments are conducted. The tabular output without further condensation delivers benefit-risk information to patients, healthcare professionals and regulators as a basis for own decisions based on individual preferences. The parameters and results are acceptable and are easily interpreted by readers without or with little statistical expertise. It is therefore an easy-to-implement tool to structure simple decision problems on daily basis. However, its recommendation to use odds ratios to summarise benefits and risks is somewhat controversial in comparative benefit-risk assessment. The updated BRAT tool also allows users to use risks differences as the summary statistics.

We present the steps involved in PhRMA BRAT Framework and their descriptions in Table 21 for quick reference.

Figure 4 An example of benefit-risk summary table and the associated forest plot – Natalizumab case study

		Outcome	Natalizumab Risk / 1000 pts	Placebo 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%)	1	0	1	(-, -)
	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 Natalizumab 🟨
Higher for Placebo 🟦

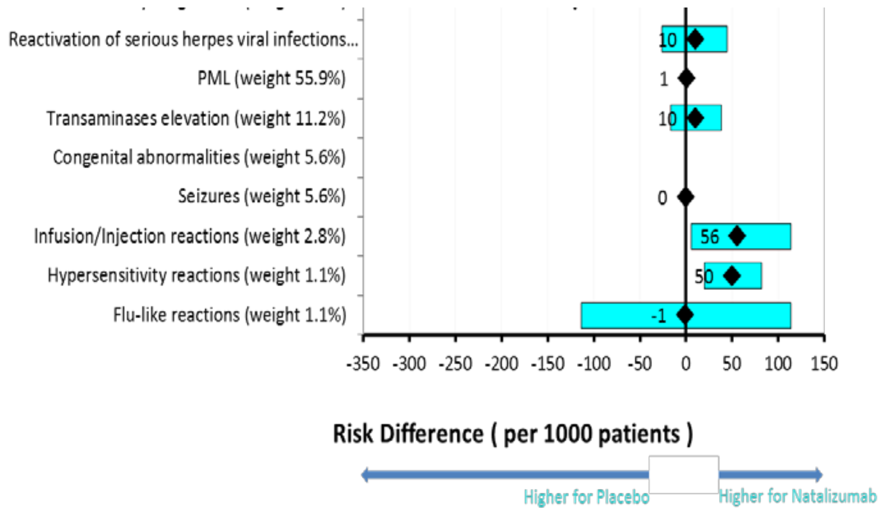


Table 21 The PhRMA BRAT Framework [6]

Step	Description
1. Define the decision context	Define drug, dose, formulation, indication, patient population, comparator(s), time horizon for outcomes, perspective of the decision makers (regulator, sponsor, patient, or physician)
2. Identify outcomes	Select all important outcomes and create the initial value tree. Define a preliminary set of outcome measures/endpoints for each. Document rationale for outcomes included/excluded
3. Identify and extract source data	Determine and document all data sources (e.g. clinical trials, observational studies). Extract all relevant data for the data source table, including detailed references and any annotations, to help the subsequent interpretations create summary measures
4. Customise the framework	Modify the value tree on the basis of further review of the data and clinical expertise. Refine the outcome measures/endpoints. May include tuning of outcomes not considered relevant to a particular benefit-risk assessment or that vary in relevance by stakeholder group
5. Assess outcome importance	Apply or assess any ranking or weighting of outcome importance to decision makers or other stakeholders
6. Display and interpret key benefit-risk metrics	Summarise source data in tabular and graphical displays to aid review and interpretation. Challenge summary metrics, review source data, and identify and fill any information gaps. Interpret summary information.

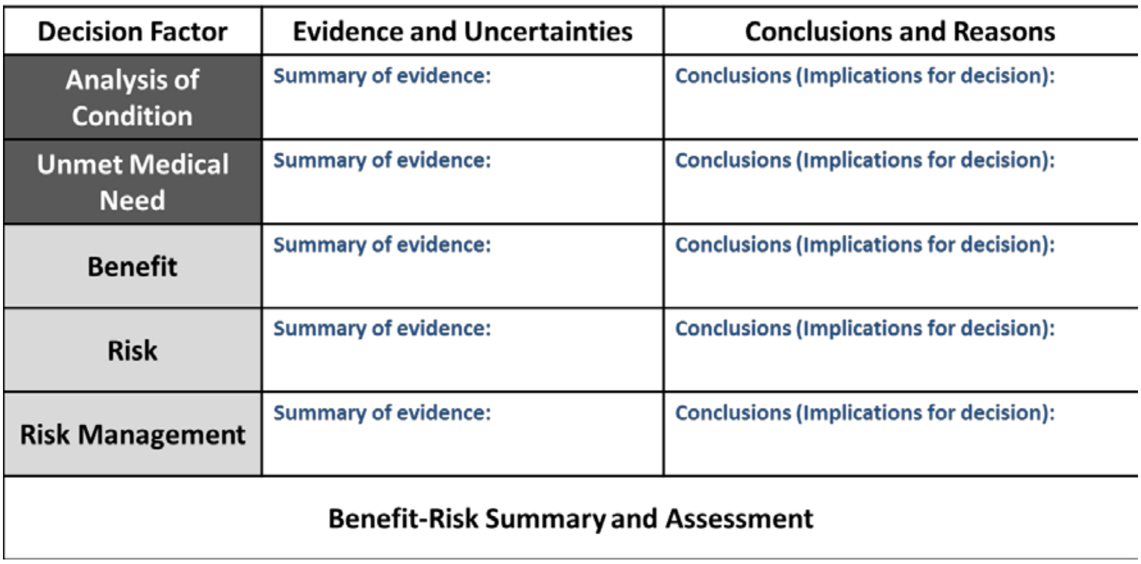
A.5.4 Other frameworks still under development

Description

We are aware of four frameworks that are still under development at this time. These are the FDA Center for Drug Evaluation and Research Benefit Risk Framework (BRF) framework, the CMR CASS framework which later evolved into the Consortium on Benefit-Risk Assessment (COBRA), the Southeast Asia Benefit-Risk Evaluation (SABRE) Initiative and the Unified Methodologies for Benefit-Risk Assessment (UMBRA) Initiative. The details of the frameworks are not yet available in full, so we briefly discuss them here.

The FDA BRF addresses the issues relevant to regulatory decision making [7]. FDA BRF aims to aid communication within and outside the FDA by approaching benefit-risk decision making through a set of guidelines that are described as simple and user-friendly; can address critical issues; are capturing expert views faithfully; can represent issues transparently; are compatible with quantitative analysis of clinical benefit and risk; can facilitate communications; and are broadly applicable. FDA BRF would "tell the story" of a decision problem through the framework. A proposal for the framework is shown in Figure 5.

Figure 5 An improved framework being considered in FDA BRF (reproduced from [56])

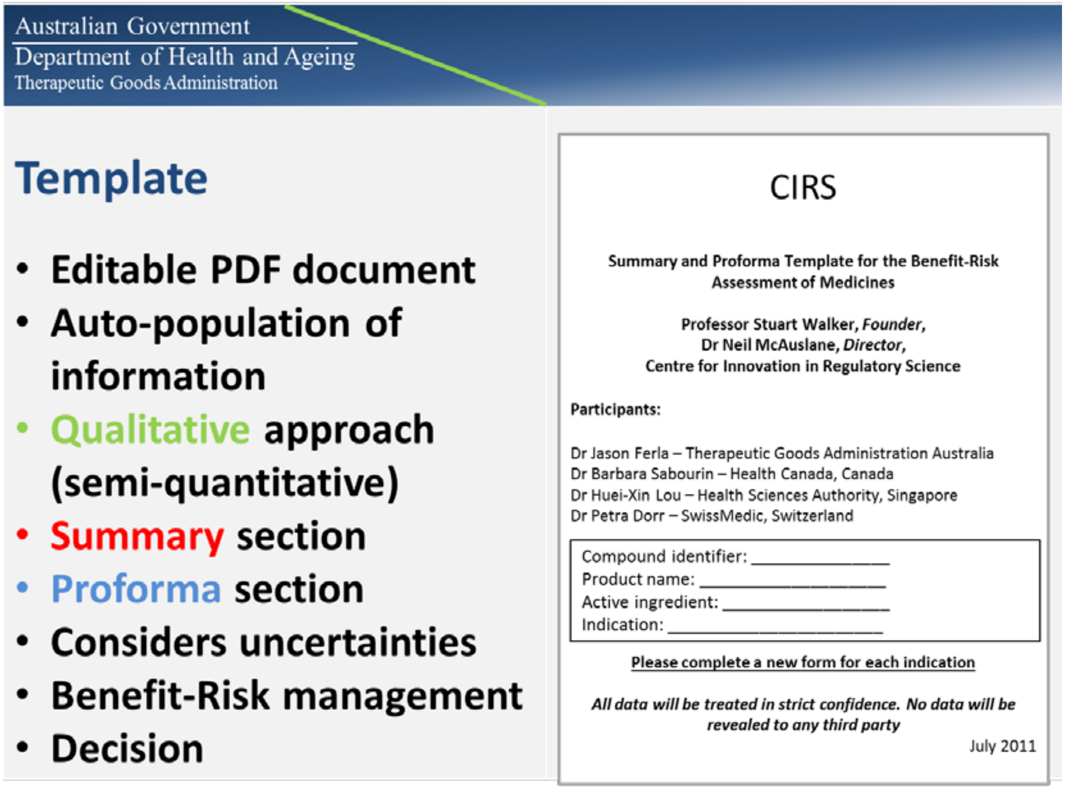


Decision Factor	Evidence and Uncertainties	Conclusions and Reasons
Analysis of Condition	Summary of evidence:	Conclusions (Implications for decision):
Unmet Medical Need	Summary of evidence:	Conclusions (Implications for decision):
Benefit	Summary of evidence:	Conclusions (Implications for decision):
Risk	Summary of evidence:	Conclusions (Implications for decision):
Risk Management	Summary of evidence:	Conclusions (Implications for decision):
Benefit-Risk Summary and Assessment		

We suggest reading Frey (2012) [56] and Jenkins (2010) [7] for more details on FDA benefit-risk framework.

The CMR CASS framework was initially described in a paper by Stuart Walker in 2008 [8]. The important points being considered and tested within CMR CASS are: (1) a universal framework for every parties is the target and a quantitative BR model is the ultimate goal; (2) the changes in benefit risk balance need to be accommodated and benefit-risk assessment should be revisited through the product life-cycle; (3) the development of the framework needs the involvement of a wide range of stakeholders; (4) the challenges in assessing benefit and risk in post-approval phase are to be tackled [8]. It is initially intended for small regulatory agencies (Canada, Australia, Switzerland, Singapore → CASS) and is also known as the 4-Agency Consortium. The CMR CASS framework later developed into COBRA coinciding with the establishment of the UMBRA initiative. COBRA developed a framework "proforma" which was tested in a retrospective study (Figure 6), with plans to further improve the template to reflect more of the actual practice, to integrate visualisations of data and to initiate a prospective study [10]. CMR has also been succeeded by the Centre for Innovation in Regulatory Science (CIRS) which is guided by the same objectives (http://cirsci.org).

Figure 6 Template for COBRA (reproduced from [10])



Australian Government
Department of Health and Ageing
Therapeutic Goods Administration

Template

- Editable PDF document
- Auto-population of information
- Qualitative approach (semi-quantitative)
- Summary section
- Proforma section
- Considers uncertainties
- Benefit-Risk management
- Decision

CIRS

Summary and Proforma Template for the Benefit-Risk Assessment of Medicines

Professor Stuart Walker, *Founder*,
Dr Neil McAuslane, *Director*,
Centre for Innovation in Regulatory Science

Participants:

Dr Jason Ferla – Therapeutic Goods Administration Australia
Dr Barbara Sabourin – Health Canada, Canada
Dr Huei-Xin Lou – Health Sciences Authority, Singapore
Dr Petra Dorr – SwissMedic, Switzerland

Compound identifier: _______________
Product name: ___________________
Active ingredient: _________________
Indication: ________________________

Please complete a new form for each indication

All data will be treated in strict confidence. No data will be revealed to any third party

July 2011

We suggest reading Walker (2009) [8] and the CIRS Workshop Synopsis on benefit-risk (2012) [10] for more details on CMR CASS and COBRA.

SABRE is another recent initiative in Southeast Asia to promote better assessment of benefits and risks of medicine, but has not yet released any details to date.

UMBRA is the leading initiative to bring together the expertise such as COBRA, PhRMA BRAT (Section A.5.3) and SABRE in order to establish a unified framework containing common elements across different methodologies for benefit-risk assessment. The current framework model currently proposed by UMBRA contains eight steps in four stages:

(1) Stage 1 – Framing the decision
  (a) Step 1: Decision context
(2) Stage 2 – Identifying benefits and risks
  (a) Step 2: Building value tree
  (b) Step 3: Refining the value tree
(3) Stage 3 – Assessing benefits and risks
  (a) Step 4: Relative importance of benefits and risks
  (b) Step 5: Evaluating the options
(4) Stage 4 – Interpretation and recommendations
  (a) Evaluating uncertainty
  (b) Concise presentation of results (visualisation)
  (c) Expert judgement and communication

The Further details on the UMBRA initiative can be found on http://cirsci.org/UMBRA.

Evaluation

The current version of FDA BRF focusses on structured qualitative approach that provides a high-level snapshot and the concise bottom-line descriptions of the relevant issues to the regulatory decision in words [56]. The rows in Figure 5 are designed to “tell the story” of the regulatory decision by asking five relevant questions: “what is the problem?”, “what other potential interventions exist?”, “what is the benefit of the proposed intervention?”, “what am I worried about?” and “what can I do to mitigate/monitor those concerns?” [56].

CMR CASS’s ultimate goal was to develop an approach which is universal, quantitative, lifecycle covered, and flexible with various stakeholders. However, its successor COBRA takes a more qualitative approach (semi-quantitative) to benefit-risk assessment. UMBRA works actively with COBRA, PhRMA BRAT and SABRE to align common concepts and requirements for benefit-risk assessment but the details on this ongoing work have not been published yet. Therefore we are unable to comment further on these frameworks. We understand that these frameworks are being tested for their feasibility by several working groups and being improved with each iteration.

A.6 Quantitative frameworks

Authors: Shahrul Mt-Isa, Nan Wang, Sinan B. Sarac, and Lawrence D. Phillips

A.6.1 Benefit-Less-Risk Analysis (BLRA)

Description

BLRA [11] is a relatively recent multi-criteria analysis approach but it has been suggested that it is a “reinvention of MCDA without knowing of its existence” [1]. It is true that BLRA has many aspects in common with MCDA, but BLRA provide more definitive framework around the issues of adverse events experienced by individual patients (in clinical trials). BLRA is an extension of previous work on GBR (see Appendix A.9.3) [36]. BLRA provides both a measure and a framework. The BLRA measure is defined as $\sum(\text{benefits}) - f \times \sum(\text{risks})$, where f is a proportionality constant to place risks on the same scale as benefit for direct assessment.

We suggest reading Chuang-Stein (1994) [11] for more details and worked example..

Evaluation

- Principle

The BLRA framework helps to increase transparency in decision-making through a seven-step process: (1) organise safety data into body functions; (2) score the risk in each class; (3) build the risk component; (4) define benefit outcome; (5) discount the benefit; (6) compare risk-adjusted benefit measures for options; and (7) run sensitivity analysis. Uncertainties in input parameters are dealt with similar to MCDA (Appendix A.6.4). BLRA also requires explicit value judgments for the model. Although the proposed BLRA is estimated without uncertainties in the resultant measure, and is generally based on its point estimates, it is conceptually straightforward to extend this, for example by fitting in a Bayesian framework to take advantage of Bayesian analysis which could then provide uncertainty estimates in the final BLRA measures.

- Features

Multiple benefits and risks can be estimated simultaneously to produce an integrated BLRA measure. Multiple sources of evidence are dealt with as different criteria similar to MCDA (Appendix A.6.4). The standard BLRA approach only deals with two options at a time but can be easily extended to account for more than two options through statistical modelling. This is possible because BLRA is estimated at individual level therefore standard regression techniques can be used to compare BLRA for different options. The primary sensitivity analysis concentrates on varying the choice of proportionality constant f.

- Visualisation

There is none proposed.

- Assessability and accessibility

Many aspects of BLRA are similar to those of MCDA. However, to date, there is no specialist software to perform BLRA benefit-risk assessment. The emphasis on individuals’ value judgments in BLRA naturally demonstrates its capability for individual decision-making. This does not hinder the capabilities of BLRA for use in regulatory decision-making; although for simple decisions, as MCDA, the application of BLRA may also be too cumbersome.

A.6.2 Net Clinical Benefit (NCB)

Description

Net clinical benefit is a conceptually simple approach to balancing benefits and risks for decision-making. NCB is a B-R framework with a metric also called the NCB, albeit implicitly expressed in the primary paper [12]. The primary application of NCB is given in a Bayesian framework which contributes to its strength. NCB is expressed as sum of benefits less the sum of risks, which would commonly be on odds ratio or relative risk scale. Alternatively, established health indices e.g. QALY (see Section A.8.1) can also be used.

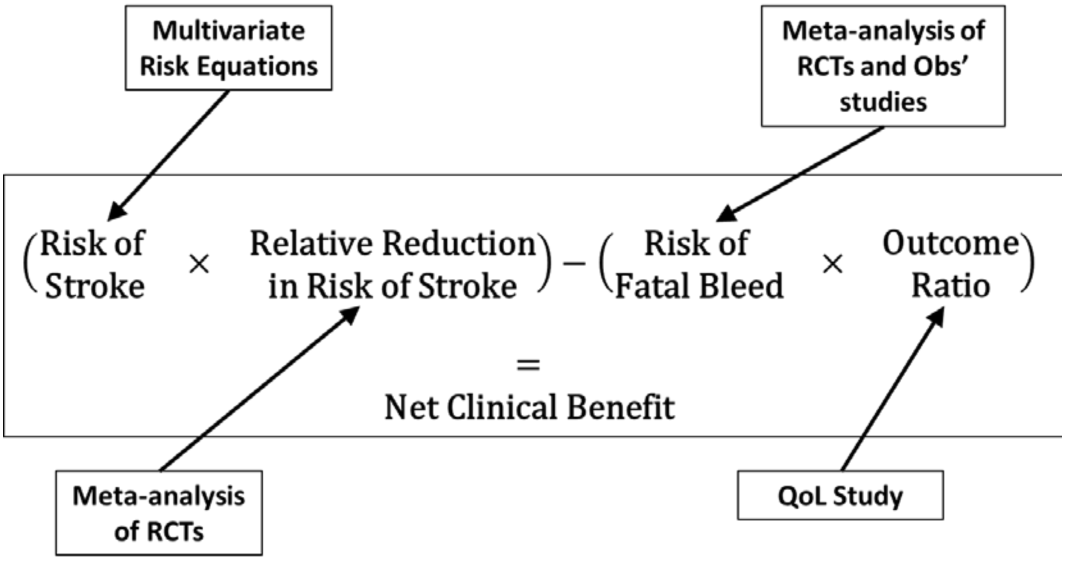
We suggest reading Sutton (2005) [12] for more details and worked example.

Evaluation

- Principle

Transparency is increased through the application of the NCB framework which is divided into three steps [12;12]: (1) define decision problem and data sources; (2) establish the functional form of NCB equation; and (3) estimate the NCB. An example of a functional form for a model [12] is illustrated in Figure 7.

Figure 7 Schematic illustration of the sources of evidence synthesised in the net clinical benefit model (reproduced from [12])


Multivariate Risk Equations
Meta-analysis of RCTs and Obs' studies
(Risk of Stroke × Relative Reduction in Risk of Stroke) − (Risk of Fatal Bleed × Outcome Ratio)
=
Net Clinical Benefit
Meta-analysis of RCTs
QoL Study

The final step (3) is divided into smaller pieces to provide transparency in the estimation of NCB, where fuller details can be found in Sutton's paper with a worked example [12]. NCB inherits the strengths of Bayesian modelling where statistical uncertainties in various parameters are naturally taken into account and the impact of the observations with these uncertainties is propagated into the end results in the form of posterior distributions.

- Features

Many features of NCB are similar to BLRA (see Appendix A.6.1). Sutton (2005) suggested that sensitivity analysis is performed on individual parameters in the NCB equation whilst acknowledging full sensitivity analysis is important but impractical [12]. Being estimated in Bayesian framework, NCB models naturally allow formal updating of results based on current knowledge and new data. An appealing feature of NCB is that it can be easily extended to derive an NNT measure, therefore aiding communicability of the results to a more general audience. NCB can also be extended to estimate the minimum event rate for which treatment is favourable (MERT), an idea similar to MCE (see Appendix A.7.4).

- Visualisation

A graphical representation of NCB model is shown in Figure 8 as a threshold plot where it illustrates that benefit only occurs when risk is above a certain threshold. Figure 9 illustrates the NCB as a function of patient risk. Additionally, the probability of benefit being greater than zero (or above certain cut-point) is represented by posterior distributions through simulation (Figure 10).

Figure 8 Graphical representation of a net clinical benefit model (reproduced from [12])

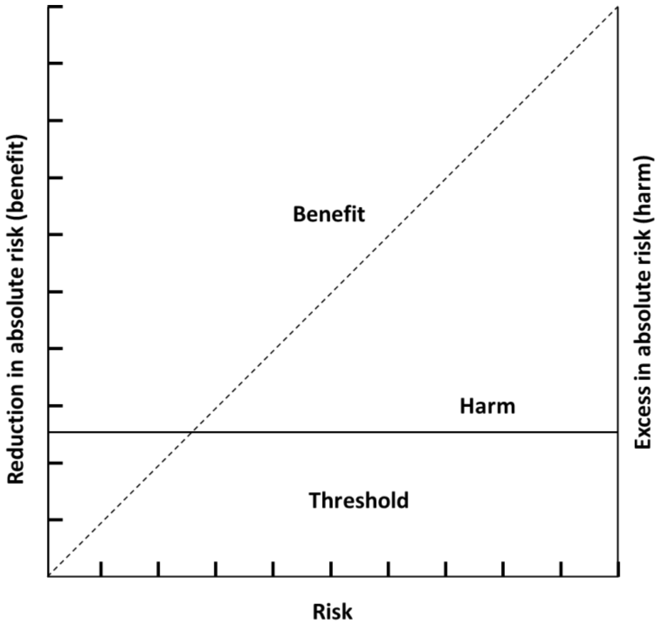


Figure 9 Median and 95% credible intervals net clinical benefit expressed as a function of patient risk (reproduced from [12])

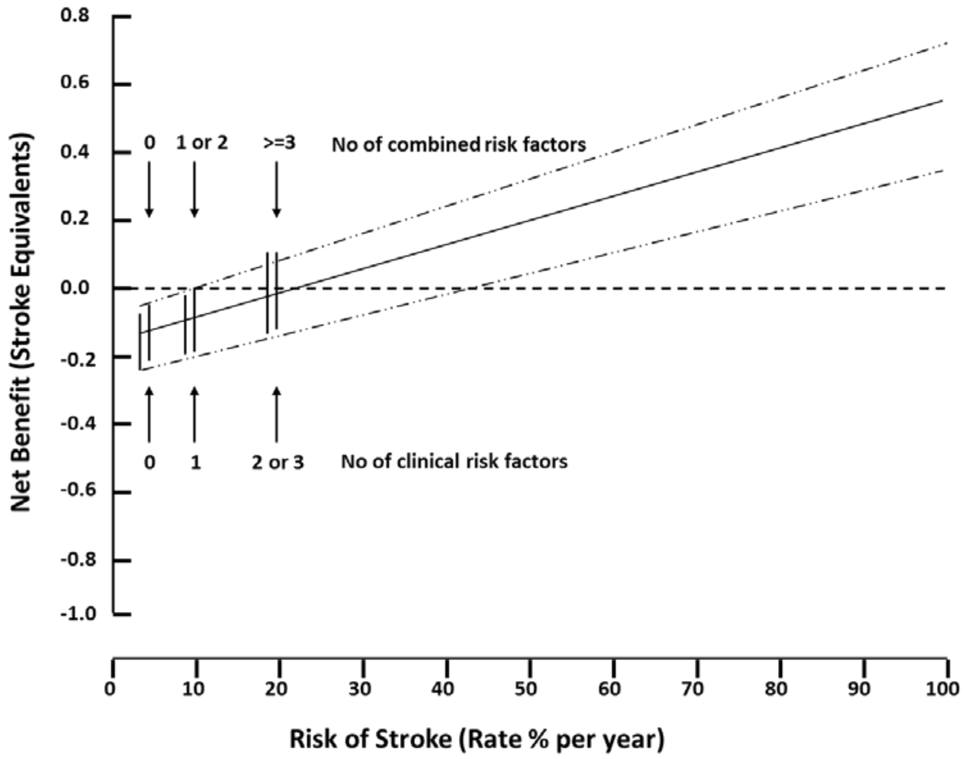
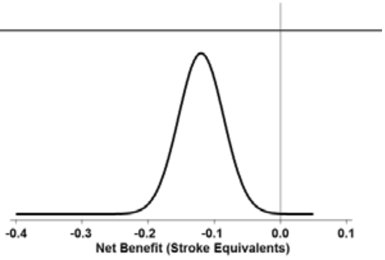
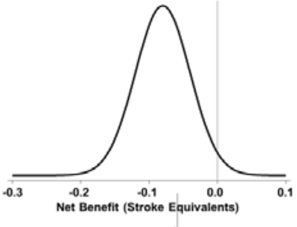
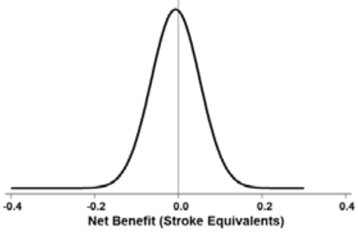


Figure 10 Simulated posterior distribution plots of net clinical benefit for different levels of risk of stroke (reproduced from [12])

Number of risk factors	*Median (95% CrI)*	*Probability of Benefit >0*	*Simulated Posterior Distribution*
Clinical risk factors			
0	-0.12 (-0.21 to -0.05)	0.04%	
1	-0.08 (-0.18 to -0.001)	1.39%	
2 or 3	-0.007 (-0.12 to 0.11)	44.74%	

- Assessability and accessibility

There are a number of resultant metrics from NCB analysis: NCB, NNT, and MERT. This increases the interpretability of the measures when they are being communicated to different stakeholders. The acceptability of NCB would depend on the chosen functional form and its underlying assumptions, but are very similar to BLRA. Acceptability of NNT and MERT is subjected to the same scrutiny as the classical NNT approach (see Section A.7.1). Complex Bayesian models have always been associated with long computing time, a feature that NCB inherits. The complexity of the model and parameterisations can increase analysis time of an NCB model therefore it may not be suitable for real-time decision-making. Nevertheless, disregarding the analysis time factor, NCB is a potential contender for transparent decision-making required by regulatory agencies.

A.6.3 Decision trees

Description

A decision tree is a diagram, like a tree laid on its side, with decisions as roots, uncertain events with their outcomes, and further decisions and outcomes as branches, with consequences at the ends of the branches [57]. These act-event sequences represent the unfolding nature of difficult problems as the outcomes become known to the decision maker. Just two quantitative assessments are required: utilities that characterise the judged value of the consequences and probabilities that the consequences will occur. Repeated application of the expected utility rule, multiplying utilities by their probabilities and adding the products of the branches at each node of the tree ('averaging out'), then choosing the decision with the highest expected utility ('folding back'), provides a guide to action. Decision trees are not domain specific, so they can represent a wide variety of decisions, including medical decision making [13]. Many textbooks provide guidance on how to apply this approach to real-life decisions [58-60].

We suggest reading Spiegelhalter (2004) [61] for introduction, Raiffa (1968) [57] as core reference, and Hunink (2001) [13] for worked example.

Evaluation

- Principle

Decision trees are easily-visualised graphical representations of the expected utility rule. It is possible with decision tree software to show the expected resources and rewards at each branch of the tree, with the final net change displayed as the overall consequences at the end of each path through the tree. The algorithm for 'averaging out' and 'folding back' the decision tree is implemented by the computer, with the relevant expected utility results shown at each of the nodes. Thus, this is a fully transparent representation of the inputs and the results. As such, it shows which initial decision should be most favoured, and it indicates what subsequent decisions are the best policies depending on the outcomes of the future uncertain events.

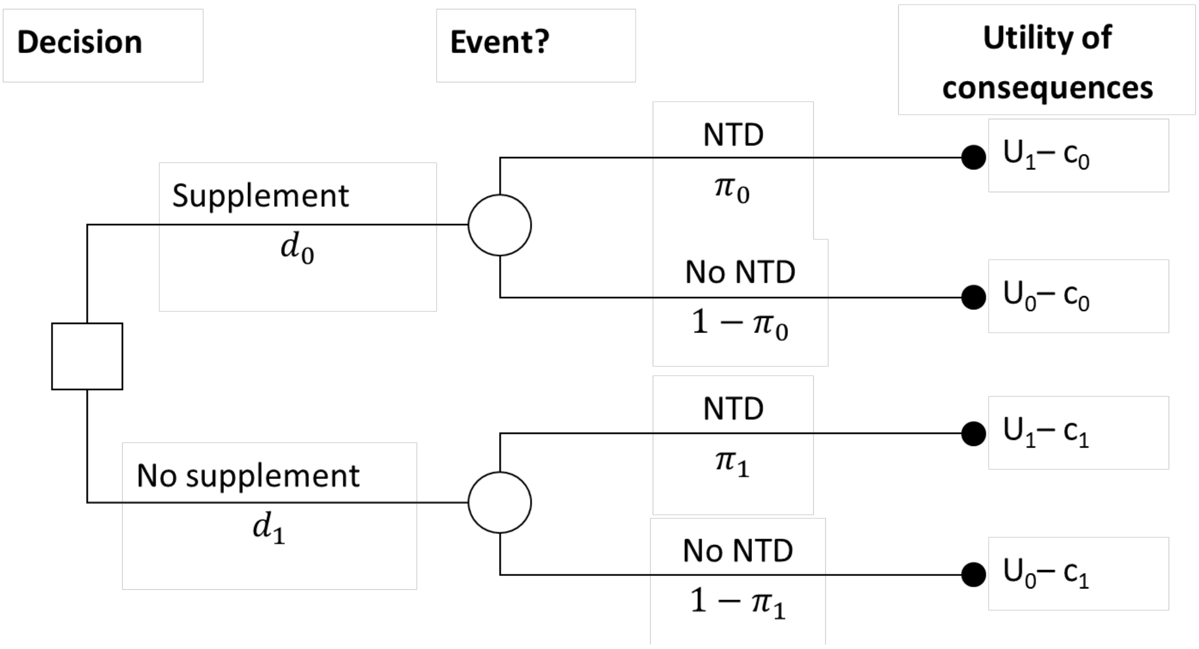
- Features

Current decision tree software provides a range of features that can accommodate even the most complex of problems, involving many options, numerous uncertain events and multiple objectives. Typical features include tornado diagrams (see Figure 19 for example) for carrying out deterministic analyses, which identify those uncertain quantities that most affect the final expected utilities, sensitivity graphs that enable individual utilities or probabilities to be varied, either singly or in combination, for their effects on the results, and value of information calculations to determine whether or not it would be worthwhile to commit resources to gain further information before making a decision.

- Visualisation

A simple decision tree for deciding whether or not to take a folic-acid supplement that could reduce the chances of a neural tube defect (NTD) in an unborn child is shown in Figure 11. The two choices, d_0 and d_1 are represented as branches emanating from a square node (called a 'choice node'). Uncertainty about the outcomes, NTD and No NTD, are shown on branches emanating from the circular node (an 'event node'), with the associated uncertainties about the NTD outcome shown as probabilities, π_0 and π_1, the latter assessed as higher than the former. The utilities associated with a child free of NTD are U_0, and with having NTD are U_1. The cost to the mother of taking or not taking the supplement is represented by c_0 and c_1, respectively, though these costs could be any mixture of monetary and non-monetary considerations. Once specific values are given to the probabilities, utilities and costs, the tree can be folded back to provide a probability-weighted utility for each choice, showing which should be more preferred.

Figure 11 Decision tree of folic acid supplement for pregnant women (reproduced from [4])



(Reconstructed from Spiegelhalter *et al.*[4] with the square representing a decision, hollow circles representing chance events, and solid circles representing the utilities)

Expected utilities $E(.)$ for $d_i, i = (0,1)$ are calculated as $\mathrm{E}(d_i) = \pi_i U_1 + (1 - \pi_i) U_0 - c_i$

- Assessability and accessibility

Familiarity with decision theory and decision analytic software is required for modelling problems more complex than the one shown in Figure 11. Once that knowledge is acquired, or brought in, developing a structure of the problem can prove challenging because most people are not accustomed to thinking in terms of clearly-defined decisions, events and their outcomes, and the possible consequences. With experience, it becomes easier, but at this stage we could not find any example of how a decision tree has been applied to the benefit-risk evaluation of drugs at the stage of a regulatory decision. They have been used by pharmaceutical companies, and by post-hoc analyses of approved drugs. However, this does not mean that the expected utility rule is irrelevant to benefit-risk considerations. In modelling the benefit-risk of several drugs being considered by the Committee on Human Medicines (CHMP), the expected utility rule was used to capture both the utility of a favourable or unfavourable effect of a drug and the probability that a patient would experience that effect [62]. In short, the expected utility calculation captured the relative desirability or undesirability of an effect weighted by the incidence of the effect. Thus, this aspect of decision theory provides a means for expressing all effects and their uncertainties as expected utilities, enabling benefits to be balanced with risks. Coupled with the weighted utility calculation explained in the next section, Multi-Criteria Decision Analysis, this balance can take account of multiple benefits and multiple risks. For this reason, there is substantial potential for decision theory to contribute to determining the benefit-risk balance of drugs, while they are in development, at the approval stage, and post-approval as more information is received about their effectiveness and risks.

A.6.4 Markov Decision Process (MDP)

Description

Markov chain combined with decision tree produces Markov decision process (MDP) which is a tool for multi-stage decision making with finite states and options. Transition probabilities among states of different stages and utilities (or costs) at all stages are the elements for decision making and the goal is to find an option (or combination of options at different stages) to maximize the expected utility of entire process.

We suggest reading Sonnenberg (1993) [14] for introduction and as core reference, and Thompson (2008) [63] for worked example.

Evaluation

- Principle

Markov decision process is a rigorous mathematical tool for multi-stage decision making with Markov dependence between the states in different stages. To apply the model in medical decision making, one has to estimate the transition probabilities associated with each option. Those probabilities may not be directly available in source data such as RCTs, for example, the probability of response to treatment without AEs (in RCTs, the response rate and AE rate are usually given separately). Also, the dynamic nature of MDP implies that not all medical decision problems are ready to be described as a MDP.

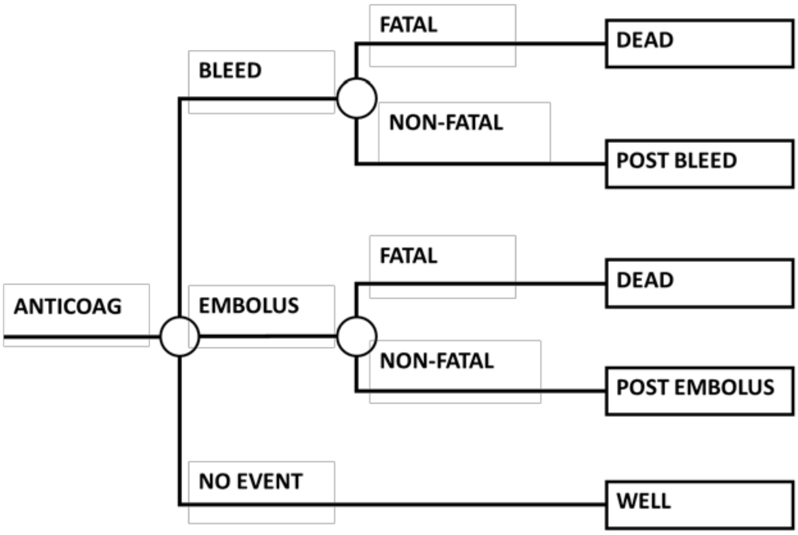
- Feature

MDP allows multiple criteria and multiple options. Multiple benefit and risk criteria however may make the structure of MDP very complicated. MDP also allows the benefit and risk status change with time, but the transition probabilities may not ready to elicit from the reports of clinical trials and epidemiology studies. Estimating transition probabilities from different sources of evidences could also be a difficult task. When transition probabilities are known, MDP could perform population level analysis, otherwise MDP only capable of performing individual level analysis. Updating the model with new data is not straight forward.

- Visualisation

A tree type decision structure usually is used to illustrate a MDP as shown in Figure 12.

Figure 12 A tree fragment modelling of anticoagulant therapy (reproduced from [14])



- Assessability and accessibility

MDP can be used by any stakeholders to include their preferences into utility at each stage. Understanding of the model may need some educations. This model may be suitable for post-market decision based on long-term data. Also, this model is also suitable for patient's decision to undergo a therapy considering all the possibilities and follow-up consequences. However, it is often very difficult to clearly define health states which are central to the application of MDP, therefore can restrict its use in medical decision making.

A.6.5 Multi-Criteria Decision Analysis (MCDA)

Description

Recognising that many decisions are characterised by multiple, often conflicting, objectives, the decision theory had been extended so that the consequences of decisions could be described by their relevant criteria (Keeney-Raiffa approach) [15]. MCDA and other multi-criteria decision making approaches are realisations of the multi-attribute utility theory (MAUT) which is an extension of unidimensional utility theory that allows multiple criteria to be combined in a logical way. The criteria are depicted in a value tree, with nodes representing objectives and the branches emanating from the nodes showing criteria that represent different ways of realising the objectives. Consequences of decisions are assessed separately for each of the criteria, and the criteria weighted to ensure the comparability of utilities across the criteria. By itself, a value tree can be used to create a deterministic model, for example, to balance benefits against costs, as in choosing to buy a car, when there is little or no uncertainty to be considered. The result is an overall preference value for each option, which provides an easily-understood ordering of the options.

When uncertainty is taken into account, then a decision tree to model the uncertainty and a value tree to represent the multiple objectives can be used in combination. The weighted utility rule combines the separate utilities on the various criteria into one total utility for each consequence, while the expected utility rule weights the total utilities for all the consequences to give an overall preference value for each option. Thus, the combination of expected utility and weighted utility can model benefits and risks as well as their uncertainties.

It is worth pointing out that MCDA is often used to refer to a collection of other multi-criteria approaches, which are well described elsewhere [64]. Many of those are not as comprehensive as the Keeney-Raiffa approach, so are not included in this review.

We suggest reading Mussen (2009) [65] for introduction in medicine and Dodgson (2000) [66] for more general introduction, Keeney (1976) [15] as core reference, and Mussen (2009) [67] for worked example.

Evaluation

- Principle

Because the principles of MCDA are logically sound, being based on decision theory, the resulting models are transparent insofar as the structure of the model (options, outcomes, consequences, objectives and criteria) and all quantitative inputs (utilities, probabilities and criterion weights) are made explicit. A generic eight-step process, based on the PrOACT-URL (Section A.5.1) – first proposed for use in health-care decisions in 2001 [13], but only ten years later being framed in the context of benefit-risk assessment of drugs [62;68] – provides a framework for both qualitative and quantitative assessments.

The first five steps, PrOACT, result in a deterministic model, which for benefit-risk assessment means that point estimates are input for all the effects. Once a result has been calculated, uncertainties of input values are dealt with through sensitivity analysis: changing utilities and criterion weights over ranges of concern to see what effect the changes have on the overall results. This may be sufficient for assessors to draw conclusions, but if not, further analysis, such as in combination with probabilistic simulation (Appendix A.10.2), may be helpful.

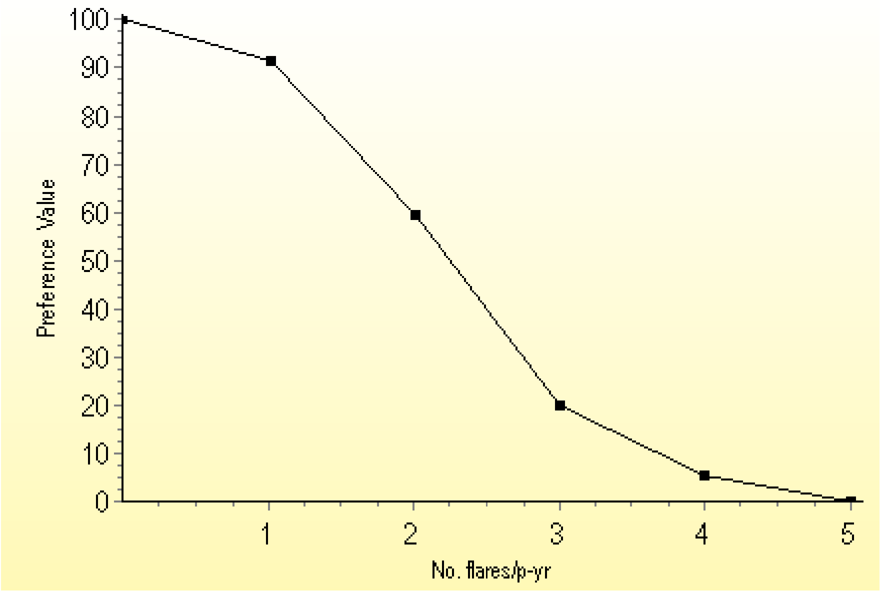
It is important to realise that certain conditions should be met if the simple calculations applied in MCDA are to give valid results. Briefly, the criteria should be:

1) Requisite in number – a complete set showing how the options differ in ways that matter to the decision maker(s), without double-counting.

2) Understandable as preferences – an unambiguous definition for each criterion, which indicates a clear direction of preference (more is preferable to less, or less to more).
3) Mutually preference independent – so that scores can be assigned on each criterion without having to know the scores on any of the other criteria.
4) Accommodating of preferences over time – fully covering the time horizon of the model (possibly a mixture of short-term and long-term criteria).

A key question for applying MCDA to the benefit-risk assessment of drugs is who does the scoring and weighting. Measurable data are usually available, but these must then be translated into preference scores through the use of value functions. These are often linear, direct or inverse, but may be non-linear. For example, Figure 13 is a value function determined by clinical assessors for flares per patient-year of lupus disease. There it can be seen that a drug which lowers the average number of flares per year for a population of patients from 3.5 to 2 would be considered much more effective than one which lowers it from 3.5 to 3. The ratio of differences in flare rate is 1.5 to 0.5, or 3 to 1, but the preference value differences are 45 to 5, or 9 to 1. This non-linear relationship is, of course, a matter for clinical judgement, but it could be different for industry, regulators, practitioners or patients. Criterion weights are also a matter of clinical judgement, and, again, could be made differently by different constituents. MCDA modelling requires these judgements, which can be seen as negative in the sense of including subjectivity into benefit-risk assessment, or positive as revealing and making explicit assessments that cannot be made objectively; data don't speak for themselves.

Figure 13 Number of flares per patient-year for lupus disease



- Features

MCDA appropriately weighs each criterion so that the units are equivalent across all the criteria; therefore the option scores are directly comparable and combinable. The overall scores for a typical MCDA benefit-risk model have two components: overall benefits score and overall risks score. Benefits and risks generally form high-level criteria which can naturally accommodate several sub-criteria. Evidence from different sources or time dimension is just another criterion in MCDA. The obvious advantage of MCDA over other benefit-risk assessment approaches is that it is capable to simultaneously compare more than two options in one model; differences are not inputs, they are outputs. Sensitivity analysis is a required step in MCDA. Specialist software such as Hiview 3 (http://www.catalyze.co.uk), V•I•S•A (http://www.visadecisions.com), Intelligent Decision System (http://www.e-ids.co.uk) and Logical Decisions (http://www.logicaldecisions.com) facilitate the implementation of MCDA.

- Visualisation

There is no specific visualisation technique for MCDA per se, but Hiview 3, for example, provides some useful visuals to communicate the results, shown in Figure 13 to Figure 17. Throughout, the terminology now favoured by the European Medicines Agency (2010) of ‘Favourable Effects’ and ‘Unfavourable Effects’ replaces the more common ‘Benefits’ and ‘Risks’.

Figure 14 The value tree for Drug X, used in combination with Methotrexate, for adult lupus disease

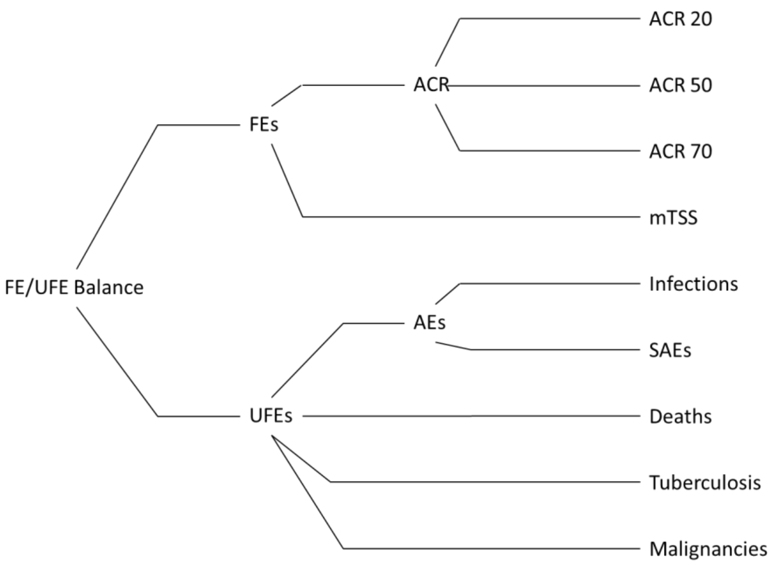
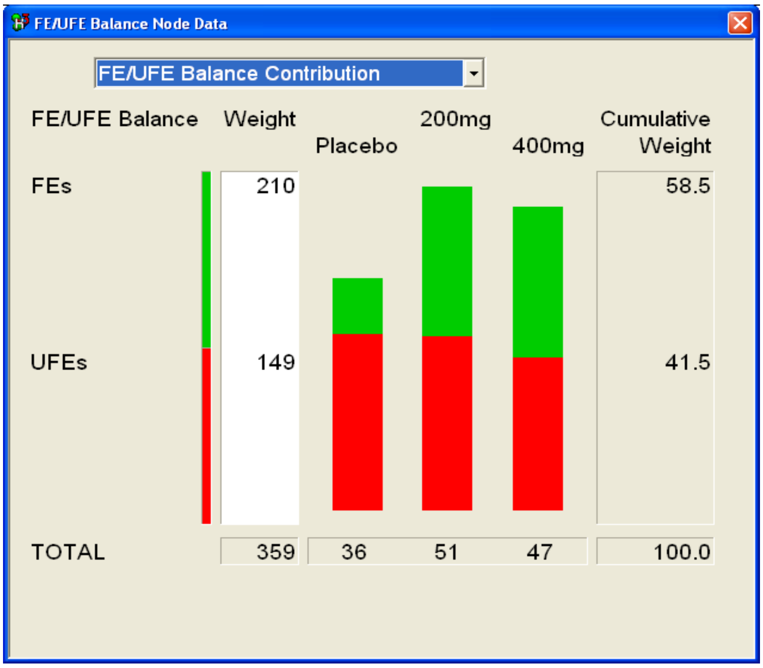
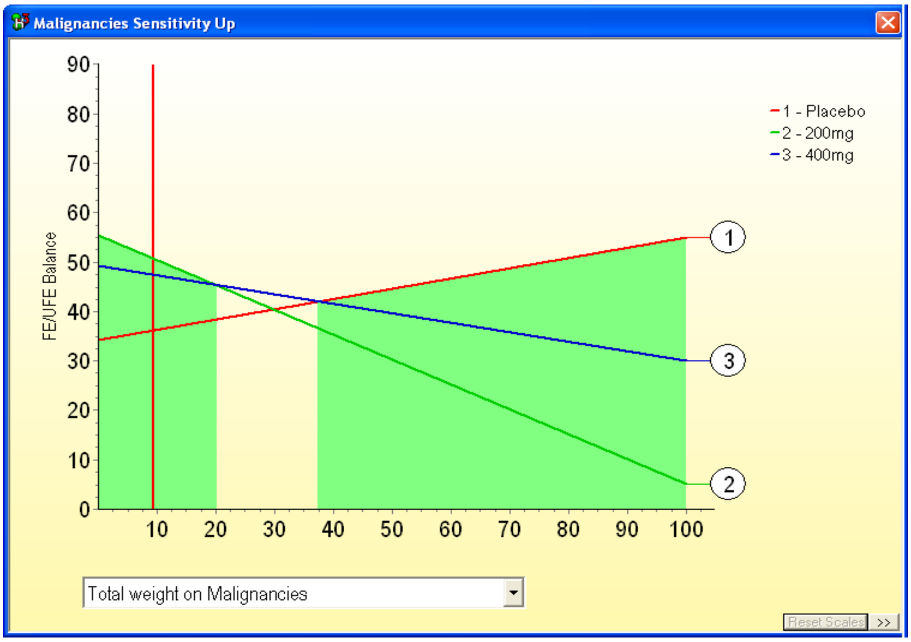


Figure 15 The overall weighted preference values for the 200mg and 400mg doses of Drug X, and the placebo, all used in combination with methotrexate



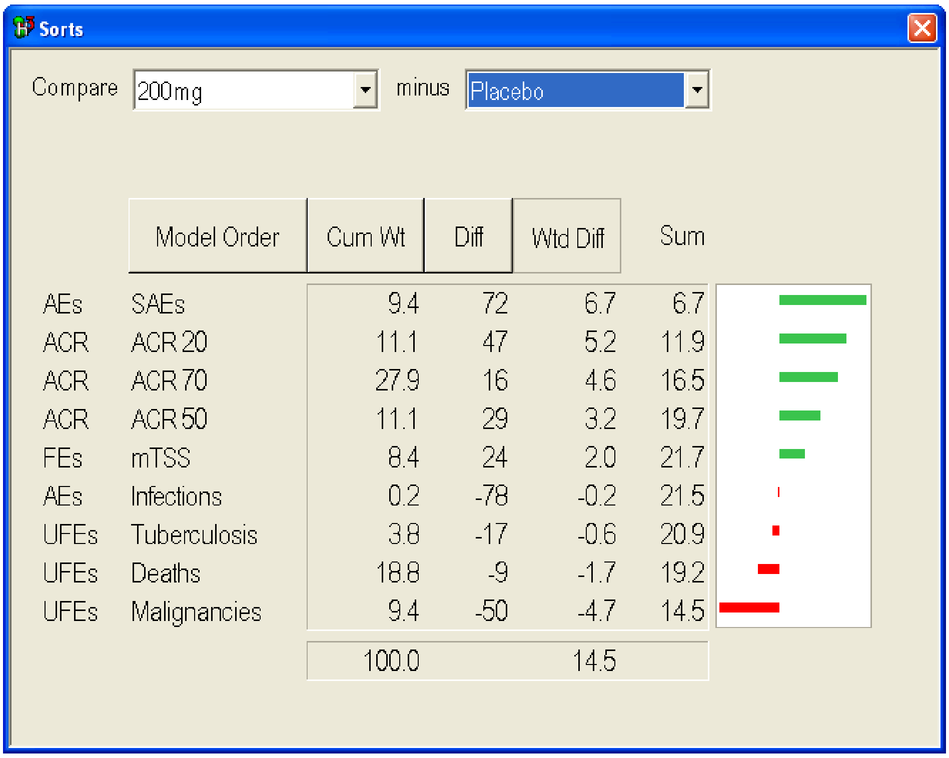
Longer green bars (upper portions) indicate more benefit (more favourable effects), longer red bars (lower portions) indicate more safety (less unfavourable effects). Overall, the 200mg dose is most preferred because it is safer than the 400mg dose.

Figure 16 A sensitivity analysis on the weight given to malignancies



The vertical red line shows the weight of 9% currently given to the malignancy criterion. The intersection of that line with the three lines above it – 36, 47 and 51 – correspond to the overall scores of the placebo, 400mg and 200mg options, as shown in the previous Figure 15. Increasing the weight on malignancies would first cause the 400mg dose, then the placebo to become most preferred.

Figure 17 Comparison of the more preferred 200mg dose with the placebo



Compare 200mg minus Placebo

	Model Order	Cum Wt	Diff	Wtd Diff	Sum
AEs	SAEs	9.4	72	6.7	6.7
ACR	ACR 20	11.1	47	5.2	11.9
ACR	ACR 70	27.9	16	4.6	16.5
ACR	ACR 50	11.1	29	3.2	19.7
FEs	mTSS	8.4	24	2.0	21.7
AEs	Infections	0.2	-78	-0.2	21.5
UFEs	Tuberculosis	3.8	-17	-0.6	20.9
UFEs	Deaths	18.8	-9	-1.7	19.2
UFEs	Malignancies	9.4	-50	-4.7	14.5
		100.0		14.5	

Green (right) bars show the five advantages of the 200mg dose, and the red (left) bars the placebo. The columns show the Cumulative Weight for each criterion (summing to 100), the Difference column the difference in preference scores between the 200mg dose and the placebo, the Weighted Difference column shows the weighted difference scores, and the last column of figures gives the cumulative sum of the weighted difference scores. Thus, each row shows the 'part scores' associated with each criterion such that their sum equals the difference between their overall scores, 14.5.

- Assessability and accessibility

Building an MCDA model requires a working knowledge of decision analysis, but, contrary to the worries of some commentators, Phillips et al reported that his team were able to build each of five different models about drugs under review by the CHMP in six hours or less [62]. This was possible because in each case a team of four to six clinical assessors and experts, working as a team, had already carefully examined the dossier and were reasonably well acquainted with the issues facing them in their task of writing an assessment report. The options were clear and the criteria were available, if not already clearly set out in a single table. The purpose of the modelling focussed on the process of incorporating both data and clinical judgements, then exploring the sensitivity of the results to imprecision in the data, uncertainty and differences of opinion. At the conclusion of this process, participants felt better informed about the benefit-risk balance. MCDA improves decision-making through its requirements for documentation of the process, but the accuracy of outputs are still dependent on the knowledge and reliability of the experts whilst the final decision is still subjectively made by the decision-maker (as in any other approaches).

A.6.6 Stochastic Multi-criteria (-objective) Acceptability Analysis (SMAA)

Description

SMAA is a family of methods to deal with MCDA problems when the consequences of the actions under concern are random and the choices of weights across criteria are unclear [69]. SMAA can be seen as an extension of MCDA with an added advantage of being able to include parameter uncertainties due to sampling variations, and the ability to characterise typical benefit-risk trade-offs without preference information. Here one should be aware that MCDA includes a variety of different methods which may roughly be classified into two classes: MAUT model and outranking model. SMAA was originally developed as MAUT model (SMAA, SMAA2) and later was adapted for outranking model (SMAA3, SMAA-TRI). The MCDA approach currently used in medical benefit-risk analysis is additive utility model (MAUT). We introduce mainly SMAA and SMAA-2, whilst briefly mentioning the other varieties for awareness.

In additive utility model, the consequences of an action on all criteria are converted into a sum of weighted utilities (or scores), which is also called value function. An action is ranked first among all the actions in additive utility model if its value function has the largest value. In SMAA, since the consequences are random, the probability of an action i being ranked first – or the “acceptability index” – is:

$$b_i^1 = \int_{\xi \in X}^{\cdot} f_X(\xi) \int_{w \in W_i^1(\xi)}^{\cdot} f_W(w) dw d\xi$$

Here $f_X(\xi)$ is the density function on the space of all consequences *X*, a subset of $R^{m \times n}$ (*m, n*: number of actions and number of criteria), $f_W(w)$ is the density function on weight space $W = \{w \in R^n : w_1 + \cdots + w_n = 1\}$ and $W_i^1(\xi)$ is the action i favourable weight space; that is, given consequence ξ, action i is ranked first for any $w \in W_i^1(\xi)$. The density $f_W(w)$ can be a uniform density on weight space W if the DM does not know the weights on the criteria. The expected centre of gravity of set $W_i^1(\xi)$ is called the “central weight” vector for action i, and is defined as:

$$w_i^c = \frac{1}{b_i^1} \int_{\xi \in X}^{\cdot} f_X(\xi) \int_{w \in W_i^1(\xi)}^{\cdot} w f_W(w) dw d\xi$$

The probability that action i is ranked first with weight vector w_i^c is the “confidence factor” p_i^c; which is then compared against each other to determine the better action. The larger the p_i^c, the better the action i. SMAA-2 is an immediate extension of SMAA which considers, in addition to the chance of being ranked first, the chance that an action is ranked second, third etc., depending on the total number of actions.

SMAA-3 and SMAA-TRI are the methods of SMAA family based on the outranking MCDA methods ELECTRE III and ELECTRE TRI respectively. For each possible consequence, the ranking is based on ELECTRE outranking methods and, similar to SMAA, the favourable weight space for each action is to be defined accordingly. SMAA-3 calculates the probability of an action being ranked first, and SMAA-TRI calculates the probability that an action being ranked into “first class” for the purpose of comparison between multiple actions.

In SMAA family methods, all the probabilities are high dimensional integrals. Therefore they are calculated by simulations.

We suggest reading Tervonen (2008) [70] for general introduction or Tervonen (2011) [17] for introduction for drug benefit-risk assessment with worked example, and Lahdelma (1998) [69] as core reference.

Evaluation

- Principle

The principles of SMAA are very similar to MCDA and are based on decision theory. The transparency in decision-making process is increased by its ability to elicit preference information post-hoc therefore the model built can be prior to knowing the specific situation it will be used in. Statistical uncertainties of the posterior distributions are propagated from the statistical uncertainties in input values expressed as arbitrary stochastic variables, and the uncertainty in observed data producing confidence intervals for the point estimates which are then used when making decisions. Understanding the principles of SMAA require mathematical understanding of stochastic phenomena and uncertainty. Value judgments in SMAA can be implicit when preference information are missing but become explicit when these are obtained.

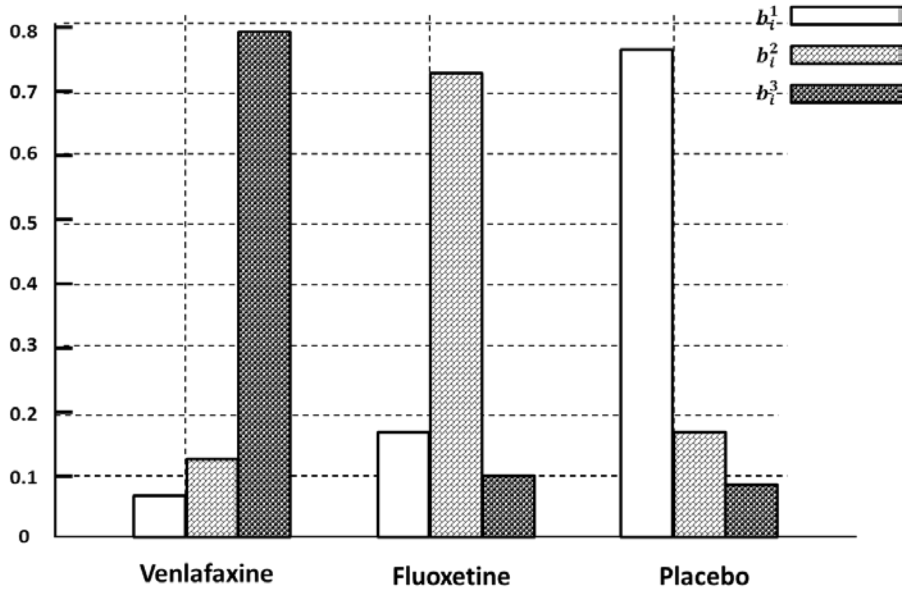
- Features

As in MCDA, SMAA provides integrated benefit-risk weighted utility scores for each option and ranks the different options. Multiple benefits and risks as well as multiple sources of evidence can be dealt with in the same way as MCDA. More than two options can be assessed simultaneously and ranked the overall weighted utility scores.

- Visualisation

Bar charts have been proposed to illustrate the components acceptability (Figure 18).

Figure 18 Visualisation of rank acceptability index for SMAA-2 model without preference information for therapeutic group of antidepressants (reproduced from [17])



b_i^r indicates that alternative i is ranked at place r

- Assessability and accessibility

The final SMAA results rank different options being compared similar to MCDA therefore making it easily interpretable. The practicality of applying SMAA in comparison to MCDA requires considerably more mathematical and computational knowledge despite the availability of open-source software, JSMAA, to implement the model (http://smaa.fi/jsmaa.php). SMAA is suitable and is readily-built for real-time interactive decision-making.

A.6.7 Sarac's Benefit-Risk Assessment Method (SBRAM)

Description

This approach is proposed by Dr Sinan Sarac in the Novo Nordisk A/S and the Technical University of Denmark [18]. The Sarac's Benefit-Risk Assessment method involves eight successive steps in its defined framework: (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. Some elements of this approach are shared with MCDA (Appendix A.6.4). To allow comparison 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 objective information (data) wherever possible. Weights and scores are multiplied, and the results are visualized in standardized diagrams.

We suggest reading Sarac (2010) [18] for more details and worked example.

Evaluation

- Principle

The scores are objectively based on descriptive statistical comparison to the comparator, whenever possible. The -1, 0, 1 system, representing inferior, equivalent, superior (to comparator), cannot distinguish between two candidates that are all superior (or all inferior) to the comparator, if the two candidates do not have direct comparisons. However, an assessment can be conducted between two candidates, where one of the candidates act as the comparator and all differences between the two are captured, similar to the concept of mixed treatment comparison method (Appendix A.10.4). Uncertainty is estimated both quantitatively and qualitatively and may result in an interval-score. Quantitative uncertainty evaluation is conducted according to the well-known principles of "bootstrapping", [71]. The scores in all criteria are not integrated; but they are graphically presented in a tornado-like diagram where judgments on the multiple benefit and risk criteria can be made visually.

- Feature

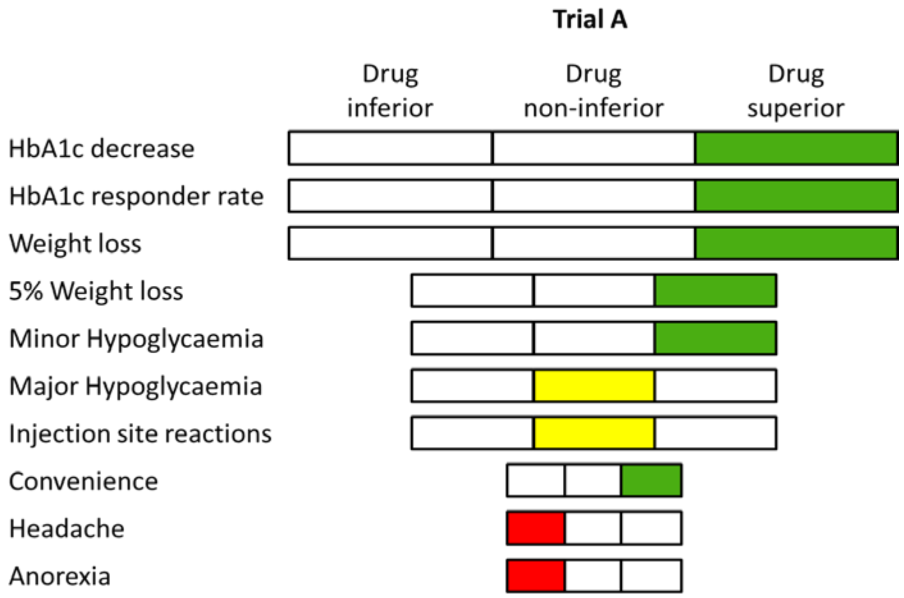
This method is designed for both pharmaceutical companies and regulatory agencies. It uses clinical trial data, but can incorporate other data as well, e.g. external data on an already marketed comparator, in the evaluation of uncertainty and evidence. The method is based on data from experimental and clinical studies as well as from other sources of information. While supported by professional statistical analysis, the method itself is of a qualitative nature in order to properly allow for uncertainties and differences in opinion. This qualitative feature also serves to focus the assessment on clinical and toxicological issues and it allows subjective comparison of benefits and risks to be made.

In principle the method can also be developed into a dynamic assessment that follows a drug from its first conception to the end of its life. Clinical relevance plays an important role in the assessment of a drug or drug candidate. In present context, we define the clinical relevance of a drug by the proportion of patients who experience a clear effect of a prescribed treatment with the drug. The criterion to be able to say that a drug effect is clinically relevant is when the conditions are improved for more than two out of three patients, but this principle may not be applicable for different medical conditions, so an appropriate ratio should be chosen. This criterion differs from the criterion of statistically significance of a positive drug effect in that it measures the fractions of patients helped by the drug rather than demonstrating an improvement for the average patient. A very small difference in effect can be statistically significant if the number of patients is high. On the other hand, a significant effect of the drug can be statistically insignificant if the number of patients examined is small.

- Visualisation

Visualisation of the weighted scores for all criteria simultaneously is a step in the procedure for judgement of benefit-risk balance. This is illustrated in Figure 19.

Figure 19 An example of tornado diagram. The width of the box indicates the weight of the criterion; the colours represent scores: red = -1, yellow = 0, green = 1 (reproduced from [18])



- Assessability and accessibility

The scores in this approach are based on descriptive statistics whenever possible. Different stakeholders (e.g. industry and regulatory agencies) are likely to have different views on weighting and scoring of the various criteria, and they may also disagree about the choice of decision criteria. It is therefore important that this information is fully maintained and clearly presented in the final decision process. On the other hand it is clearly also important that the underlying statistical analysis is available in a clearly understandable form.This approach is designed for, e.g. pharmaceutical companies to evaluate new candidate drugs or to prepare new drug application. It seems fit for these purposes.

A.6.8 Clinical Utility Index (CUI) and Desirability Index (DI)

Description

The idea and application of DI came long before CUI, where it has been used extensively in industrial quality management [72]. CUI came much later [21] and it was introduced as a means to assess the therapeutic index of new drugs [20]. CUI is defined as the weighted sum of benefits and risks where benefits take positive values and risks take negative values, which takes the functional form $CUI = \sum_{i=1}^{m} w_i \times U_i$ for utilities U_i and weights w_i when there are $1 \leq i \leq m$ criteria. Its relation to DI can be seen as the geometric equivalence, where DI is defined as $(\prod_{i=1}^{m} U_i)))^{1/w}$. So far, CUI/DI is only known to be used in drug development where the utility index is expressed as a function of dose [19;73]. The resultant metrics from CUI/DI framework are also known as CUI and DI respectively.

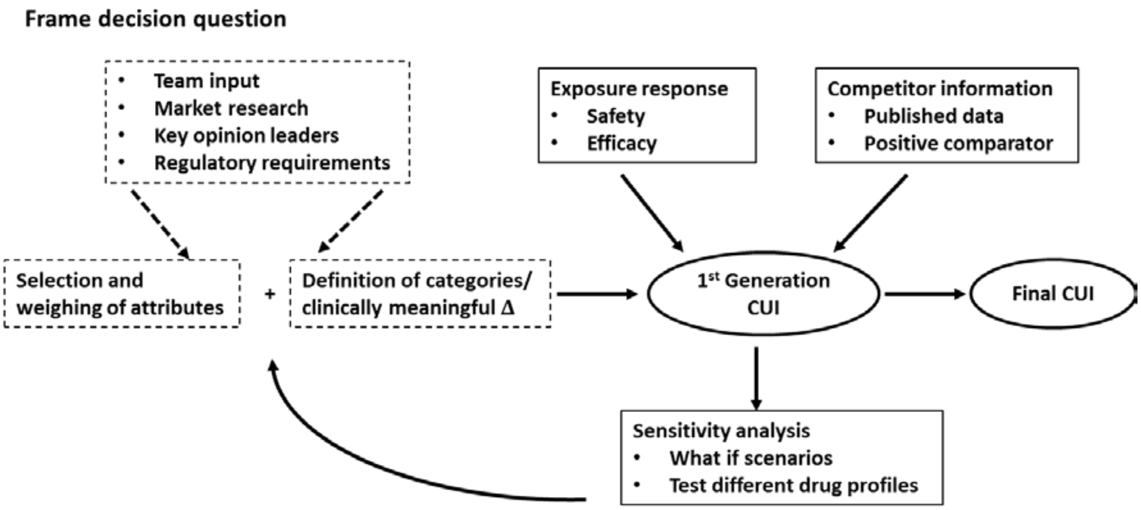
We suggest reading Ouellet (2010) [20] for more details, and Ouellet (2009) [73] and Renard (2009) [19] for worked examples.

Evaluation

- Principle

The original desirability index is derived from specific desirability functions [72] where its statistical distributions have been studied [74]; unlike CUI. The derivation of DI is generalised (without proof) to any function on (0,1) scale [19] and shares the strengths and weaknesses of CUI, where the functional forms of combining the benefits and risks are arguably interchangeable. CUI comes within a four-step framework [20]: (1) exposure-response analysis of benefits and risks endpoints; (2) definition of criteria to define clinically meaningful changes; (3) selection of important attributes and definition of relative weights; and (4) sensitivity analysis and measurement of uncertainty. This is illustrated in Figure 20 [20]. However, the transparency of CUI (and DI) is hampered from having a rather too general framework. Uncertainties of CUI estimates can be obtained through simulations and other uncertainties can be addressed in a similar way to MCDA. In general, the principle of CUI is straightforward and easily understood by end users, but the actual derivation of CUI requires more extensive knowledge and understanding of the concept of functional analysis.

Figure 20 Key aspects in the development of a CUI (reproduced from [20])



- Features

CUI and DI provide integrated measures of benefits and risks. They can accommodate several benefits and risks criteria, as unidimensional criteria which are later combined. Multiple sources of evidence are dealt as criteria similar to MCDA (see Section A.6.5). It is possible to incorporate time dimension in place of or in combination with dose range. When a criterion is totally unacceptable (valued as zero), it is ignored in CUI because CUI is a function of the sum of the expected utilities. However, the same criterion would render DI to be totally unacceptable because DI is a function of the products of the expected utilities.

- Visualisation

Clinical utility index can be visualised over a range of exposure (dose) and is illustrated in Figure 21. A three-dimensional surface plot of desirability versus efficacy and safety has also been suggested as in Figure 22, and the choice of weightings can be aided using equi-desirability contour plots illustrated in Figure 23[19].

Figure 21 Dose-response relationship for efficacy, toxicity and utility index (reproduced from [20])

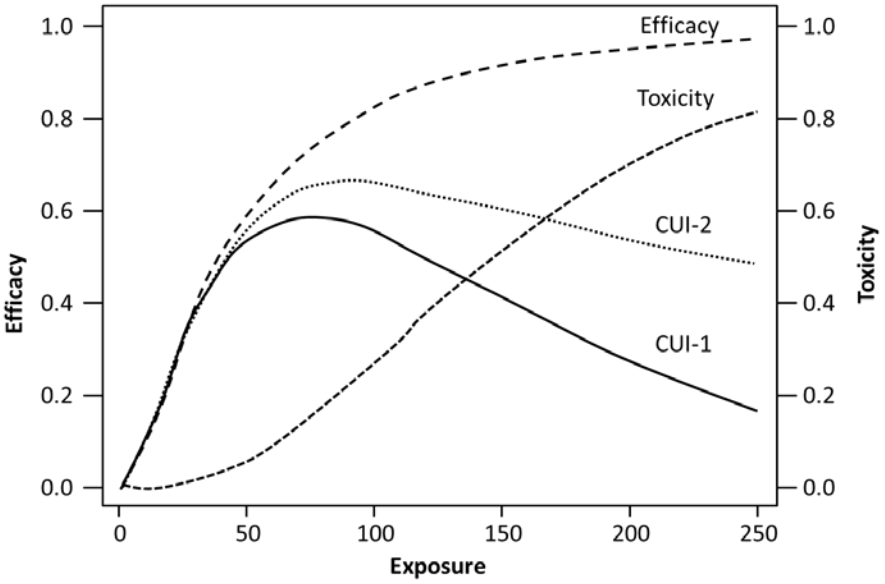


Figure 22 Desirability surface (reproduced from [19])

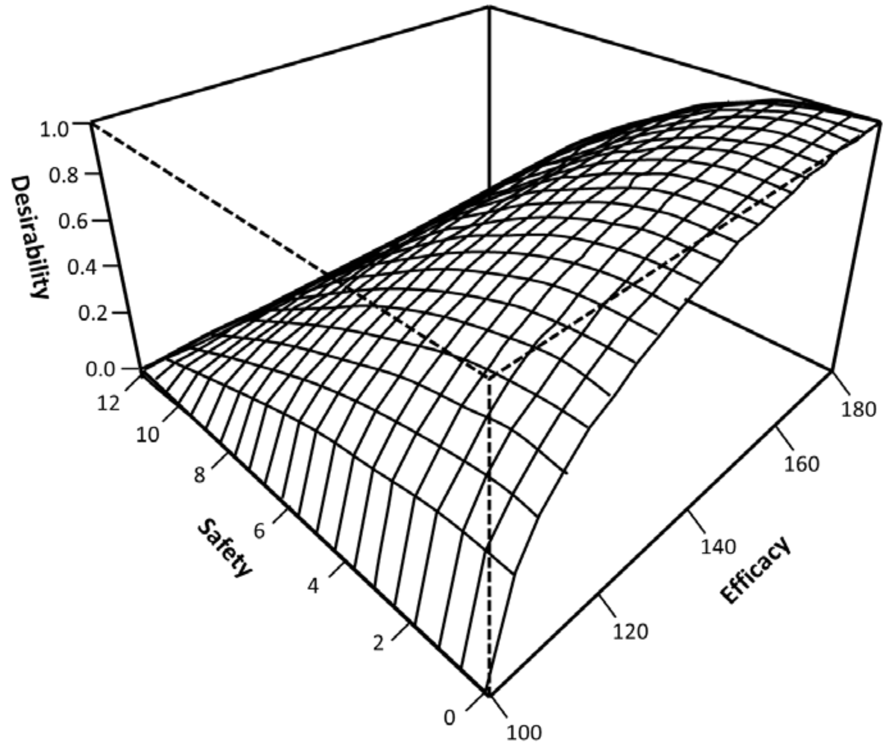
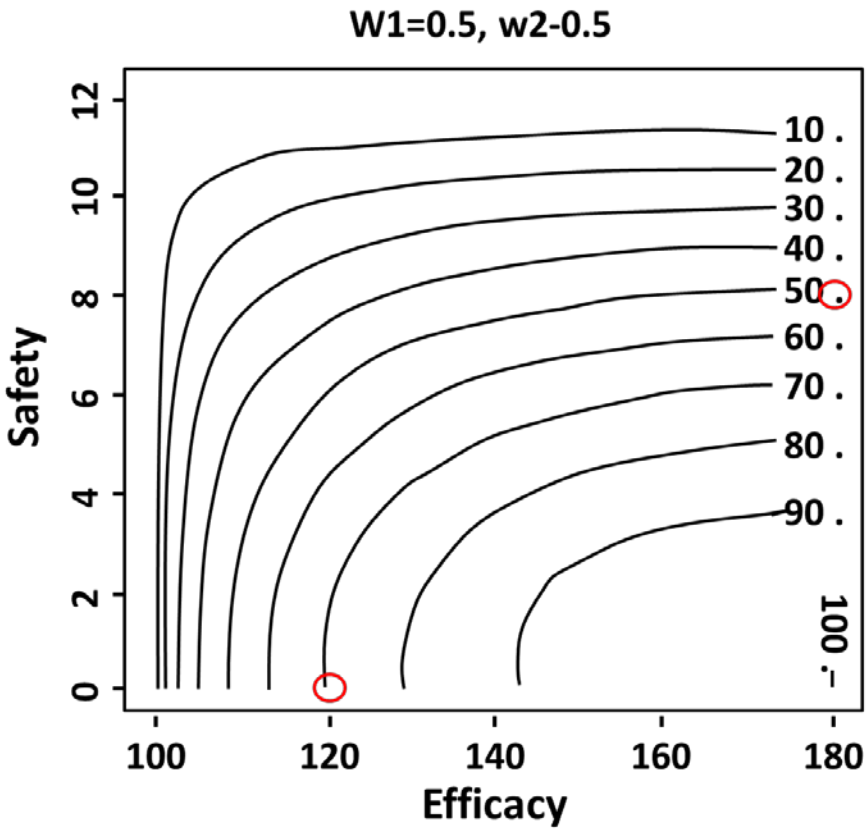


Figure 23 Equi-desirability contours and weightings (reproduced from [19])



- Assessability and accessibility

The parameters and results are acceptable given that the choice of utility functions used is acceptable. They can be interpreted easily as the excess benefit having discounted risks (CUI) or the multiples of benefits per unit risk (DI). Other aspects of assessability and accessibility are similar to MCDA but there is no specialist software to aid its application.

A.7 Quantitative threshold indices

Author: Shahrul Mt-Isa

A.7.1 Number needed to treat (NNT) family approaches

Description

The idea of measuring benefits in terms of number of patients needed to be treated to prevent one patient from an adverse event or a disease originated from Laupacis (1988) and has been termed the number needed to treat (NNT) [24]. NNT is derived from the probabilities of events in two treatment group of interest, where the difference between the two probabilities p_1 and p_2 gives the "excess" events in one group, i.e. $p_1 - p_2$. NNT is then calculated as the reciprocal of this difference, produces a metric that follows a geometric waiting time distribution $1/(p_1 - p_2)$. In epidemiological terms, an NNT is just the reciprocal of an absolute risk reduction. A parallel but opposite metric, number needed to harm or NNH, gives an estimate of number of patients needed to be treated to observe one patient developing an adverse event or a disease using probabilities q_1 and q_2 denoting risks in treatment 1 and 2. In a benefit-risk assessment, NNT and NNH are calculated independently and directly compared against each other. A treatment with $\mathrm{NNH} > \mathrm{NNT}$ is favourable; and a greater ratio of $\mathrm{NNH}/\mathrm{NNT}$ signifies a better treatment option. However, comparing NNT directly to NNH is not logically sound as this implicitly assumes equal weighting when they are most commonly not.

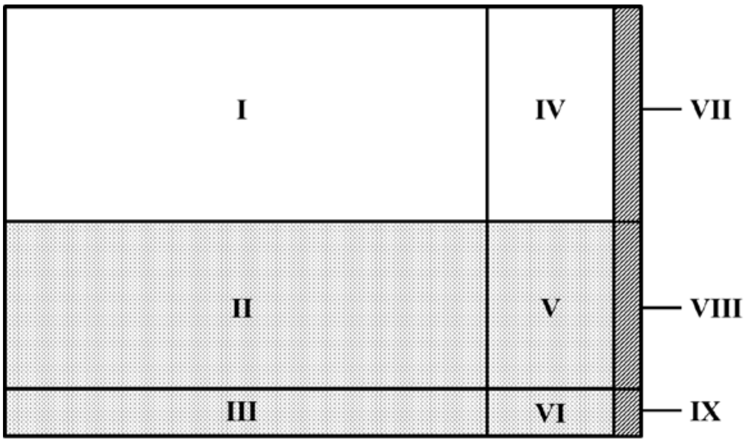
We suggest reading Holden (2003) [30] as an introductory reference with worked example for NNT, and Laupacis (1988) [24] as core reference.

An approach to adjust NNT for utility and the timing of benefits and risks has been proposed and termed the utility- and time-adjusted NNT (UT-NNT) [34]. UT-NNT has similar flavour to INHB approach (Appendix A.9.1). UT-NNT is calculated from the reciprocal of the difference between the two "incremental" benefits and "incremental" risks as in INHB. The adjustment for time is made by multiplying time saved or lost due to treatment by the probabilities of benefits or risks. Adjustment for utilities is made in the same way. UT-NNT is more a trade-off index than a threshold as it provides the way to integrate and trade off benefits and risks, but is described here because it is a direct extension of NNT and fit within the NNT family approaches.

We suggest reading Riegelman (1993) [34] for more details and worked example on UT-NNT.

NNT approach was also extended to account for adverse events occurred on treatment known as the adverse event adjusted NNT (AE-NNT) [26]. AE-NNT estimates the number of patients to be treated to observe one patient in whom treatment was successful without inducing treatment-related adverse events ("unqualified success"). It is most useful in a trial where the outcomes of a treatment are well defined with respect to what is considered a successful treatment and what are considered as the adverse events. It cross-classifies the group of patients based on these characteristics as illustrated in Figure 24. AE-NNT corresponds to the reciprocal of the total probabilities in groups IV and VI (Figure 24). A synonymous but opposite measure is the number of patients to be treated to observe one patient in whom treatment was a failure but had induced adverse events ("unmitigated failure"), which corresponds to group VIII (Figure 24).

Figure 24 A treated population, suffering (independently) from condition A (lined; p_2=5%) and from adverse event B (stippled; q_2=50%). Broken lines indicate the corresponding boundaries in untreated population [26]



I: never developed condition A or AE B, whether treated or not
II: never developed A, whether treated or not, and in whom treatment has induced B
III: never developed A, whether treated or not, and developed B unrelated to treatment
IV: A has been prevented by treatment, and never developed B, whether treated or not
V: A has been prevented by treatment, and in whom treatment has induced B
VI: A has been prevented by treatment, and developed B unrelated to treatment
VII: treatment failed to prevent A, and never developed B
VIII: treatment failed to prevent A, and in whom treatment has induced B
IX: treatment failed to prevent A, and developed B unrelated to treatment

We suggest reading Schulzer (1996) [26] for more details and worked example of AE-NNT.

Another generalisation of the extension of NNH approach to include utilities of benefits and risks comes as the relative values adjusted NNH (RV-NNH) which focuses on risks [27]. Relative value is defined as the ratio of risk utility to benefit utility. RV is then multiplied by the absolute risk difference of a particular AE to give RV-NNH. A composite RV-NNH for multiple AEs can be derived by taking the reciprocal of the sum of the products of absolute risk difference for particular risks and their relative values. This is given by the following equation for AE k with probabilities q_{ik} in group $i = 0,1$:

$$\text{RV-NNH} = \left\{\sum_{k=1}^{K} (q_{1k} - q_{0k}) \times RV_k\right\}^{-1}$$

We suggest reading Guyatt (1999) [27] for more details and worked example on RV-NNH.

Another family of approaches based on NNT known as the impact numbers to assess the benefits and risks of a given treatment in a population is described separately in Appendix A.7.2.

Evaluation

- Principle

In general, the calculations of the original NNT approach and its extension are transparent due to its apparent simplicity but the choice of input values need to be more transparent e.g. in explicitly stating and justifying the source of data used. Input values for NNT must be rates/probabilities from present knowledge. Although prior work such as combining rates in a meta-analytic framework etc. can be done, it is not a standard NNT approach.

The NNT approach can only deal with one criterion at a time, whether benefit or risk. AE-NNT deals with one benefit and one risk criteria but is conditional on the risk being null. UT-NNT and the more generalised RV-NNH can deal with multiple criteria through weighted summation of criteria. However, the sum of the probabilities in the composite RV-NNH has an infinite upper bound, therefore the resultant RV-NNH measure approaches zero contributing to implausible interpretation of the measure.

NNT extensions incorporating utilities, UT-NNT and RV-NNH, provide a way to incorporate value judgments, a vital parameter in decision-making that is missing from NNT approach. However, they violate decision theory when the reciprocal of the product of utilities and probabilities (i.e. expected utilities) is interpreted in the same way as in NNT.

- Features

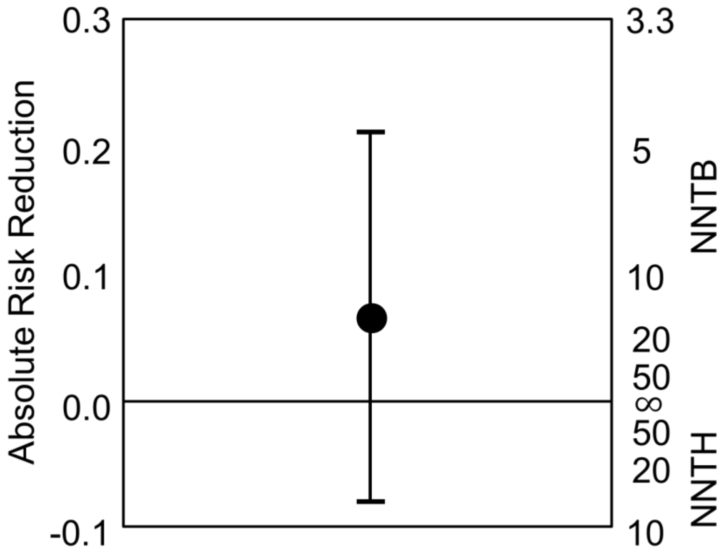
The benefit and risk are described by NNT and NNH separately. AE-NNT integrates benefit and risk but only estimates the NNT in the best case scenario; whilst the reciprocal of unmitigated failure estimates the NNT in the worst case scenario. UT-NNT can integrate benefit and risk into a single measure for analysis; and particularly projected itself to being superior to NNT due to its ability to incorporate the time dimension as well as utilities. RV-NNH does not properly integrate benefit and risk into a single measure, but only penalise the excess risk by the ratio of risk utility to benefit utility. In order to accommodate multiple sources of data, the NNT family approaches require meta-analyses to be performed.

Some work improving these approaches for pragmatic applications include the considerations of correlations between events in AE-NNT [26], and empirical posterior distribution estimation of the credible intervals for NNT under Bayesian framework (to some extent similar to confidence intervals) [75]. Additionally, NNT family approaches can be readily used for resource evaluation because it estimates the number of patients needing treatment thus resources like cost of treatment can be easily assigned and computed.

- Visualisation

Only the means and confidence intervals plots have been used for NNT. Altman (1998) proposed to reverse the y-axis to show that mean is within the continuous region of its CI as shown in Figure 25.

Figure 25 Relationship between absolute risk reduction, number needed to treat and their confidence intervals (reproduced from [76])



- Assessability and accessibility

The parameters required for NNT analysis are straight-forward i.e. rates/probabilities of events. However, the source of the rates/probabilities can be questionable due to the quality of data sources and is subjected to bias. The point estimates of the NNT are generally easy to interpret (given some technical knowledge) but the CIs have been criticised when rates include zero, the CI includes infinity (see also relevant work described above by Altman (1998) [75]). NEAR avoids the null point and the "signs" problem in the NNT approach being a ratio. However, NEAR OR and RR must be interpreted rigorously within the context of the problems and with great caution exercised when extrapolating NEAR results [29]. NNT family approaches although simple, can aid decision making based on evidence provided that decision-maker understands the relevance of underlying data/assumptions that were used to construct the measure.

A.7.2 Impact numbers

Description

The concept of NNT has been extended to provide several measures based on the population health perspective; and are known as "impact numbers" [77]. Adjustments are made on the rates of events (benefits or risks) to correct for the population being represented. One strict assumption is that causation is established between the exposure and outcome. The first two introduced were the disease impact number (DIN) and population impact number (PIN) [77]; and further three impact numbers were also described in a follow-up article, which are the case impact number (CIN), exposure impact number (EIN), and exposed cases impact number (ECIN) [77]. In the following year, the number of events prevented in the population *(NEPP)* and the population impact number of eliminating a risk factor over time *t (PIN-ER-t)* were introduced [78;79]. An online probabilistic simulation program to calculate NEPP and PIN-ER-*t* along with their confidence intervals is available at http://www.phsim.man.ac.uk/.

We suggest reading Attia (2002) [77] and Heller (2002) [80] as introductory and core references, and Heller (2003) [78] for worked example.

Evaluation

- Principle

Impact numbers are conceptually the same as NNT because they are derived from and calculated in a similar way to NNT. The interpretation is also similar. Therefore impact numbers share many of the strengths and weaknesses of NNT. The importance of justifying data sources used when using impact numbers is emphasised, which when compared to NNT approach is more transparent [77]. The idea of describing different populations is not new, and is generally based on the more well-known epidemiological measures such as attributable risks, attributable fractions and their population counterparts. It is also apparent that impact numbers are undefined when there is no difference in risks between options.

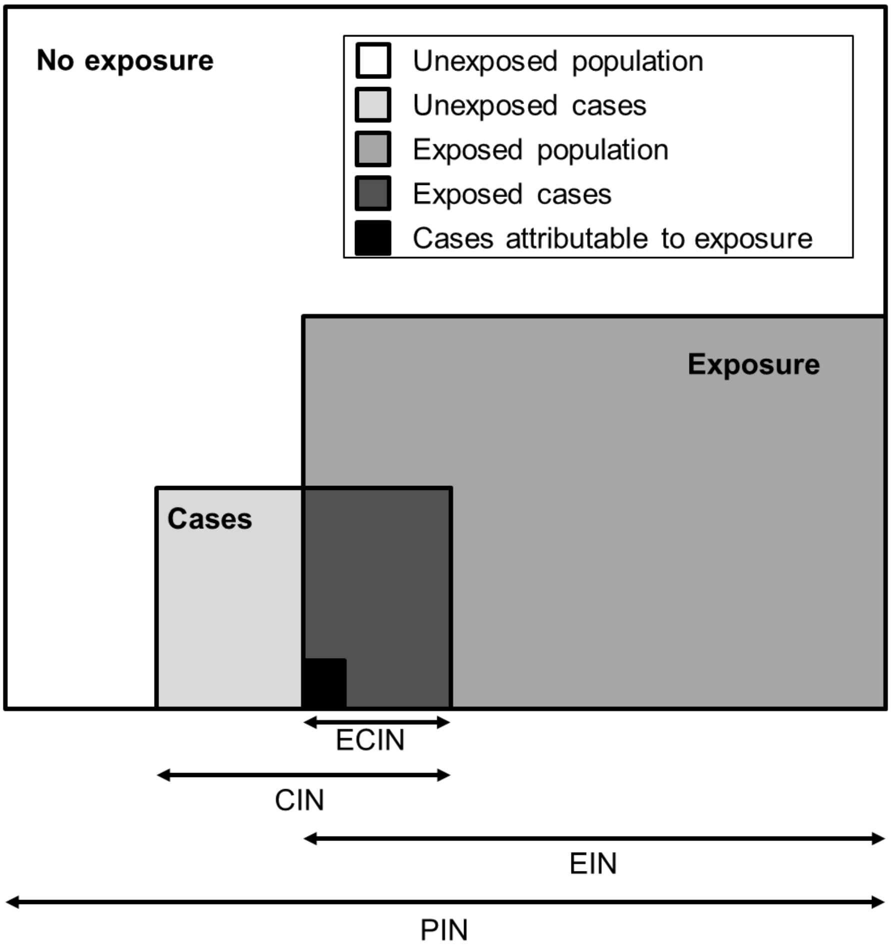
- Features

The features are very much similar to NNT (Appendix A.7.1) but most importantly, impact numbers were developed to give population health perspective with specific applications in RCTs, cohort and case-control studies.

- Visualisation

The visualisation of the results would be the same as or very similar to those presented for NNT or NEAR (Appendix A.7.1). A visual representation of the relations of the impact numbers to the population they are related to is shown in Figure 26.

Figure 26 Visual representation of the respective population to which each impact number relates (reproduced from [80])



- Assessability and accessibility

Given that it is made clear which population each impact number is describing in an analysis, the interpretation is straightforward (and appealing to many) with the same rigour as that of NNT. In real-life decision-making, the use of impact numbers may not be very suitable for formal regulatory decision-making but can give an idea of the “impact” of risk factors of interest in the population of interest.

A.7.3 Net efficacy adjusted for risk (NEAR)

Description

The net efficacy adjusted for risk is conceptually similar to "unqualified success" in the situation when marginal probabilities are not available. It builds a χ^2-like table tabulating the number benefit or not benefit from treatment against number with and without ADRs. Expected frequencies are calculated using the cells counts and odds ratios or relative risks are then calculated.

We suggest reading Boada (2009) [29] for an introduction and worked example, and Boada (2008) [28] as the core reference.

Evaluation

- Principle

The principle of the approach is logically sound for estimation of odds ratios or relative risks. It is unclear how NEAR increases transparency in the benefit-risk assessment. However, only the use of data from RCTs and observational studies including their meta-analysis results are recommended when applying NEAR. Statistical uncertainties of the point estimates are based on confidence intervals of OR and RR. The concept of expected frequency and OR/RR in the best case scenario can be easily understood.

- Features

The features are mostly similar to NNT (see Section A.7.1). It is unclear whether several benefits and risks can be included in deriving the final NEAR measure from the original paper [28] but the follow-up paper [29] demonstrates that only one of benefit and risk can be estimated in one NEAR measure. Multiple sources of evidence can be dealt with through meta-analysis [28]. NEAR has also been extended to take into account intention-to-treat or per protocol analysis [29].

- Visualisation

The organisation of the expected frequencies can be visualised in a 2 × 2 table as in Figure 27; and NEAR results can be visualised using a forest plot, as shown in Figure 28.

Figure 27 Theoretical distribution of results of a clinical trial in which two drugs have been studied [29]

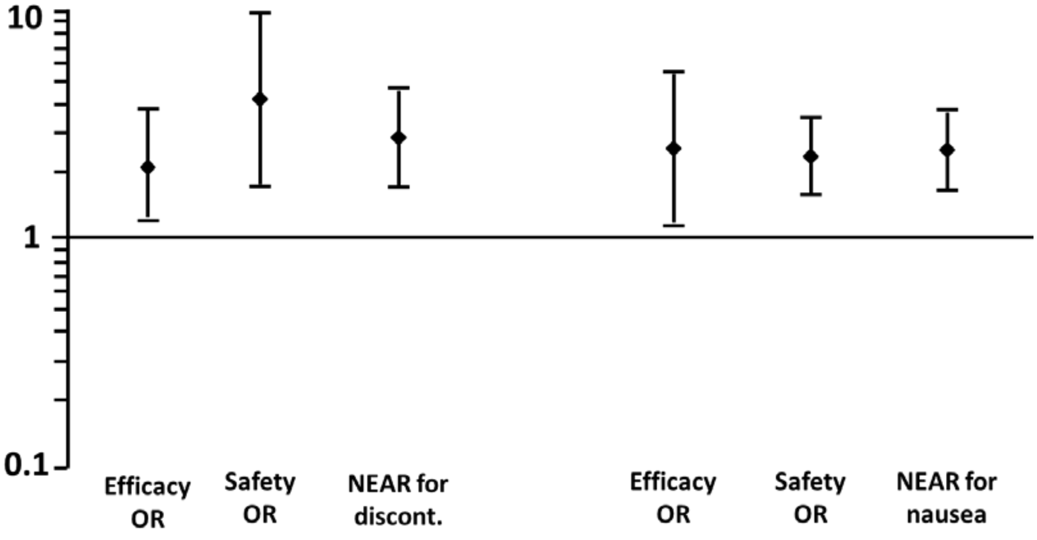
	Responders without ADRs	Other results	Total
Treatment A	a1	b1 + c1 + d1	n1
Treatment B	a2	b2 + c2 + d2	n2

In the second column, the expected frequencies for optimal results obtained with each drug are noted; in the third column, the sum of the remaining expected frequencies for each drug are noted, that is, patients who do not respond and moreover suffer ADRs plus those patients who respond and suffer ADRs plus those patients who do not respond and do not suffer ADRs. Now, let b1+c1+d1=S1 and b2+c2++d2=S2. Then, (a1*S2)/(a2*S1) expresses NEAR OR and (a1/n1)/(a2/n2) expresses NEAR RR. Finally, the CI for these new parameters may be calculated in the following manner:

$$\text{NEAR OR} \pm \text{CI95\%} = \text{NEAR OR} \times e^{\pm 1.96\sqrt{\frac{1}{a1}+\left(\frac{1}{a2}\right)+\left(\frac{1}{S1}\right)+\left(\frac{1}{S2}\right)}}$$

$$\text{NEAR RR} \pm \text{CI95\%} = \text{NEAR RR} \times e^{\pm 1.96\sqrt{\frac{1}{a1}-\left(\frac{1}{n1}\right)+\left(\frac{1}{a2}\right)-\left(\frac{1}{n2}\right)}}$$

Figure 28 Forest plot for NEAR in an application to a clinical trial comparing cabergoline versus bromocriptine for the treatment of hyperprolactinaemia (reproduced from [29])



On the left, intention-to-treat analysis is considered: traditional efficacy and safety OR together with NEAR OR, taking discontinuation as the ADR, are presented. On the right, per-protocol analysis is considered: traditional efficacy and safety OR together with NEAR OR, taking nausea as the ADR, are presented. Vertical lines represent 95%CIs.When the lower limit is > 1, the proband drug is preferable.

- Assessability and accessibility

These are similar to NNT in many respects (Appendix A.7.1). NEAR avoids the null point and the “signs” problem in the NNT approach being a ratio. However, NEAR OR and RR must be interpreted rigorously within the context of the problems and with great caution exercised when extrapolating NEAR results [29].

A.7.4 Minimum Clinical Efficacy (MCE) and relative values adjusted MCE (RV-MCE)

Description

Minimum clinical efficacy describes the minimal therapeutic benefit threshold at which a treatment is still worth administering by taking into account the natural characteristics of the disease in the untreated population [53]. MCE only compares two active treatments to be because the probability of benefits or risks in the untreated population is used as denominator i.e. $e_1 > \frac{q_1 - q_0}{p_0}$ where e_1 is the benefit (the "minimum clinical efficacy" required) in active treatment, p_k and q_k are probabilities of benefits and risks respectively in group k $(k: 0 = \text{control}, 1 = \text{active})$. As e_1 is calculated as $1 - p_1/p_0$, MCE reduced to $1/(p_1 - p_0)$ $< 1/(q_1 - q_0)$ which is simply NNT<NNH. MCE is analogous to the "minimal clinical difference" that is often specified for sample size calculations, with an extension to incorporate both benefits and risks into one measure. RV-MCE is MCE adjusted for utilities in the same way as RV-NNH (see Appendix A.7.1).

We suggest reading Holden (2003) [30] for more details and worked example.

Evaluation

- Principle

It bears heavy resemblance to absolute risk difference with an additional parameter in the denominator for rates in the untreated group. The use in benefit-risk assessment is achieved by simple rearrangement of the expressions in the inequality comparing the difference in the rates of benefits and the difference in the rates of risks. However, RV adjustment as in RV-NNH threatens the logical soundness of MCE.

- Features

MCE gives a combined threshold estimate based on one benefit and one risk, whilst RV-MCE allows several benefits and risks to be included. (RV-)MCE is used solely based on the point estimates (under standard circumstances) and unable to handle uncertainty in the measurements of benefits and risks [23]. Other features are similar to NNT (Appendix A.7.1).

- Visualisation

There is none specific to (RV-)MCE but other visual representations may be easily adapted.

- Assessability and accessibility

Whilst being very closely similar to NNT, MCE is not widely used in epidemiology and very rarely used for benefit-risk assessment [53] and its statistical properties have not been well-studied.

A.7.5 Maximum Acceptable Risk (MAR)

Description

Maximum acceptable risk is analogous to "willingness-to-pay" measure [31] and is a natural extension of SPM. MAR compares two options based on several benefits criteria being traded off by one risk of choice. It also has the same flavour as the MCE approach where the increased risk is accounted for an option having offset the benefits of using that option (see Appendix A.7.4).

We suggest reading Johnson (2009) [31] for more details and worked example.

Evaluation

- Principle

MAR uses SPM as the basis for data collection. The only difference is in the formulation of the utility function. MAR explicitly assume that benefits only occur when the risk does not i.e. the rates of risk occurring p is multiplied by risk utility and $1 - p$ is multiplied by the utilities of benefits. MAR of one treatment is subtracted from MAR from its comparator for direct comparison based on a set of criteria.

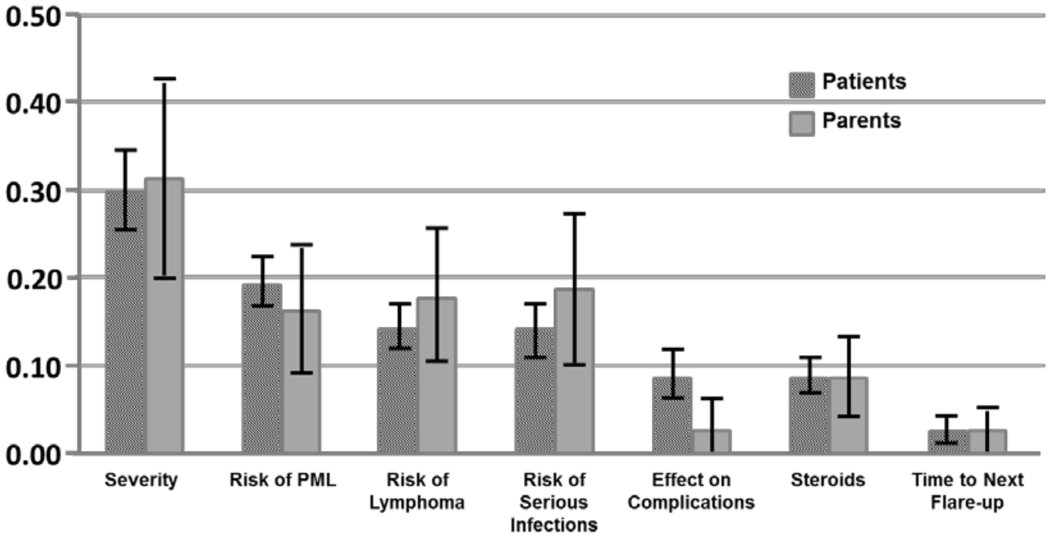
- Features

A slight dissimilarity to SPM, a MAR only allows one risk to be considered for a set of benefits criteria. In order to be comparable, the same set of benefits criteria must be used when estimating a MAR for another risk.

- Visualisation

A visual representation of the components of patient utilities is shown in Figure 29 for a worked example in the original paper [31].

Figure 29 The relative contribution of each treatment attribute to patient utility and 90% confidence intervals (reproduced from [31])



- Assessability and accessibility

These are the same as SPM (see Appendix A.11.1) – the only difference is that MAR is interpreted as the maximum acceptable risk for a specific AE to warrant a treatment over its comparator.

A.8 Quantitative health outcome indices

Authors: Nan Wang and Shahrul Mt-Isa

A.8.1 Quality-Adjusted Life-Years (QALY) and related health indices

Description

QALY is a measure of life time with quality of life incorporated into the measurement. For example the QALY of a subject who lives four years in QoL (quality of life) 0.75 has $QALY = \ 4 \times 0.75 = 3$. In general

$$QALY = \sum_{i=1}^{n} Q_i t_i$$

where i represents a specific health state, Q_i represents the QoL associated with health state *i*, and t_i represents the proportion of time spent in health state *i*. The QoLs, or health utilities, are numbers between 0 and 1 derived from health outcomes by agreed methods (for example rating scale). Since QoLs represent the balanced health outcomes in the period of time under concern, in benefit-risk context, QALY is already a measure combined both benefit criteria and risk criteria. So, QALY can be directly used in benefit-risk assessment for the comparison of different options. QALY outcomes in a benefit-risk assessment can also be separated into health improvement with QALY gain and health adverse impact with QALY loss [35]. The two QALYs can then be brought into INHB for benefit-risk assessment.

Following QALY, which appeared first in 1970s, there are other health index introduced as well such as DALY (disability-adjusted time, early 1990s) and HALE (health-adjusted life expectancy, early 1990s). DALY is a measure for year loss compared to the national life expectancy. It weighs the level of disability and counts both the year of loss due to disability and the year short to national life expectancy. For example a subject has a disability (weight level 0.3) for 10 years and dies at age 70, given the life expectancy 82, the DALY of the subject is $10 \times 0.3 + (82 - 70) \ = \ 15$. DALY can also act as a population measure defined by the summation of DALY of each individual in the population. HALE is simply a summary of QALY in a concerned population. In QALY, benefit and risk criteria affect QoL, in DALY, benefit and risk criteria affect disability weights. To use QALY and DALY or other health indices in benefit-risk assessment, one needs to know not only the benefit and risk effects of the options but also the time lengths of those effects in order to derive the health indices. Viewing from this point, we would say that health indices are suitable in diseases of chronic nature or cancer and oncology.

We suggest reading Weinstein (2009) [32] as introductory reference. A theoretical core reference and worked example can be found in Pliskin (1980) [81].

Evaluation

- Principle

Health indices such as QALY are widely used in health care studies for cost-effectiveness analysis. In benefit-risk assessment, establishing a 'common currency' for both benefit and risk is crucial in evaluating the benefit-risk balance, and health indices can take this role. For individual health index such as QALY, standard statistical analysis can be applied to estimate the benefit-risk balance and the difference of balances between different treatments. For population health indices, the uncertainty in parameters can be dealt with by Bayesian statistics or probability simulation as described in Appendix A.10.2.

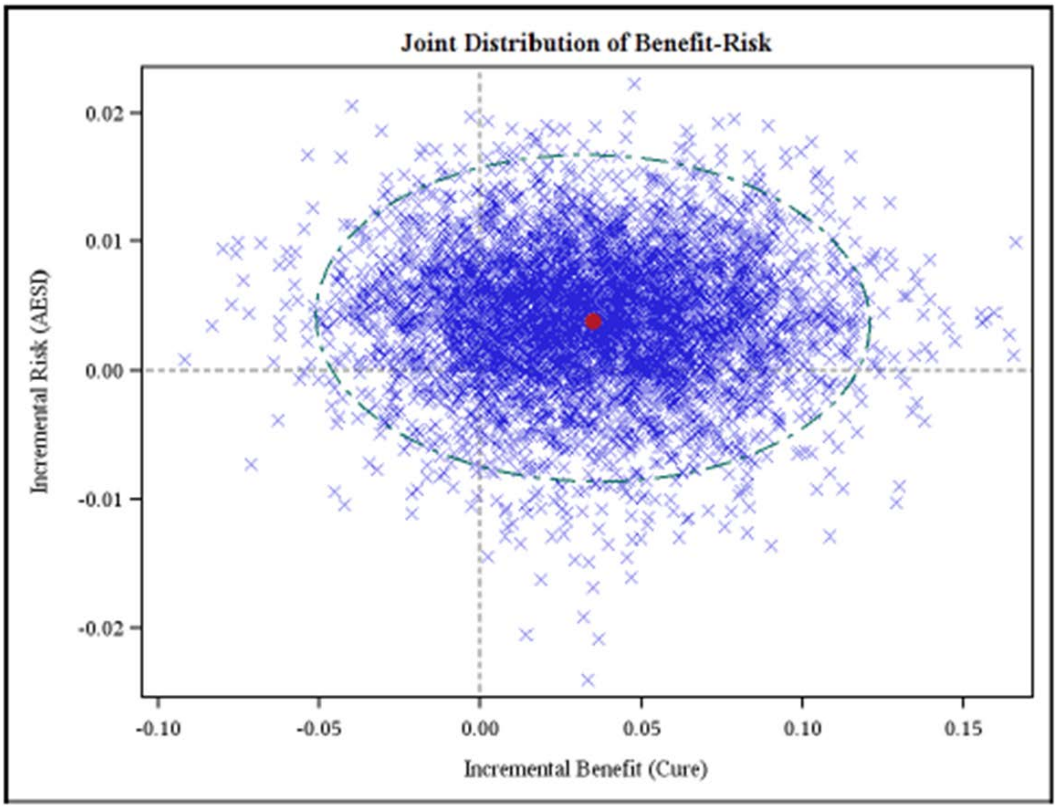
- Feature

With health indices, benefit and risk are integrated and the time dimension is included. A benefit-risk assessment with health index can accommodate multiple criteria and multiple options. For health index based benefit-risk analysis, in principle, a sensitivity analysis can be performed on the derivation of QoL, since weights or utilities of different criteria appear only in the derivation of QoL. Incorporating multiple sources of data for health index based analysis may need a meta-analysis given that the health index is derived in the same way in all sources.

- Visualisation

The incremental benefit QALY and incremental harm QALY can be presented on a XY-plane similar to that in benefit-risk threshold indices (Appendix A.7) and probabilistic simulation method (Appendix A.10.2). An example of scatter plot is shown below.

Figure 30 A scatter diagram of incremental risk (AESI) versus incremental benefit (Cure) of telithromycin relative to placebo in a simulation



Monte Carlo simulation in IMI-PROTECT telithromycin case study for ABS indication plotted on the risk–benefit plane: the incremental probability of AESI (Hepatic, Syncope, Visual Cardiac) versus. the incremental probability of Cure, with 95% confidence interval. The red dot mark the point estimate of BRR of telithromycin versus placebo

- Assessability and accessibility

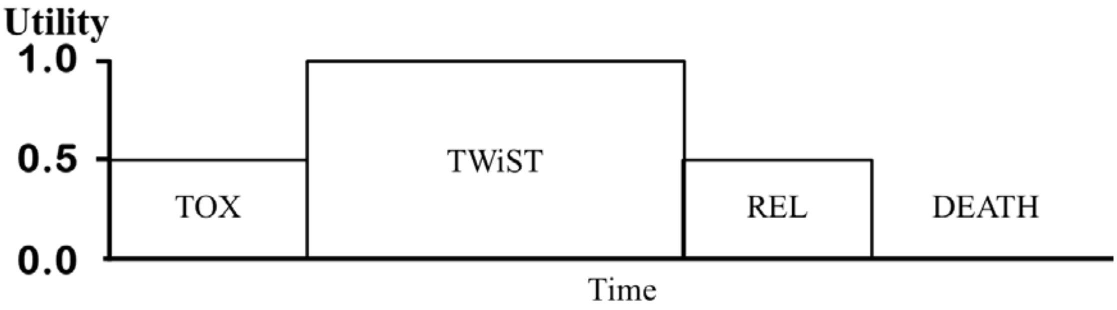
Health indices are widely used in health care studies and are easy to understand. In principle, stakeholders' preference in this approach is reflected from the derivation of QoL (weights or utility assigned to different criteria and different levels of each criterion). However the QoL usually is derived by standard way such as EQ-5D which may be considered as the average preference of population level. Therefore, there are some blames on health indices that they are 'risk-neutral'.

A.8.2 Quality adjusted Time Without Symptoms and Toxicity (Q-TWiST)

Description

Q-TWiST is an extension of QALY specifically developed for the application in cancer treatments based on discrete health states experienced by the patients. It was first proposed in breast cancer trials [82]. Q-TWiST is obtained by dividing survival time into discrete health states: TOX (time subject to toxicity effect), TWiST (time without toxicity and disease), and REL (time of relapse to death). The resultant time-utilities in the three health states are recombined using utility-weighted sum [33]. Mathematically, Q-TWiST is represented by $\text{Q-TWiST} = u_{\text{TOX}} \times \text{TOX} + \text{TWiST} + u_{\text{REL}} \times \text{REL}$. The components that make up Q-TWiST are illustrated in Figure 31.

Figure 31 The breakdown of survival time spent at different health states and the associated utilities for that state (reproduced from [33])



Clearly, Q-TWiST itself provides a summary measure incorporating both benefit and risk over time. Q-TWiST comparison of different treatments is naturally a benefit and risk comparison of different treatments [83].

We suggest reading Gelber (1995) [33] for introduction and worked example, and Goldhirsch (1989) [82] as core reference.

Evaluation

- Principle

Q-TWiST is individual measurement for each patient. Variability of Q-TWiST from different treatment or population can be handled by standard statistical analysis to provide confidence estimations. Utility in deriving Q-TWiST is judged by the disease and treatment applied. Q-TWiST can deal with any number of treatment options but seems confined to survival endpoints only.

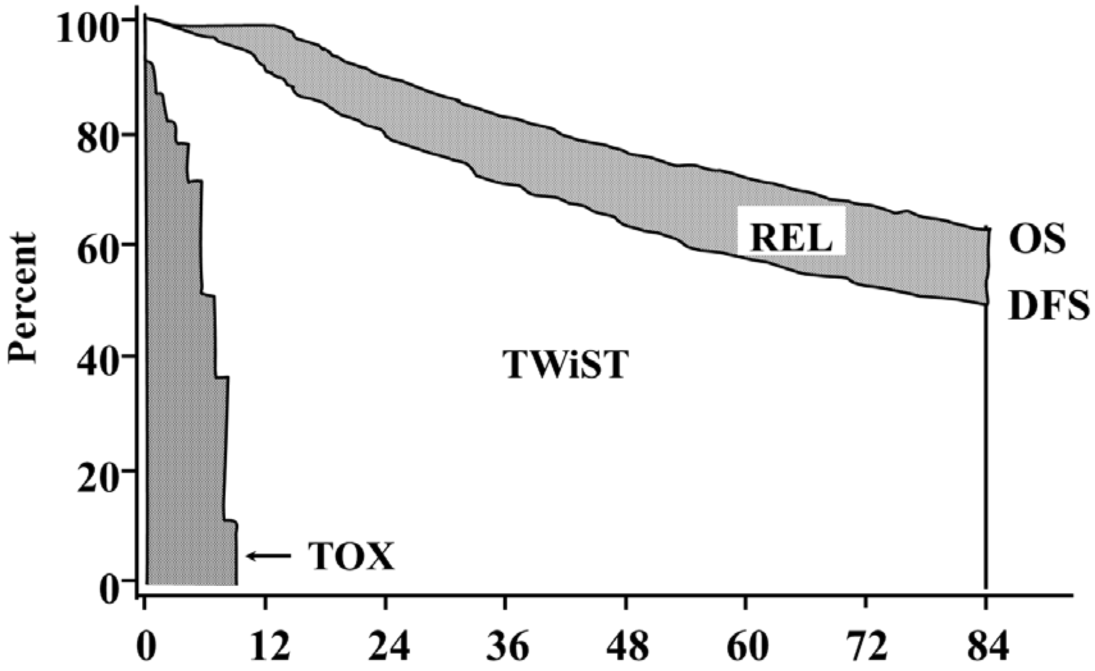
- Feature

With Q-TWiST, benefits and risks are naturally integrated and time dimension incorporated. Sensitivity analysis can be performed on choices of utility. Incorporating multiple sources of evidence using Q-TWiST certainly requires all sources of data to have the Q-TWiST derived in the same way. The analysis is then purely a statistical meta-analysis, similar to that with other health indices.

- Visualization

Currently Q-TWiST is conventionally shown by a stratified survial curve as below. This is however a graphical representation for one treatment (Figure 32). Graphical representation of the contrast between two treatments should still be considered.

Figure 32 Partitioned survival plot (reproduced from [33])



- Assessability and accessibility

The concept of Q-TWiST is easy to understand. Preference in this approach is reflected in the utility in Q-TWiST derivation. Although Q-TWiST was developed within the oncology domain, its application can be generalised to other diseases or medical conditions with similar characteristics, for example, Q-TWiST has been applied in measuring quality of life in patients with HIV infections [84], multiple sclerosis [85], and epilepsy [85].

A.9 Quantitative trade-off indices

Authors: Nan Wang and Shahrul Mt-Isa

A.9.1 Incremental Net Health Benefit (INHB)

Description

It is most commonly used with health indices like QALYs. The "incremental" change in risks is subtracted from the "incremental" change in benefits [35]. This implicitly assumes equal weights for benefits and risks, or requires common metrics for benefits and risks to be established before using INHB. The general form of INHB is the incremental net benefit (INB) [86], commonly used for economic evaluation, is not specific to health outcome indices but is not reviewed here as it is conceptually similar to INHB.

We suggest reading Garrison (2007) [35] as an introductory reference, Lynd (2010) [87] as core reference with worked example, and Minelli (2004) [88] for another worked example.

Evaluation

- Principle

As in the QALY (Appendix A.8.1), statistical uncertainty in benefit -risk balance can be dealt with by standard statistical inference if individual benefit and risk measurements are involved, or by simulation means if population benefit and risk parameters are involved.

- Feature

The benefit and risk are integrated in INHB. It can be extended to multiple benefit criteria and multiple risk criteria. Actually health indices are all derived by considering multiple factors affecting health states. INHB compares two options each time. Updating the model with new data may need a meta-analysis first, or a Bayesian type of updates.

- Visualisation

The type of Visualisation described in QALY and health indices section can be used here.

- Assessability and accessibility

INHB is easy to perform and understand. The acceptability and interpretability of the results depend on the health index used in INHB.

A.9.2 Benefit-Risk Ratio (BRR)

Description

BRR is a simple ratio measure (probably the simplest) of comparing benefit to risk analogous to relative risks. For a given benefit with rates p and a given risk with rates q, $\text{BRR} = p/q$. This is conceptually the expected “multiples” of benefit p per unit risk q. BRR is closely related to benefit-risk difference which looks at the “difference” of benefit from risk, instead of the ratio. Other benefit-risk indices framed these ideas in more defined contexts for use in benefit-risk assessments; including the health indices described in Appendix A.8.

We suggest reading Chuang-Stein (2008) [89] as an introductory reference, and Korting (1999) [90] as the core reference. Worked example is not necessary due to its simplicity but Payne (1975) [91] could be referred to if required.

Evaluation

- Principle

It is a very simple approach based on probabilities and is easy to understand. However, BRR is not transparent for benefit-risk assessment when used in its simplest form where no weighting or scoring is involved. The origin of the BRR generally used and as understood today can perhaps be traced back to the work of benefits and risks in radiology [91]. In a benefit-risk assessment, an equilibrium point at which the benefit and risk are considered equal should be established. The equilibrium point is then used to determine whether benefits outweigh risks, insufficient to conclude, or that risks outweigh benefits [89] – this implicitly incorporates weighting into the expression. This exercise is analogous to establishing ‘regions of equivalence’ in Bayesian analysis.

- Features

The features of BRR are very similar to those of NNT (Appendix A.7.1) and can only deal with one benefit and one risk at a time.

- Visualisation

There is none.

- Assessability and accessibility

BRR is only acceptable and interpretable if the assumption of equal weightings for benefit and risk hold. It provides a straightforward comparison of benefit and risk but may be unreliable when used in its simplest form. However, it is important to understand that BRR can be derived from good statistical models based on high quality evidence data or through simulations. In these cases, the use of BRR is more attractive and meaningful in the context of drug benefit-risk assessment. Taking ratios of other benefit and risk metrics for example the NNT to NNH could also be regarded as BRR [23;89], where the strengths and weaknesses of the unitary metrics used are also inherited.

A.9.3 Global Benefit-Risk (GBR)

Description

Three different measures have been proposed in the primary paper to estimate the trade-off of benefits and risks [36]. Decision-makers applying GBR are presented with the choices of one linear measure and two ratio measures depend on the type of data available and inference to be made. Data are collapsed into proportions π in $j = 5$ ordered categories (Table 22), which then determine their weights w (so do the disease, symptoms, and treatment). Individual patient's score using the one of appropriate three measures are then compared for different treatment using standard statistical techniques.

Table 22 GBR classifications of individual outcomes [11]

Benefit with no ADR (π_1)	Benefit with ADR(π_2)
No benefit and no ADR (π_3)	No benefit with ADR (π_4)
ADR leading to serious complications or withdrawal from the study (π_5)	

The three measures $m, r_1$and r_2 with constants e and f are:

(a) Linear score:

$$m = \sum_{i=1}^{2} w_i \pi_i - \sum_{j=3}^{5} w_j \pi_j$$

(b) Ratio score

$$r_1 = \frac{\left(\sum_{i=1}^{2}(w_i \pi_i)\right)^e}{\sum_{j=3}^{5}(w_j \pi_j)}, e \geq 0$$

(c) Composed ratio score

$$r_2 = \frac{w_1 \pi_1}{w_5 \pi_5} \left(\frac{w_2 \pi_2}{w_3 \pi_3 + w_4 \pi_4} \right)^f, f \geq 0$$

The linear score m is simple and most useful when comparing two treatments but lacks intuitive for assessing benefit-risk balance for a single treatment. The ratio score r_1 provides a more intuitive alternative in this situation, having the same interpretation as BRR (see Section A.9.1). The composed ratio score r_2 is essentially the product of two ratios where the most important benefit $\pi_1$and the most important risk π_5 are multiplied with the other benefit and risk outcomes. The key ingredient to GBR measures is the choice of constant in the expression of benefit-risk to put the categories on the same scale, and the key paper [36] provides more detailed descriptions of the measures.

We suggest reading Chuang-Stein (2008) [89] as introductory reference, and Chuang-Stein (1991) [36] as core reference and for worked example.

Evaluation

- Principle

The mathematical principle of GBR is logically sound and simple as it aims to put measurements in different dimensions on the same scale to perform benefit-risk trading. As these are metric indices calculated for each individual, any appropriate statistical modelling techniques could be used to estimate statistical uncertainties around the GBR point estimates. Value judgments in models using GBR measures would implicitly be in the choice of what constitute benefit/risk and in the choice of constant in the expressions. Allowing individuals to choose their own weights also contribute to this.

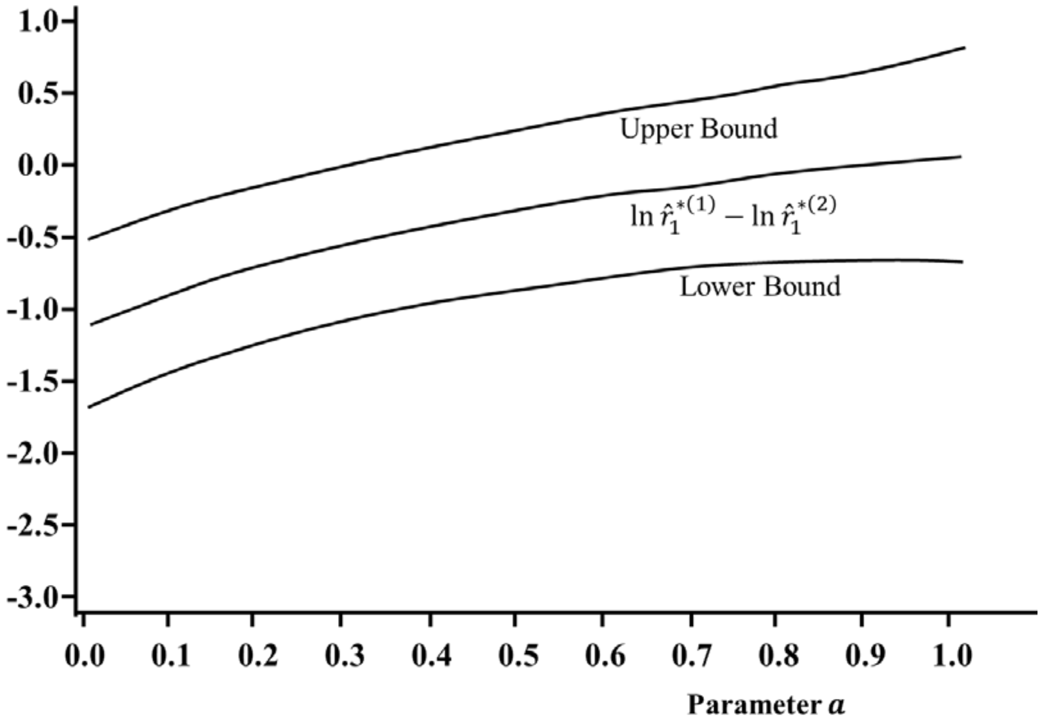
- Features

GBR gives integrated benefit-risk measures for benefit-risk assessment. Multiple benefits and risks are not differentiated using these measures. They are only regarded as collective criteria. Sensitivity analysis of GBR is suggested by varying the choice of the constants over an acceptable range.

- Visualisation

A snapshot of data for GBR analysis can be shown in a table representing the outcomes of the patients (Table 22 above). A standard line graph with CI has been suggested for sensitivity analysis by varying the constant parameter as illustrated in Figure 33.

Figure 33 Estimated difference in log risks and 95% confidence intervals by varying constant α for one of the GBR measures proposed (produced from [36])



- Assessability and accessibility

The approach is very specific to the three functional forms. Therefore knowledge of which functional form and the underlying assumptions used is essential when making comparisons or when generalising the results. Although it may help with decision-making, GBR measures do not explicitly distinguish the extent of severity or seriousness of adverse events even when individual data are available.

A.9.4 Principle of three and its modifications

Description

This approach employs three criteria – "disease", "effectiveness", and "ADRs" – for benefit risk assessment [37]. Each criterion has three attributes namely "seriousness", "duration", and "incidence"; and each attribute is scored in three levels 1, 2, and 3. Decision maker compares the total scores in disease, effectiveness and ADRs criteria to make the decision. This approach has been modified by introducing the scoring system for benefit and risk separately: (1) $\text{benefit score} = \text{cure rate} \times \text{seriousness} \times \text{chronicity/duration};$ and (2) $\text{risk score} = \text{incidence} \times \text{seriousness} \times \text{duration}$ [38]. In the case where there are multiple risk outcomes involved in a benefit-risk assessment, the average risk score of all risk outcomes is used as the risk score.

We suggest reading Mussen (2009) [65] for introduction and worked example, and Edwards (1996) [37] as core reference.

Evaluation

- Principle

This approach is a simple multi-criteria model, in that it pre-specifies the number and level of criteria, but has no restriction on number of options. The way of scoring is rough and it does not take into account the relative importance of different criteria. Statistical uncertainty in input data (variability, confidence estimations etc.) is difficult to bring into the model due to its rough scoring system.

- Feature

With the modification, this approach allows multiple risk criteria. However, all of them are taken as equally important. Risk and benefit are dealt with separately, not integrated. Updating the model with new data or accommodating multiple sources of data may need a meta-analysis first. The appealing feature of this approach is its simplicity. However the simplicity also prevents this approach from application in more complex situations.

- Visualisation

Since the approach is simple, a table presentation very clearly shows of the assessment (Table 23).

Table 23 An example of an application of principle of three to Felbamate treatment for epilepsy [92]

	Disease	Effectiveness	Dominant ADR
Seriousness	3	3	3
Duration	3	3	2
Incidence	1	0	2
Total	7	6	7

- Assessability and accessibility

The principle of three is easy to perform and understand. The scores in this approach are supposed to be evidence based. It does not take into account of stakeholder's 'preference'. It would be useful for a regulator or a pharmaceutical company to make initial assessments. A full and through assessment however should adopt more comprehensive approaches.

A.9.5 Transparent Uniform Risk-Benefit Overview (TURBO)

Description

TURBO is a very simple multi-criteria decision-making approach which gives an overview of benefits and risks in their respective dimension on a grid, which was described in Appendix E of CIOMS working group IV report on benefits and risks balance [38]. The framework given by TURBO is simple: (1) score the primary benefit on a scale 1-5; (2) score the ancillary benefit on a scale 1-2; (3) score the most important risk on a scale 1-5; (4) score additional risk on a scale 0-2; (5) add the two benefits ("B" factor) and risks ("R" factor) in their respective dimensions to obtain benefit and risk scores; (6) plot the two scores against each other to produce the "T" score. TURBO has never been applied in any real decision problems and is only an idea informally generated. A search in Google and Google Scholar (search last performed on 04 April 2011) using combination of keywords "TURBO", "benefit", "risk", "Amery", "CIOMS", "case studies", "example" did not yield any relevant results. We acknowledge that it is unfair to appraise TURBO as it can be seen as an approach "under development" (we are not aware of any development). But since TURBO has gained some attentions through many previous reviews, it is included here to point out that the concept is weak and should not be further considered in the future.

We suggest reading CIOMS Working Group IV report (1998) [38] as an introductory and core reference, and Mussen (2009) [92] for worked example.

Evaluation

- Principle

The approach assumes only two benefit criteria and two risk criteria really matter, corresponding to the primary and secondary benefits and risks respectively. The second criterion in each dimension is regarded as the correction factor to the first and scored on a shorter scale than the first. Although the scores for each benefit and risk may be determined prior to making any judgment based on the intrinsic characteristics of the disease, drug and population; the T-score is defined ad hoc. It is too simple to be transparent or of any use in drug benefit-risk decision-making.

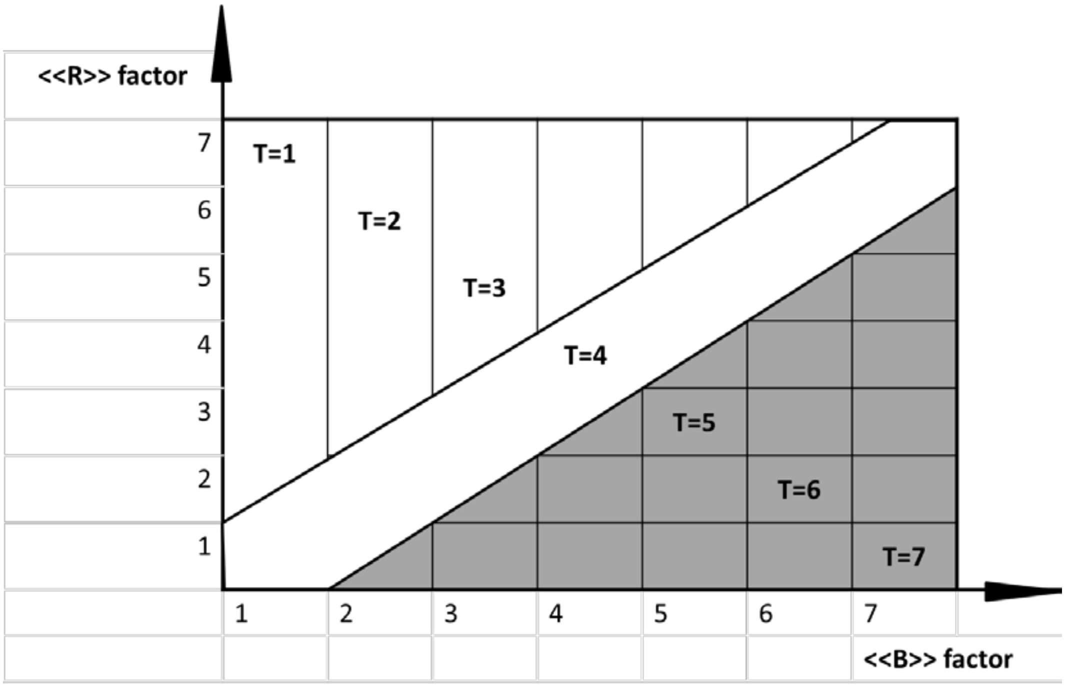
- Features

Although this is simple and quite intuitive, its restrictions to two benefits and two risks limit its usefulness for formal benefit-risk assessment. The proposed secondary (ancillary) benefit is unclear as to what it should represent. The idea of accounting for two most important risks is also unjustified as the risk profile of a treatment cannot always be solely driven by them.

- Visualisation

A TURBO grid is illustrated in Figure 34 where T-scores are defined ad hoc.

Figure 34 Intrinsic benefit-risk balance: the TURBO diagram (reproduced from [38])



- Assessability and accessibility

TURBO is too simplistic to be considered for the application in drug benefit-risk decision-making, and should not be considered further.

A.9.6 Beckmann Model

Description

This approach, proposed in Beckmann (1999), is also a simple multi-criteria model [39]. A special feature of this approach is that the quality of data (evidence) is taken as a component in benefit assessment. Benefit includes three aspects: efficacy, response rate, evidence; risk includes also three aspects: seriousness, incidence, evidence. The scoring system for benefits is:

$$\text{benefit score} = \text{efficacy} \times \text{response rate} \times \text{evidence}$$

And the scoring system for risk is:

$$\text{risk score} = \text{seriousness} \times \text{incidence} \times \text{evidence}$$

As in principle of three (see Section A.9.1), efficacy, seriousness, incidence are classified into different categories (for example the seriousness is according to WHO collaborating Centre for International Drug Monitoring 2004) and scores are based on this classification. The evidence here is classified according to the sources where the data come from e.g. RCT, observational studies, and case reports.

We suggest reading Mussen (2009) [92] for introduction and worked example, and Beckmann (1999) [39] as core reference.

Evaluation

- Principle

This approach does not take into account the relative importance of benefit and risk criteria and therefore does not integrate benefit and risk. This model is not ready to incorporate statistical uncertainty in input data into analysis due to its category scoring system.

- Feature

This approach deals with benefits and risks separately without integrating them into a single measure. The only weight here is driven by the 'evidence' parameters, and does not take into account the relative importance across criteria. The total risk of a drug, for example, is then the sum of evidence-weighted scores of those individual risks from different sources, as, evidence is the special feature of this approach.

- Visualisation

Since the approach is simple, a table presentation like that in principle of three (Appendix A.9.1) can show clearly the assessment.

- Assessability and accessibility

Analysis by this approach is easy to perform. It does not take into account stakeholder's preference. It would be useful only for a regulator or a pharmaceutical company to make initial assessments.

A.10 Estimation techniques

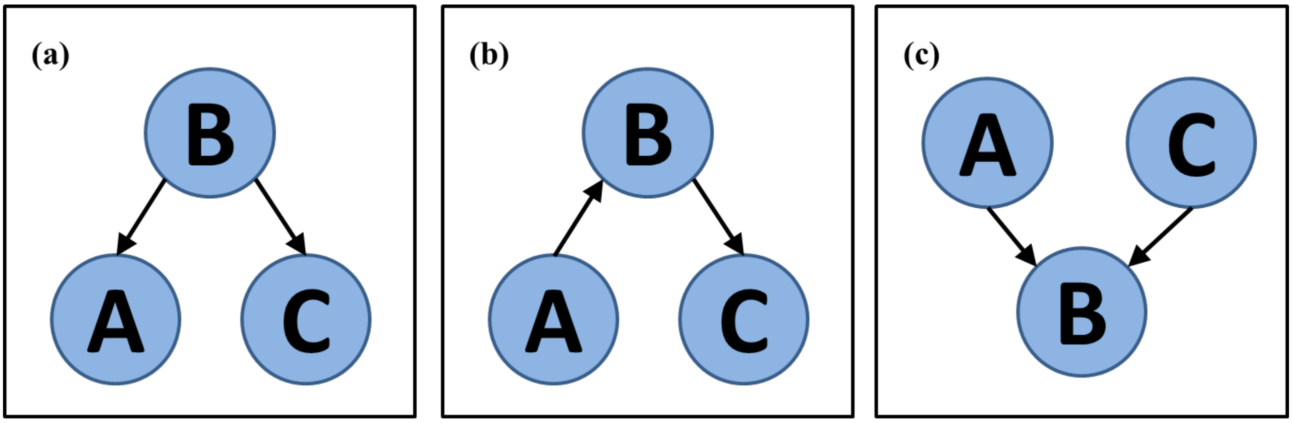
Authors: Shahrul Mt-Isa, Nan Wang, and Davide Luciani

A.10.1 Directed Acyclic Graphs (DAGs)

Description

Directed Acyclic Graphs (DAGs) are graphical structures where a collection of nodes is variably connected through directed edges (arrows). All nodes from which an arrow reaches another node are called "parents" of the latter. Since nodes cannot be reached by paths of edges by following their directions, these graphs are called "acyclic". Bayes Networks (BNs) are regarded as the prototype models based on DAGs. Therein, a DAG represents which random variables are directly associated, whereas variables are defined on a finite set of discrete values. Specifically, a DAG allows to establish when observing a variable B makes two variables A and C independent, that is, when A is independent of C conditionally on B (Figure 35a and Figure 35b), as well as when observing B makes A and C dependent (Figure 35c). So portrayed, the conditional independences allow to decompose the information about the quantitative strength of associations into distinct probability distributions, to specify as Conditional Probabilities Tables (CPTs) for each node given its parents [93].

Figure 35 Boxes(a), (b) and(c) portray three fundamental ways by which conditional independences can be read off from a DAG



We suggest reading Darwiche (2009) [94] and Jensen (1996) [95] as introductory references, Pearl (1988) [93] as core reference, and Jensen (1996) [95] and Jensen (2007) [96] for worked examples.

Evaluation

- Principle

Whenever a DAG is consistent with the conditional independences among variables, the joint probability distribution can be computed as the product of CPTs $p(X_v | \, pa(X_v))$:

$$p(V) = \sum\nolimits_{veV} p(X_v | \, pa(X_v))$$

where V is the set of nodes in the graph, v is the index of one node X in V, and $pa(X_v)$ represents the parents of the node X_v. In turn, the joint probability distribution allows to answer to any query about the uncertainty of any subset of variables conditionally on an observed subset of other variables (belief updating) [93].

- Features

The graphical structure of BNs may provide large domains with a compact probabilistic representation. However, the CPTs are often difficult to elicit from the parameters reported in the literature or from the knowledge of experts [97]. Notwithstanding, the DAG structure can be exploited to facilitate the estimate of the quantitative BN component from data. Few extensions allow continuous random variables being treated without discretisation and the DAG representing one unit of observation being meant as replicated for all available observations [98]. Then, MCMC methods can be applied to learn the distribution of the quantitative parameters involved in even complex hierarchical models, like those arising from the need to account for ADRs correlation within a body system [99] or for differences on how data are generated. The latter is often the case when uncertainties on decision outcomes must be derived from different sources of evidence, likewise when results based on animal studies or on healthy volunteers may influence the future development of a drug [100], or when spontaneous case reports are exploited to support further investigation on the drug safety profile. Although these methods have been recently introduced in the medical literature under the novel label of “network meta-analysis” [101;102], they should be regarded as part of the traditional empirical Bayes approach to hierarchical models [103]. BNs may also support personalized medical decisions on several drug options, taking into account of a possibly large set of patient characteristics. This can be done by simulating the impact of each contemplated therapy on the variables representing the benefits and the risks of interest, after the BN is entered with the observations that are available for an individual case. MACA may then be exploited to summarize the balance between risk and benefits in one single measure. Influence diagrams (IDs) extend BNs to incorporate decision nodes and utility nodes in order to avoid the need to iterate a simulation for each decision option. However, variables representing risk and benefits have to be linked to utility nodes after having been preventively defined on the same quantitative scale [96].

- Visualisation

The graphical feature of a DAG (Figure 35) may be of help in communicating the characteristics of the underlying decision domain by highlighting how differences in the eligibility criteria or in other study designs features might bias the predictions of the impact of decisions on the outcomes of interest. However, some expertise is needed in the interpretation of arrows, as their direction represents how the joint probability distribution is marginalized over the variables, not necessarily information about causal relations.

- Assessability and accessibility

Conditional independence among variables may be assessed by means of statistical tests applied on data, like in log-linear regression analysis [104]. However, conditional independencies could be also anticipated from knowledge on how the variables are causally related, especially when such knowledge is available from experimental study designs or from information about the temporal order of events [105]. Several software has been developed to support decision based on inference performed by BNs and IDs, and some of them is free of charge (for instance, Genie can be downloaded from http://genie.sis.pitt.edu/). WinBUGS and the R statistical software package JAGs are now popular tools exploiting graphical models in the estimation of uncertainty parameters from data (the relevant web sites are, respectively, http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml and http://www-ice.iarc.fr/~martyn/software/jags/).

A.10.2 Probabilistic simulation method (PSM)

Description

Probabilistic simulation is not a method for benefit and risk analysis, but it can go with most of the quantitative benefit and risk models to explore the impact of uncertainty in input data on the final benefit-risk balance. It is either a Monte-Carlo simulation (a Bayesian type approach) or a re-sampling from original data (if individual data available). It is a good means to understand the uncertainty in final benefit-risk balance [106].

We suggest reading Lynd (2004) [106] and van Staa (2008) [46] as for more details and worked examples.

Evaluation

- Principle

Monte-Carlo simulations and re-sampling methods are all sound methods in statistical inferences. With probabilistic simulations, the distribution of the benefit-risk balance will be explored and one can assess the major possibility, as well as the chance of worse cases.

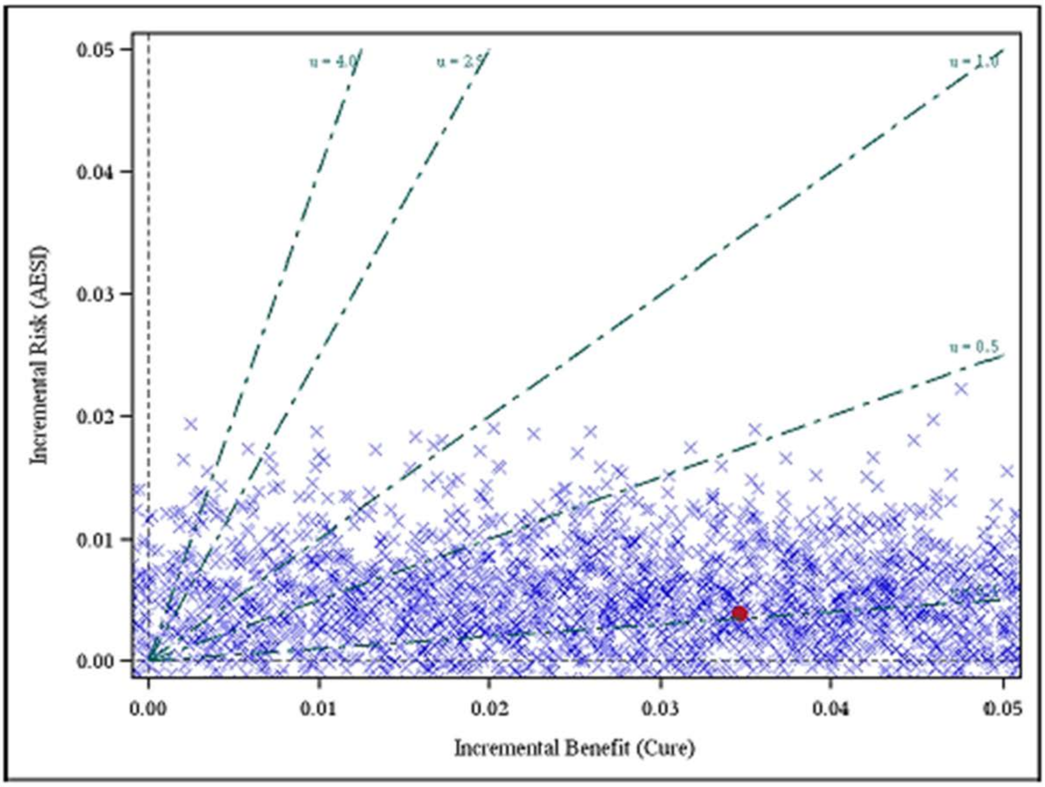
- Feature

Probabilistic simulation can be applied in any type of data, for summary results and parameter estimations, Bayesian type approach can be applied, for individual data, re-sampling methods can be applied. The application of probabilistic simulation has no restrictions on number of criteria and number of actions if they are independent or have known correlation structures.

- Visualisation

Probabilistic simulation can be presented graphically by either scatter plots or as population plots (box-plot, say). The following Figure 36 shows the distribution of risk increment and benefit increment of a new treatment over the standard treatment simulated from posterior distributions of risk increment and benefit increment.

Figure 36 Incremental risk versus incremental benefit and varying acceptability threshold obtained by probabilistic simulation



- Assessability and accessibility

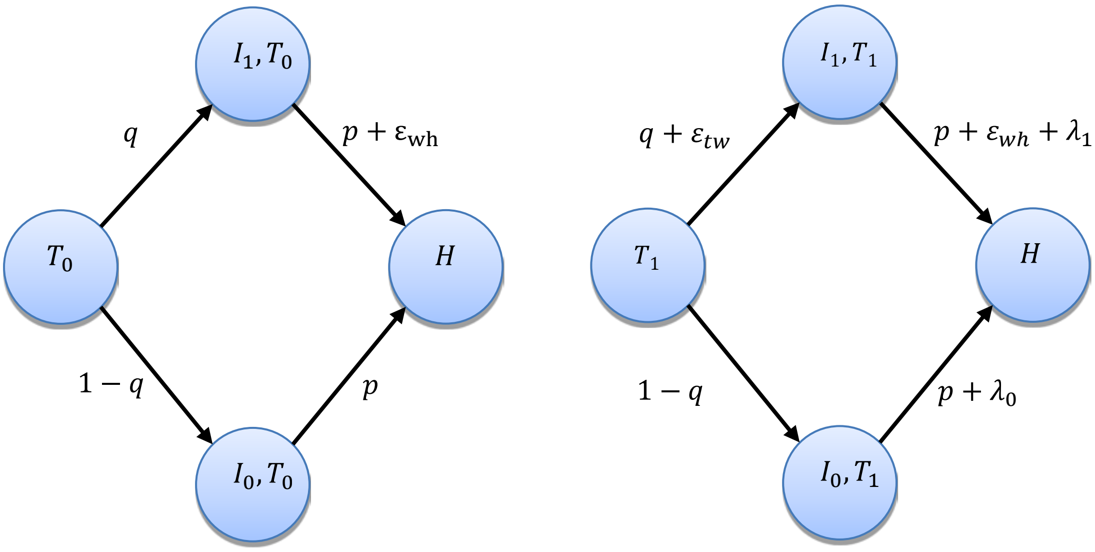
Probabilistic simulation nowadays is not difficult to perform with statistical software. The acceptability and interpretability of the results depend on the models used together with probabilistic simulation since probabilistic simulation itself is not a method for risk and benefit analysis.

A.10.3 Confidence Profile Method (CPM)

Description

The confidence profile method is intended to be applied to conventional meta-analysis or to multiparameter evidence synthesis [107].It is based on mathematical formulations of explicit forms of conditional probabilities, where evidence are structured as “chain of evidence”. Where direct evidence is available, single link chains are specified; otherwise multiple link chains are derived to connect the links (Figure 37). The core literature [44] presents the equations to be used when applying CPM together with the discussion of their operational circumstances. In particular, CPM strongly emphasises basic and functional bias adjustments. Metric indices such as rates difference (inverse of NNT) and INHB (Appendix A.9.1) are commonly used to quantify benefit-risk trade-offs, but other suitable indices are also valid. Despite the generality and the power of CPM, the approach is not as well known among statisticians and epidemiologists [107].

Figure 37 Illustration of a two-link chain showing derivation of the formula for multiple (two) link chains (reproduced from [44])



We suggest reading Ades (2006) [107] as introductory reference, Eddy (1989) [44] as core reference, and Eddy (1988) [108] for worked example.

Evaluation

- Principle

The principles of CPM are mathematically exhaustive. Evidence data are structured in chains that link the pieces of information together, thus logically sound for benefit-risk assessment. However, this can be threatened if the parameters of the underlying model – e.g. data, likelihood, dependence, and bias assumptions – are not properly specified. The equations for various formulations and circumstances are available as guidance [44]. Transparency is increased through the five basic application steps: (1) define options, the setting of application, and any factors that affect the options (collectively known as “circumstances of interest); (2) describe chains that relate options’ performance to the occurrence of the outcomes; (3) derive probability distribution for each link in each chain for the effect of options on outcomes, and then combine links within chains; (4) combine separate probability distributions across chains; and (5) reparameterise the probability distributions when the evidence of comparisons are indirect using steps (1) to (4) for each pairwise direct evidence and convolve the probability distributions. Statistical uncertainties are propagated from the likelihood functions and prior distributions (in Bayesian context), whilst other sources of uncertainties including biases as specified as stochastic parameters. CPM tries to avoid value judgments by providing formal framework for breaking the problems into parts; targeting intuitively accessible elements where empirical evidence or practical experience are often available; allowing uncertainty parameters to be expressed as probability distributions, making assumptions and judgments explicit; allowing collective judgments to be combined in a probability distributions; and allowing value of additional information about a parameter to be estimated.

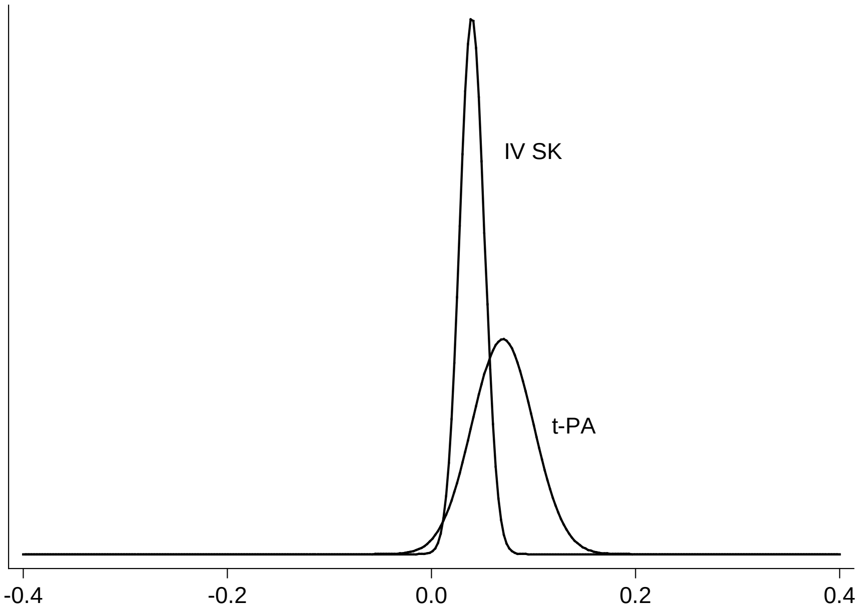
- Features

Numerical and visual representation of benefit-risk profiles are dependent on the choice of statistical model fitted. CPM can deal with multiple benefit-risk criteria and multiple sources of evidence through model parameterisation. CPM naturally allows sensitivity analysis through changes in model assumptions and specifications. The central idea of CPM is Bayesian thus ready to formally update new data and changes in assumptions.

- Visualisation

Visualisation of CPM results may vary depends on the requirements and the choice of metric indices used. Posterior distributions plots have been used to show the difference in probabilities depicting benefit-risk trade-offs (Figure 38).

Figure 38 Probability distributions for the effects of two thrombolitic agents compared with conventional care: tissue-type plasminogen activator (t-PA) and intravenous streptokinase (IV SK) (reproduced from [44])



- Assessability and accessibility

The parameters in CPM are carefully specified using probability distributions. Whilst these are conceptually acceptable and interpretable, in practice they might not be so straightforward. The results themselves, if specified correctly, are acceptable and should be easy to interpret but are still heavily dependent on the complexity of the model and the choice of benefit-risk metric indices used. CPM has close similarities to SMAA (Section A.6.6) and is a special case of CDS (Appendix A.10.5), thus shares some of their strengths and weaknesses as well as practicality of application in real-life decision making.

A.10.4 Indirect treatment comparison (ITC) and mixed treatment comparison (MTC)

Description

The origin of indirect treatment comparison (ITC) and mixed treatment comparison (MTC) can be traced back to CPM literature, with which they share the same principles (Appendix A.10.3). ITC specifically compares two treatments 1 and 2 where direct evidence is unavailable [101]. ITC exploits the networks of evidence to link the pieces of evidence when they have been compared directly to another same treatment, usually to placebo. The effect of comparison of treatment 1 against placebo is δ_{10} and for treatment 2 against placebo is δ_{20}, and therefore ITC estimates the difference between treatment 1 to 2 to be $\delta_{12} = \delta_{10} - \delta_{20}$.with variance $\mathrm{Var}(\delta_{12}) = \mathrm{Var}(\delta_{10}) + \mathrm{Var}(\delta_{20})$. The estimated $\mathrm{Var}(\delta_{12})$ associated with using indirect evidence is therefore larger than the variance if there were direct evidence of treatment 1 against 2, which is captured here.

MTC is a generalisation of ITC in a more complex network of evidence where both direct and indirect comparisons are available [101;109]. The concept of comparison networks for ITC also allows MTC [101] but is restricted to using evidence from trials with only two treatment arms. It is not uncommon for a particular trial to assess more than two treatments which leads to the development of K-comparison $(K > 2)$ MTC under the Bayesian hierarchical modelling framework [109]. MTC serves two purposes: to strengthen the inference on relative treatment effects by including both direct and indirect evidence; and to facilitate simultaneous inference for all treatments [109].

We suggest reading Lumley (2002) [101] and Lu (2004) [109] as introductory and core references with worked examples. Nixon (2007) [110] also provide a good worked example.

Evaluation

- Principle

The principles of ITC and MTC are based on probabilities and meta-analysis like the CPM (Appendix A.10.3) albeit slightly more general. ITC/MTC offer increased transparency when used in decision-making in terms of clarifying the sources of evidence, bias, and uncertainties through data structuring according to each piece of evidence. The comparison networks provide a visual way of representing the links and missing links between evidence pieces which would assist in the modelling of benefit-risk balance. The classical ITC method produces statistical uncertainties around the mean difference using the standard confidence intervals calculations, or otherwise the Bayesian version of ITC and MTC posterior estimates would inherit the uncertainties from the priors and likelihood. Value judgments in the input parameters for the modelling are not addressed in ITC/MTC literature being estimation techniques rather than benefit-risk decision-making approach. However, having well-informed clinical judgment is advisable to determine the suitability of the evidence being assessed [109]. This is directly associated with one of the main concerns in ITC/MTC – there is greater bias in using indirect comparisons than direct comparisons. Another concern is the greater uncertainties involved in using indirect evidence, which should be dealt with carefully modelling the correlation structure of the evidence's parameters.

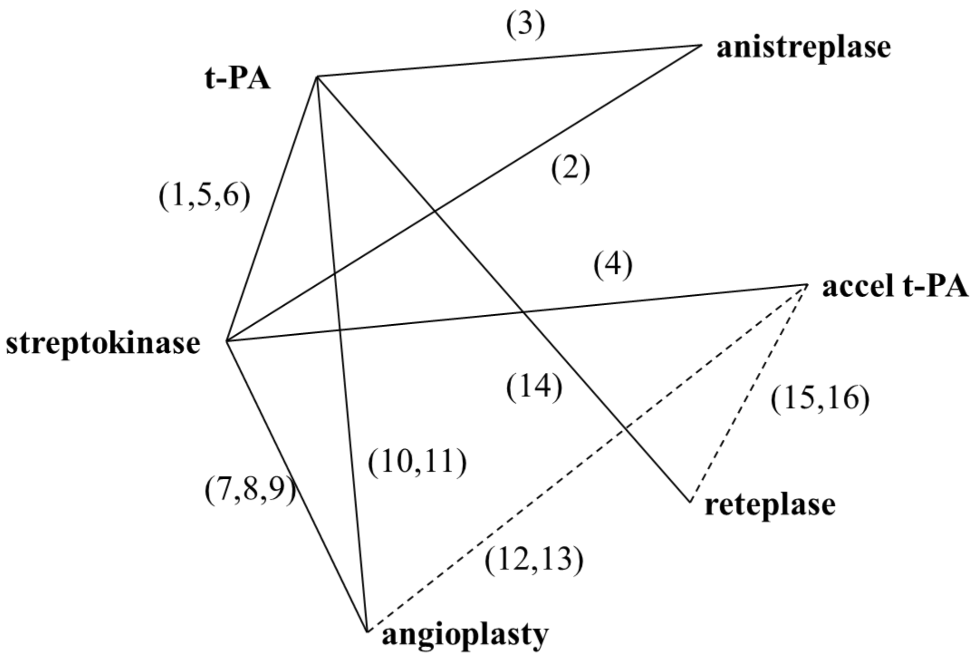
- Features

ITC/MTC "borrow strengths" from multiple trials with direct and indirect evidence to assess benefits and risks of different treatments, where multiple criteria of benefits and risks as well as multiple treatment options can also be taken into consideration and simultaneously estimated. ITC/MTC depends on the actual measure used in the models in order to describe the benefit-risk profile numerically or visually – most commonly the measures are probability-based for example odds ratio and rates difference. Sensitivity analyses are performed on the variance parameters.

- Visualisation

Comparison networks that were suggested in defining the structure of ITC serves as a visual aid to setting up a model which could also help to increase its transparency. This is illustrated in Figure 39. Visualisation of ITC and MTC results are dependent on the choice of models, therefore any suitable visual representations could be used.

Figure 39 Network of comparisons for thrombolytics and angioplasty in acute myocardial infarction [101]



The dashed lines are the two comparisons to be estimated and the numbers beside each line refer to the trials' identifier in Table II of Lumley (2002) [101] that address each comparison.

- Assessability and accessibility

The parameters and results from ITC and MTC analysis should be simple to be understood without in depth technical knowledge of probability theory. However, variance modelling requires greater level of statistical understanding. The flexibility in the application of ITC and MTC would be suitable for many stakeholders, especially pharmaceutical companies, healthcare providers, and regulators as many issues that are related to decision making for example bias and uncertainty adjustments can be addressed. ITC and MTC thus provide decision-makers with a general and flexible technique to assess benefits and risks to support decision making when used appropriately.

A.10.5 Cross-Design Synthesis (CDS)

Description

Randomised clinical trials data and clinical databases are combined in the cross-design synthesis to capture the strengths of the complementary study designs whilst minimising their weaknesses. CDS was developed based on the principles of meta-analysis, and on the 'focussed assessment' of potential biases from study designs' weaknesses [45]. The focussed assessments aim to adjust results of individual studies by correcting for known biases using statistical adjustments, and to devise benefit-risk models that would minimise the impact of unaccounted biases [45]. The idea of cross-design synthesis has also been extended to combine evidence from epidemiological studies and toxicological studies using Bayesian analysis [111].

We suggest reading Droitcour (1993) [45] for introduction, GAO/PEMD (1992) [112] as core reference, and Peters (2005) [111] for worked example.

Evaluation

- Principle

The central idea in CDS is the "extrapolation to empty cell" where available benefits and risks evidence from one population are used to predict the benefits and risks in a slightly different population ('an empty cell') by assuming proportionality in the effects measure. This "empty cell" is a population with both characteristics from the two study designs combined. Because the explicit modelling strategy is not addressed in CDS, it is naturally classified as a benefit-risk estimation technique. There are four main tasks in the cross-design synthesis approach: (1) assess the generalisability of existing randomised studies; (2) assess comparison bias; adjust individual study's results; and (4) combine results within and across design categories. The tasks were explained in details in the core CDS literature [112]. Similar to meta-analysis, uncertainties are modelled and are dependent on the data. These similarities make CDS easily understood by those who understand meta-analysis. Investigators' value judgments play important roles in determining the "value" of and which datasets to be combined. However, the elaborate tasks and over-reliance on investigators' judgments in CDS could potentially overlook inappropriate data pooling and giving false impression of scientific rigour [113]. Ades (2006) suggested that additional uncertainty arising from indirect use of evidence should also be taken into account having CDS providing a "reasonable first approximation" [107].

- Features

CDS focus on synthesising the evidence, rather than on making quantitative comparison of the outcomes; therefore does not explicitly balance benefits and risks. Multiple criteria of benefits and risks can be modelled from the multiple diverse but complementary data – which is the purpose of CDS [113]. Sensitivity analysis is addressed in the second task of CDS through the estimation of bias.

- Visualisation

Visualising benefit-risk trade-offs from CDS analysis depends on the specific model chosen, therefore is not relevant to be appraised here.

- Assessability and accessibility

Based on adjustment for bias, the results of benefit-risk assessment from a cross-design synthesis analysis would be acceptable and attractive to many stakeholders. However, the actual interpretability and acceptability rely on the specific decision model and parameters use. The practicality of CDS when used in real-life decision-making can be hampered by the heavy resources required from having clinical trials and clinical database evidence. Although CDS primarily considers evidence from clinical trials and clinical databases, other sources of evidence such as case-control studies may also provide relevant information [45]. Stakeholders such as pharmaceutical companies and regulators, and those making decision in regulatory contexts would benefit from the application of CDS. Cross-design synthesis helps decision-makers to make better decisions by making the sources of and bias in the evidence transparent, alongside maximising strengths and minimising weaknesses of the evidence.

A.11 Utility survey techniques

Kimberley S. Hockley and Shahrul Mt-Isa

A.11.1 Stated Preference Method (SPM)

Description

Stated preference method (SPM) is an approach which explores how stakeholders respond when faced with hypothetical scenarios [114;115]. Stated preference approaches include contingent valuation and discrete choice methods. A parallel approach is the revealed preference method which explores the true stakeholder preferences and decision-making based on collected data. Whilst revealed preference may be favourable, there are four compelling reasons why stated preference should be considered: (1) it may not always be possible to infer stakeholders' preferences in drug benefit-risk assessments because many aspects of healthcare do not follow market goods behaviour where trade-offs are explicit; (2) asymmetric information problem may occur in revealed preference when the actual decision from a stakeholder (say a patient) is influenced by another stakeholder (say a physician) who is more informed about the trade-offs; (3) it is not possible to specify revealed preference in advance to guarantee that an appropriate benefit-risk model can be developed; and (4) stated preference allows large data to be collected at relatively moderate cost.

We suggest reading Ryan (2008) [114] for more details with worked example.

Evaluation

- Principle

The degree to which the method is logically sound can be debated. It may be argued that choice behaviour elicited under hypothetical circumstances may not truly reflect stakeholder behaviour in real life situations [114]. Therefore it can be questioned whether the results can be transferred and applied to real decision scenarios. It is important to note that SPM is an umbrella term for a variety of benefit-risk methodologies, and there is no specific ascribed set of methodological steps and mathematical techniques [115]. Despite this, it is possible for the principles of the methods to be easily understood by end users as it simply involves transferring the value judgements of stakeholders in hypothetical scenarios to real life situations.

- Features

The method assesses benefits and risks simultaneously via describing a hypothetical scenario. The description of the scenario is flexible and easily adaptable for purpose. It may include different numbers of benefits and risks which have the potential to vary over time, and can also include numerical and/or visual representations from multiple sources of evidence.

- Visualisation

There is no standard visual representation in the application of SPM.

- Assessability and accessibility

The acceptability and interpretation of the parameters and results depends on the specific type of stated preference methodology. SPM offers the advantage that it can collect large amounts of data at a moderate cost, and has the potential to examine proposed changes from a stakeholder perspective prior to implementation [114]. Revealed preference data does not present a complete and accurate representation of stakeholders' preferences because regulators, physicians and health professionals act as gatekeepers to treatments and medicines. Therefore SPM offers unique insight into stakeholders' values, preferences and decision-making.

A.11.2 Contingent valuation (CV)

Description

CV consist of conducting surveys which describe hypothetical scenarios, and then directly asking stakeholders their willingness to pay (WTP) if the scenario offers more benefit than the current situation, or willingness to accept compensation (WTAC) if the scenario is less advantageous than the current situation.

We suggest reading Smith (2003) [116] for introduction, Mitchell (2005) [117] as core reference, and Havet (2011) [118] for worked example.

Evaluation

- Principle

The valuation of an object as a whole is theoretically sound and derives from neoclassical welfare economics. However, this means that the results are limited and it is difficult to extend the results to other situations which may be similar, but vary specific attributes. As CV is a type of SPM it also shares the critique that choice behaviour elicited under hypothetical circumstances may not truly reflect stakeholder behaviour in real life situations. Additionally, the methodology has known biases. Consumer responses have been found to be either oversensitive if the hypothetical changes affect them, or under sensitive if they feel that the changes will not affect them. It is also very important to note that some stakeholders may morally object to assigning monetary values on treatments and medicines. Protest bidders have been known place a value of 0 if they ethically disagree with the principles of CV.

Contingent valuation explicitly incorporates stakeholder values via the level of WTP or WTAC they specify. If a stakeholder values a scenario high, then the corresponding WTP will also be high. Additionally, it is possible to create hypothetical descriptions of multiple scenarios and compare then compare WTP or WTAC.

- Features

The features are largely similar to SPM where description of the scenario is flexible and easily adaptable for purpose. It may include different numbers of benefits and risks which have the potential to vary over time, and can also include numerical and/or visual representations from multiple sources of evidence

- Visualisation

There is no standard method of visualisation.

- Assessability and accessibility

The results are easily interpretable. Some statistical knowledge is required to calculate and interpret mean and median values with standard deviations and confidence intervals. If demographic and SES details are to be investigated, then knowledge of regression models is also essential. However, CV may not be practical when questioning consumers during risk-benefit decision-making. This is because some medicines and/or treatments can be free and/or subsidised to the consumer, so selecting a WTP or WTAC may seem like an abstract and difficult concept.

A.11.3 Conjoint analysis (CA) and discrete choice experiment (DCE)

Description

Conjoint analysis, similarly to CVM, also consists of conducting surveys which describe hypothetical scenarios and elicit values. However, there is a fundamental difference between CVM and CA. CVM produces an overall expected utility for the hypothetical scenario based on the options and consequences, but CA breaks down hypothetical scenarios into sets of characteristics and attributes on which utilities are elicited individually and later combined. We acknowledge that there is an ongoing debate of the relationship between CA and another more structured approach of stated preference elicitation, the discrete choice experiment (DCE) [114;119;120]. We do not attempt to join this debate other than recognise that the two terms have been used interchangeably.

Discrete choice experiment is an extension of CA where the process of elicitation is done in a structured way. DCE can be regarded a framework for eliciting utilities from relevant stakeholders with roots in the random utility theory with a strong foundation in behavioural psychology. In DCE the most important characteristics of a situation are defined and labelled as attributes. Then, each attribute is assigned levels which can be cardinal, ordinal, or categorical. The attributes and levels are then systematically varied to explore all potential configurations of attributes. These are later reduced via fractional factorial designs, where the optimal design would be orthogonal. The resultant hypothetical situations are then compiled into choice sets which contain two or more hypothetical scenarios. Stakeholders will select the most attractive scenario from the choice set, and it is assumed their selection has the highest utility out of the options provided. From this, it is possible to analyse the value each attribute via logistic regression. The utility function is specified via: $\Delta U = \sum_{j=1}^{n} \beta_j X_j$, where ΔU is the chance in utility moving from treatment A to B, X_j $(j = 1, 2 \ldots, n)$ are the differences in attribute levels A and B, and β_j $(j = 1, 2, \ldots, n)$ are the coefficients of the model to be estimated [114;121].

We suggest reading Ryan (2008) [114] for introduction and as core reference, and Ryan (1997) [122] and Ryan (2001) [123] for worked examples.

Evaluation

- Fundamental principles

DCE is both a stated preference and conjoint analysis technique and has foundations in the random utility theory and statistical experimental design. As with all SP methods, it can be argued that choice behaviour elicited under hypothetical circumstances may not truly reflect stakeholder behaviour in real life situations. Also, within a DCE the description of a hypothetical scenario is reduced to a specific number of attributes which may not accurately represent real life situations and decision-making because these attributes may not be the only driving criteria for a particular problem. Another drawback is that there have been questions raised over its internal validity, consistency and test-retest reliability [121]. Despite this, there is a high degree of transparency and all the steps involved in the process are disclosed. Additionally, the approach incorporates stakeholders' value judgements on each attribute. The practicality of its application is however limited because respondents may experience fatigue, thus affecting to what extent the number of choice sets can be completed, and the number of scenarios presented within each choice set. A compromise between being overly comprehensive and being too specific is usually necessary.

- Features

The attributes represent benefits and risks and so DCEs are capable of handling multiple benefits and risks simultaneously. Additionally, the method can incorporate time dimension as one of the attributes. The coefficient for each attribute is tested for statistical significance and it is possible to calculate marginal rates of substitution between a selected attribute and other attributes.

- Visual representation

There is no commonly used visual representation of the method.

- Assessability and acceptability

The results are not easily interpretable from the perspective of a non-statistician or those unfamiliar with the concept of experimental designs. The method also has limitations – some real life situations may not be accurately reduced and represented by a set of attributes. However, DCE can be used to investigate how specific attributes may be viewed differently by regulators, physicians and patient populations.

A.12 Systematic review process

A.12.1 Detailed search strategy

1) Search on PubMed

   A search was performed on 10th September 2010 – time span: all years, document type: all. Search was made using: “benefit risk” OR “benefit and risk” OR “risk benefit” OR “risk and benefit” OR “benefit harm” OR “benefit and harm” OR “harm benefit” OR “harm and benefit” OR “net clinical benefit” OR “net benefit”, together (AND) with any of the words model* OR method* OR analys* OR assessment OR appraisal OR balance OR ratio. The search was performed to look for keywords within title/abstract. The search was refined by AND review (also in title/abstract) and limited to articles in English.

2) Search on Scopus

   A search was performed on 10th September 2010 – time span: all years, document type: all. Search was made using: (benefit W/2 risk) OR (benefit W/2 harm) OR “net clinical benefit” OR “net benefit” ”, together (AND) with any of the words model* OR method* OR analys* OR assessment OR appraisal OR balance OR ratio. The search was preformed within author keywords. (The operator W/2 ensures that the two words are within 2 words of each other). The search was refined by AND review (in title, abstract or keywords) and limited to articles in English.

3) Search on Web of Science

   A search was performed on 10th September 2010 – time span: all years, document type: all. Search was made using: “benefit risk” OR “benefit and risk” OR “risk benefit” OR “risk and benefit” OR “benefit harm” OR “benefit and harm” OR “harm benefit” OR “harm and benefit” OR “net clinical benefit” OR “net benefit”, together (AND) with any of the words model* OR method* OR analys* OR assessment OR appraisal OR balance OR ratio. The search was performed within topic which include title, abstract and keywords. The search was refined by AND review (also in topic) and limited to articles in English.

A.12.2 Systematic literature search results

The reviews and approaches found through systematic literature search and from additional materials are listed in Figure 40 below. The reviews are listed in chronological order from left to right to give an idea of their timeline and/or the popularity. Some of the reviews are excluded in whole, and some have been excluded in part because the approaches being reviewed were out of the scope of this report. Section A.13 lists the excluded approaches and briefly describes the reasons for having excluded them. An interesting HTA framework in prioritising health research is described in greater capacity but is not considered as a relevant approach for this report; therefore has not been formally classified in Section 3.

Figure 40 Literature search results

	Linnerooth, J. 1979	Carrin G. 1984	Fletcher, A. 1991	Sonnenberg, F.A. 1993	Tom, E. 1997	Ried, W 1998	McAlearney, A.S. 1999	Harris, R.P 2001	Briggs, A.H. 2002	Cates, C.J. 2002	Hahn, S. 2003	Holden 2003	Shih, Y.C.T. 2003	Wong, E.Y. 2003	Rich, K.M. 2005	Linkov, I. 2006	Sassi, F. 2006	Ades, A. 2006	Parexel Report 2007	Garrison, L.P. 2007	Mussen, F. 2007	Hoch, J.S. 2008	Chuang-Stein C. 2009	Mussen, F. 2009	Boada, J. 2009	Citrome, L. 2009	Khan, A.A. 2009	Guo, J.J. 2010	Ouellet, D. 2010	Pennington, C. 2010	Phillips, L. 2010	CIRS Workshop, 2012
Conley’s model	🟥																															
Consumer maximising models	🟥																															
Jones-Lee’s model	🟥																															
Linnerooth’s model	🟥																															
Usher’s model	🟥																															
Cost-benefit analysis (CBA)		🟥													🟥																	
Linear programming (LP)		🟥													🟥																	
Cost-effectiveness analysis (CEA)		🟥																														
Economitric modelling (EM)		🟥																														
Multi-attribute problem analysis (MPA)		🟥																														
Quality-adjusted life year (QALY)			🟥			🟥	🟥							🟥			🟥			🟥											🟥	
Markov Model				🟥	🟥							🟥							🟥												🟥	
Monte Carlo simulations (MCS)					🟥																							🟥				
Decision Trees					🟥																										🟥	
Survival Analysis / Hazard function / Kaplan-Meier					🟥																										🟥	
Fuzzy Logic					🟥																											
Health Years Equivalent (HYE)						🟥	🟥							🟥																		
Disability-adjusted life year (DALY)							🟥										🟥														🟥	
Quality adjusted time without symptoms and toxicity (Q-TWiST)							🟥												🟥				🟥					🟥			🟥	
Multi-attribute utility function							🟥																	🟥								
USPSTF method								🟥																								
Bayesian approach									🟥		🟥		🟥																		🟥	
Incremental cost effectiveness ratio (ICER)									🟥													🟥										
Cost effectiveness (CE) plane									🟥																							
Frequentist approaches											🟥																					
Relative-Value-Adjusted (RV) NNT and RV-NNH												🟥																🟥				
Relative-Value-Adjusted MCE												🟥																				
Computable general equilibrium (CGE)															🟥																	
Input-output models (I-O)															🟥																	
Multimarket models															🟥																	
Partial equilibrium analysis															🟥																	
Social accounting matrices (SAMs)															🟥																	
Multi Criteria Decision Analysis (MCDA)																🟥			🟥		🟥							🟥			🟥	
Confidence Profile Method (CPM)																		🟥														
Cross Design Synthesis (CDS)																		🟥														
Mixed and indirect treatment comparison																		🟥														
NNT/NNH																			🟥				🟥	🟥		🟥		🟥			🟥	
Benefit less risk analysis																			🟥				🟥	🟥				🟥			🟥	
Principle of Threes																			🟥					🟥							🟥	
Transparent Uniform Risk Benefit Overview (TURBO)																			🟥					🟥							🟥	
Incremental Net Health Benefit (INHB)																			🟥									🟥			🟥	
Maximum acceptable risk (MAR)																			🟥									🟥				
Minimum clinical efficacy (MCE)																			🟥									🟥				
Benefit risk model																			🟥													
MCDA combined with a decision conferencing process																			🟥													
Net Clinical Benefit (NCB)																			🟥													
Comparative risk assessment (CRA)																					🟥											
Average cost effectiveness ratio (ACER)																						🟥										
Benefit:Risk ratio																							🟥			🟥						
Global Benefit:Risk (GBR) Score																							🟥									
Benefit risk Contour																								🟥				🟥				
Beckmann Model (BM)																								🟥							🟥	
Principles of Threes modified to quantitative																								🟥								
Net efficacy adjusted for risk (NEAR)																									🟥							
Probabilistic simulation methods (PSM)																												🟥			🟥	
Stated Preference Method (SPM)																												🟥			🟥	
Risk-Benefit Acceptability Threshold (RBAT)																												🟥				
Risk-Benefit Plane (RBP)																												🟥				
Clinical Utility Index (CUI)																													🟥			
Clinical Quality Value Analysis (CQVA)																														🟥		
Bayesian belief networks																															🟥	
CMR-CASS																															🟥	
Conjoint Analysis																															🟥	
Discrete-event simulation																															🟥	
FDA BRF																															🟥	
PhRMA BRAT																															🟥	
System dynamics																															🟥	
Consortium on Benefit Risk Assessment (COBRA)																																🟥
Southeast Asia Benefit Risk Evaluation (SABRE)																																🟥
Unified Methodologies for Benefit Risk Assessment (UMBRA)																																🟥

A.13 Excluded approaches

A.13.1 Preliminary Assessment of Technology for Health Services (PATHS model)

Description

PATHS model is designed for organisations that fund health technology assessments (HTA) to help prioritise funding decisions. It assesses whether the cost in terms of likely health gain by adopting a technology is justified. The result of a health technology assessment could be 'favourable', 'unfavourable' and 'inconclusive' which are the three scenarios considered in PATHS model. The probabilities, potential costs, and potential gains in terms of health indices (QALY, for example) of the three scenarios are estimated and compared to the costs and health outcome without adopting this technology to obtain incremental costs and incremental effectiveness. The expected incremental cost-effectiveness ratio from the three scenarios is derived and used to prioritise HTAs.

We suggest reading Townsend (2003) [124] for more details on the HTA PATHS model.

Evaluation

- Principle

PATHS model is not a benefit-risk assessment, but a cost-effectiveness assessment. The effectiveness is indicated by health indices which already have the benefit and risk associated with the technology integrated. Since the decision is before the trial, the information used in PATHS is from previous available data on the technology and experts' opinion.

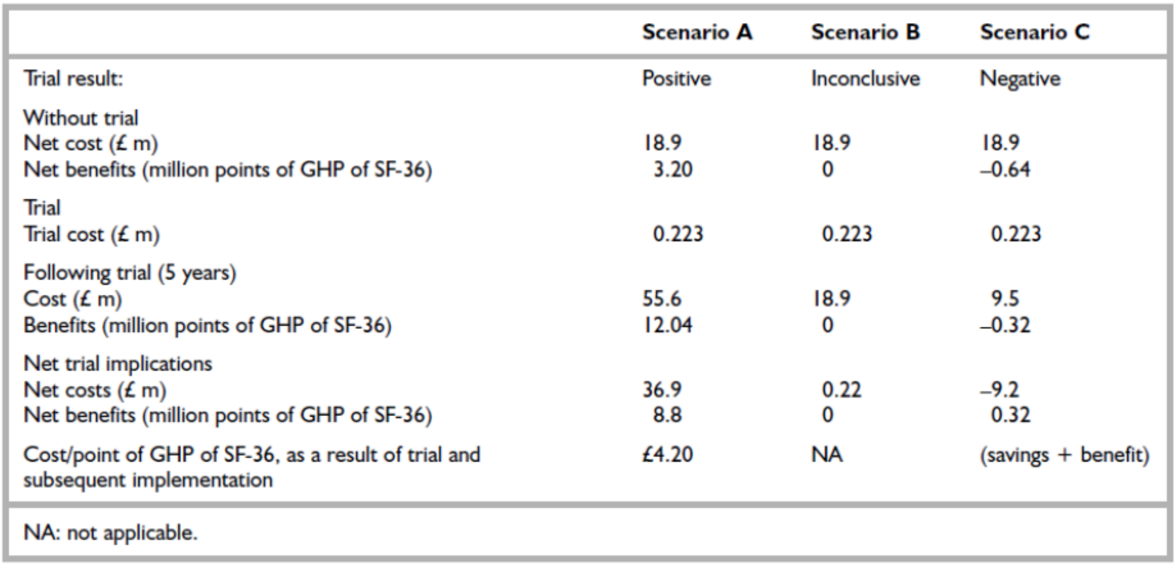
- Features

The effectiveness evaluation of PATHS model resembles a benefit-risk balance assessment, and this is done through health indices such as QALY. So disregarding costs, the PATHS model does not significantly add value to benefit-risk assessment in its own right.

- Visualisation

There is no graphical representation of PATHS model, but the results can be visualised in a table for the three scenarios as illustrated in Table 24.

Table 24 Evaluation of the postnatal midwifery support [124]



	Scenario A	Scenario B	Scenario C
Trial result:	Positive	Inconclusive	Negative
Without trial			
Net cost (£ m)	18.9	18.9	18.9
Net benefits (million points of GHP of SF-36)	3.20	0	–0.64
Trial			
Trial cost (£ m)	0.223	0.223	0.223
Following trial (5 years)			
Cost (£ m)	55.6	18.9	9.5
Benefits (million points of GHP of SF-36)	12.04	0	–0.32
Net trial implications			
Net costs (£ m)	36.9	0.22	–9.2
Net benefits (million points of GHP of SF-36)	8.8	0	0.32
Cost/point of GHP of SF-36, as a result of trial and subsequent implementation	£4.20	NA	(savings + benefit)

NA: not applicable.

- Assessability and accessibility

PATHS model is easy to perform and understand. The critical aspect of this approach is over-reliance on experts' opinions since there could be no data available (for example the probability of the three possible outcomes of the trial).

-- End of report --