This report is based on research conducted by the XXX Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. XXX). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

*Methods Research*

Number XXX

# Methods for Benefit and Harm Assessment in Systematic Reviews

**Prepared for:**
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
http://www.ahrq.gov

**Contract No.**

**Prepared by:**

*Investigators*

AHRQ Publication No. XX-XXXX
Date

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted for which further reproduction is prohibited without the specific permission of copyright holders.

**Suggested Citation:**
XXX. Methods for Benefit and Harm Assessment in Systematic Reviews. Reviews of Evidence. Evidence Report/Technology Assessment No. XXX. (Prepared by XXX Evidence-based Practice Center under contract No. XXX). AHRQ Publication No. XX-XXXX-X. Rockville, MD Agency for Healthcare Research and Quality. In press

**No investigators have any affiliations or financial involvement (e.g., employment, consultancies, honoraria, stock options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in this report.**

# Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

Comments are welcomed on this evidence report. They may be sent by mail to the Task Order Officer named below, at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D.
Director, EPC Program
Agency for Healthcare Research and Quality

Parivash Nourjah, Ph.D.
Task Order Officer
Agency for Healthcare Research and Quality

The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services of a particular drug, device, test, treatment, or other clinical service.

# Acknowledgments

# Abstract

**Introduction**. Systematic reviews are challenged by how to report and synthesize benefits and harms in order to inform the assessment of the balance of benefits and harms by decision makers with varying preferences. A number of quantitative approaches for benefit harm assessment exist but it is unclear whether they are applicable in the context of a systematic review.

**Objective**. To describe key challenges and methodologic characteristics of existing quantitative approaches for assessment of benefits and harms, to identify useful quantitative approaches and formulate principles for the assessment of benefits and harms in the context of a systematic review

**Process.** A multidisciplinary team reviewed the literature on quantitative approaches to assessment of benefits and harms to understand the key characteristics of these approaches and identify challenges. The team explored the role of values and preferences in systematic reviews, and formulated principles for reporting, analysis and presentation of benefits and harms in systematic reviews so that decision-makers are better able to weigh the benefits and harms (including adverse outcomes and burdens) for a population, as well as for subgroups for whom this balance may vary. This process was informed by input from a Technical Expert Panel.

**Results.** A number of challenges for the conduct (e.g. selection of outcomes, time horizons, and sources of evidence for benefits and harms), analysis (e.g. putting different outcomes on same scale), grading of the strength of evidence (e.g. asymmetry of quality for benefit and harm outcomes) and incorporation of preferences in a systematic review were identified. Among 15 quantitative approaches for assessment of benefits and harms, most could be used in a systematic review because they use aggregate data and do not require individual patient data. Simpler approaches such as the ratio of the number-needed-to-treat and number-needed-to-harm may be suitable for relatively simple decision making contexts where very few benefit and harm outcomes are relevant, which are similarly important and for which there is evidence of similar quality. More complex methods seem more adequate if multiple benefit and harm outcomes are important. The team formulated principles to address these issues in the context of a systematic review.

**Conclusion.** Careful attention to the common as well as unique characteristics of the outlined quantitative approaches for the assessment of the balance of benefit and harm and the principles may bridge the gap between evidence synthesis and its translation to evidence based public policy.

# Contents

Evidence Report ..... 1

Chapter 1. Introduction ..... 3

Chapter 2. Overall Process ..... 7
- Identification of Challenges ..... 7
- Identification of the Role of Values and Preferences ..... 7
- Identification of Key Characteristics and Desired Properties of Quantitative Approaches ..... 8
- Development of Principles ..... 8
- Role of the Technical Expert Panel ..... 9

Chapter 3. Identification of Challenges ..... 11
- Background ..... 11
- Process of Identification of Challenges ..... 11
- Results ..... 12
- Summary ..... 16

Chapter 4. Values and Preferences ..... 17
- Background ..... 17
- Process ..... 18
- Definitions ..... 18
- Patient Preferences ..... 19
- Patient Values ..... 20
- Decision-maker preferences ..... 20
- The role of preferences in benefit and harm assessment ..... 20
- Conclusion ..... 24

Chapter 5. Quantitative approaches for benefit and harm assessment in the context of systematic reviews: A framework for organizing and selecting quantitative approaches ..... 25
- Background ..... 25
- Processes ..... 26
  - Review of Literature ..... 26
  - Identification of key characteristics of existing quantitative approaches for benefit and harm assessment and development of unifying framework ..... 26
  - Defining desired properties of a benefit and harm assessment ..... 27
- Results ..... 27
  - Results Part I: Key characteristics of quantitative approaches for benefit and harm assessment ..... 27
  - Results Part II: Key characteristics of benefit and harm assessment ..... 29
  - Review of the literature: key characteristics of existing quantitative approaches for benefit and harm assessment ..... 30
  - Results Part III: Desired properties of quantitative benefit and harm assessment ..... 39
- Discussion ..... 41

Chapter 6. Development of Principles .....45
Background .....45
Process .....45
Principles for Benefit and Harm Assessment in Systematic Reviews .....46
Overarching Principles .....46
Systematic Review Protocol Development .....46
Conduct and Reporting of Systematic Reviews .....49
Limitations .....49

Chapter 7. Conclusion .....51

Reference List .....53

List of Acronyms and Abbreviations .....57

Figure 1. Stages of benefit and harm assessment .....5
Figure 2. Definitions: benefits and harms assessment .....6
Figure 3. Preferences .....19
Figure 4. Key characteristics of benefit and harm assessment .....29
Figure 5. Framework for organizing quantitative approaches .....30

Table 1. Listing of Challenges of Assessment of benefit and harm in systematic reviews .....15
Table 2. Organizing framework of existing quantitative approaches for benefit and harm assessment .....34

# EVIDENCE REPORT

# Chapter 1. Introduction

Systematic reviews conducted by Evidence-based Practice Centers (EPC) assess the effectiveness and safety of health care interventions and are useful for a variety of decision makers. A systematic review usually considers a range of outcomes which are relevant to patients and stakeholders. Systematic reviews sometimes consider a range of study designs including trials and observational studies that inform the Key Questions. Systematic reviews also attempt to identify subgroups (e.g. elderly and ethnic minorities) in whom the effectiveness and safety of interventions may vary. In certain comparative effectiveness reviews, the benefits and harms obtained from the trials and observational studies may be described separately, and may be addressed in separate Key Questions.[1 2]These results may be reported on separate tables and in separate sections of the report.[3-5] There is usually information asymmetry with more reliable and robust data on benefits as compared to harms in comparative effectiveness reviews. A recent review of 104 comparative effectiveness studies reported that only 19% of comparative effectiveness studies of medications focused on safety.[6] The metrics with which benefit outcomes are reported are usually mean differences between study groups, odds ratios or relative risks that are pooled across studies. Harm outcomes associated with treatments are commonly reported as dichotomous outcomes and summarized as pooled odds ratios (various forms) or relative risks and if they are rare maybe modeled on a relative scale. Systematic reviews may describe the strength (quality) of the evidence for each benefit and harm outcome separately. [1] The balancing of benefits and harms (i.e do benefits of an intervention outweigh harms?) is usually under the scope of decision makers, and is currently usually qualitative.

Several formal quantitative approaches for assessing benefits and harms of interventions have been proposed.[7] However, these approaches have primarily been used to evaluate benefits and harms within an individual study. Because these approaches focus on individual studies, some of these approaches have, to date, relied on information collected from individual study participants that is unavailable to most systematic reviewers. These approaches have not been rigorously applied to comparative effectiveness reviews which synthesize information across multiple studies. There are several methodological issues inherent in utilizing these quantitative approaches for assessment of benefits and harms in systematic reviews. The role of preferences for assessment of benefit and harm in systematic reviews has not been formally evaluated. There has not been an assessment of how these approaches could be utilized in the context of systematic reviews to quantitatively assess benefits and harms.

The specific objectives of the first part of this EPC methods project were to: **To describe key challenges and methodologic characteristics of existing quantitative approaches for assessment of benefits and harms, to identify useful quantitative approaches and formulate principles for the assessment of benefits and harms in the context of a systematic review** so that decision makers are better able to weigh the benefits and harms (including adverse outcomes and burdens) for a population, as well as for subgroups for whom this balance may vary. While fulfilling these objectives the team addresses the role of preferences and how quantitative approaches for an assessment of benefit and harm may be applicable to systematic reviews. In an upcoming report the team will illustrate the results of implementation of various quantitative approaches in the context of a recent systematic review.

Our first objective was to identify challenges relevant to the assessment of benefit and harm in systematic reviews (evidence synthesis). (**Figure 1)** The team organizes its findings around an organizing framework for approaches to risk benefit analysis, in 4 stages: Evidence generation

(Randomized Controlled Trial (RCTs) and observational studies), evidence synthesis (systematic reviews and meta-analyses), processes from evidence to generation of Evidence-based Medicine (EBM) tools (modeling or simulation considering decision-making context, and of EBM tools, (clinical practice guidelines, decision aids). Ultimately, these evidence-based medicine tools together with individual preferences of patients in clinical practice will influence decision making, but this is not covered in this report. However, several of the challenges of assessment of benefit and harm in evidence synthesis, such as the selection of relevant outcomes, are inextricably linked to challenges that occur in the primary trials and observational studies that inform a systematic review (stage of evidence generation). Other challenges, such as the need to incorporate preferences, are applicable in all stages including the final stages of generation of clinical practice guidelines in the appropriate decision making context (Figure1). This report deals largely with the role that quantitative assessments of benefits and harms may play in the balancing of benefits and harms.( **Figure 2 Benefit and Harm Assessment**).This process of benefit and harm assessment relies on several sequential steps **(Figure 2**) including the reporting of benefits and harms as absolute or relative measures of benefits and harms in systematic reviews or meta-analysis [Step I] and the use of various quantitative approaches that provide explicit quantitative information on the benefit and harm outcomes [Step II]. Policy makers make judgments at the population level after an assessment of the balance of benefit and harm as described in Figure 2. [Step III]. Providers and patients make shared decisions at the individual-level using information from decisions at the population-level after incorporating their values and preferences [Step IV].

**Figure 1. Stages of benefit and harm assessment**

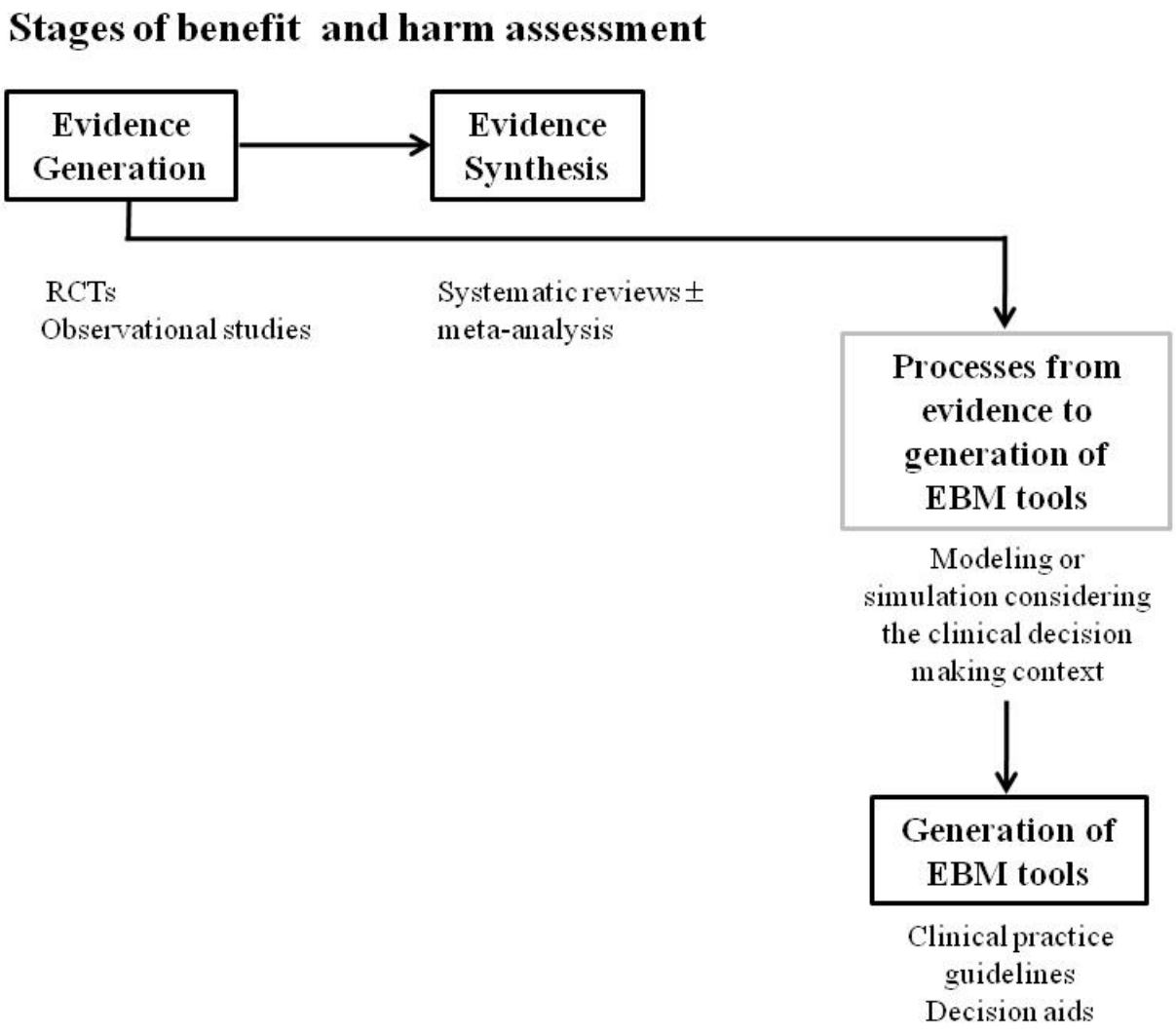
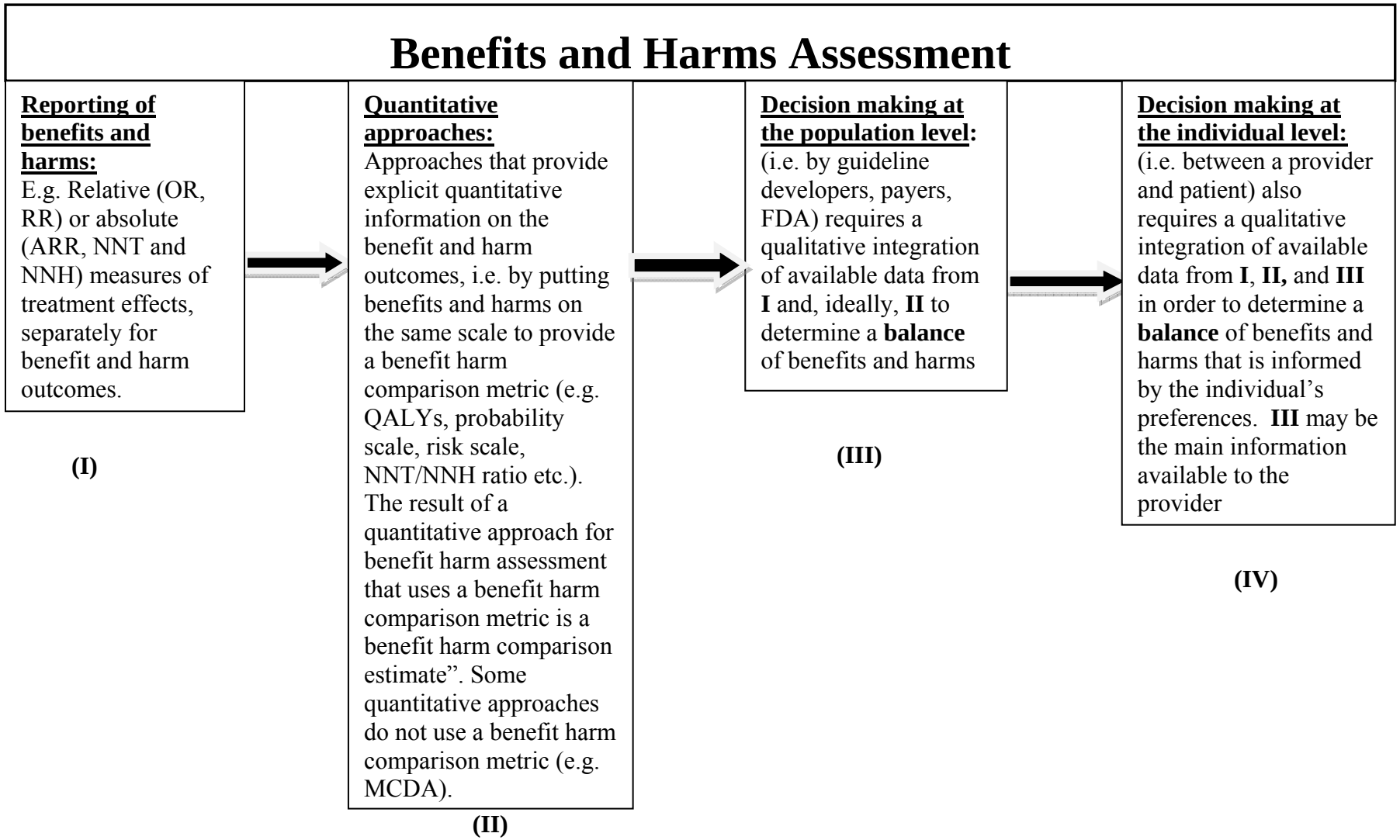


**Figure 2. Definitions: Benefits and Harms Assessment**




Abbreviations: AR= Absolute Risk; ARR= Absolute Risk Reduction; FDA= Food and Drug Administration; NNT=Number Need to Treat; NNH= Number Needed to Harm OR = Odds Ratio; QALY= Quality Adjusted Life Years; RR= Relative Risk; MCDA = Multicriteria decision analysis

# Chapter 2. Overall Process

Here the team provides a brief overview of the overall process to describe key challenges and methodologic characteristics of existing quantitative approaches for assessment of benefits and harms, to identify useful quantitative approaches and formulate principles for the assessment of benefits and harms in the context of a systematic review**.** The details of the process for each aspect can be found in subsequent sections of the report.

## Identification of Challenges

The team, consisting of clinicians, epidemiologists and statisticians identified the methodological challenges of assessing benefits and harms in systematic reviews by reviewing the literature on quantitative approaches and trying to apply them to the various steps of a systematic review. Although several reviews on the topic of benefit and harm assessment have recently been published, these reviews have been limited to applying the various quantitative approaches to data from a single study.[7 11 12] To maximize efficiency and avoid duplication, the team evaluated the quantitative approaches within these reviews, key articles culled from the investigators' reference libraries, and in citing literature and considered their application to benefit and harm assessment in systematic reviews. [7 11 12] The team used a sample of systematic reviews for guidance during the process. [3 4] The team selected a convenience sample of recent EPC reports of interventions from the effective healthcare website.[3 4 5 8]In order to improve the generalizability of our report to the wide community of systematic reviewers the team also reviewed a sample of other recent systematic reviews which reported on benefits or harms.[9 10] To inform the challenges the team evaluated quantitative approaches for assessment of benefits and harms with key articles culled from the investigators' reference libraries, including prior work on methods for describing benefits and harms. Each key article was reviewed as a group to inform the challenges of reporting, analysis and assessment of benefits and harms in systematic reviews until saturation of themes was achieved.

The team generated a preliminary list of challenges and then solicited input on the challenges from the TEP. The team listed the challenges according to the PICOTS framework, with slight modifications. Several of these challenges were applicable at various overlapping stages of the systematic review process, such as the analytic framework, inclusion and exclusion criteria for articles, quantitative data synthesis, narrative synthesis and grading the overall strength of the body of evidence, assessing applicability and presenting the results.

## Identification of the Role of Values and Preferences

The team identified the incorporation of preferences as a major challenge to the assessment of benefits and harms in systematic reviews. These could be either the unconscious influence of preferences of systematic reviewers on the conduct of the systematic review, or an explicit effort by the systematic review team to incorporate the preferences of various decision makers. The team devoted three weekly sessions to evaluate the implicit and explicit role of values and preferences during the steps of evidence generation, evidence synthesis and generation of EBM tools in the assessment of benefit and harms in systematic reviews. Apart from patient preferences, the team also evaluated how preferences of decision makers, including investigators,

systematic reviewers and policy makers, may impact the assessment of benefit and harms in systematic reviews.

## Identification of Key Characteristics and Desired Properties of Quantitative Approaches

Our team discussed each of the existing quantitative approaches, identified through the ways described above, for the assessment of benefit and harm in 12 weekly sessions. The discussion served to develop an understanding of each approach but also to define properties that characterize various quantitative approaches for the assessment of the balance of benefit and harm. The team did not exclude any quantitative approach because of the type of data used, type of statistical analysis or modeling, the way preferences of decision makers were considered or the way the analyses were presented. The team screened the reference lists of all included articles for additional relevant articles. Each article that appeared relevant was then discussed by the group as described below.

The team iteratively defined key characteristics with which each existing approach for the assessment of benefit and harm can be described. These characteristic(s) also allow comparisons across approaches. In referring to an assessment that combines benefit and harm in a single number, the team refers to a "benefit harm comparison metric". In contrast, an EPC Comparative Effectiveness Review (CER) may sometimes present a side-by-side assessment of benefits and harms. For example, the team identified the type of data needed for each approach, what assumptions underlie these approaches, what metrics were used to estimate the result of a benefit and harm comparison and how the results of the assessment were communicated to decision makers. When applicable, the team recorded limitations inherent to each of the quantitative approaches that may threaten their usefulness and their underlying implicit and explicit assumptions. All approaches were then described using the final list of key characteristics and unified for a comparison using a structured table.

The team also identified the ideal desired properties of a benefit-harm quantitative approach based on our review of the approaches and the challenges and key characteristics. Starting with the list of key characteristics, through group discussion, the team identified which properties a quantitative approach for benefit and harm assessment should possess in order to be valid and relevant to decision makers. Recognition of such desired properties helps to recognize advantages and disadvantages of existing quantitative approaches and to refine them or to develop new quantitative approaches that fulfill as many desired properties as possible.

In addition, the team defined desired properties of quantitative approaches for benefit and harm assessment that were beyond statistical considerations such as what populations would ideally be considered, the comprehensiveness and quality of data to be collected or what sources of evidence should be considered.

## Development of Principles

To develop principles for systematic reviews to consider and to guide our work, the team reviewed the above list of challenges, the role of values and preferences, the previously defined key characteristics and the desired properties of the various quantitative approaches. The team sought to develop principles that would facilitate a robust assessment of benefits and harms in

systematic reviews. These principles addressed various stages of the systematic review process including protocol development, conduct, reporting, analysis, presentation and the role of preferences in systematic reviews.

## Role of the Technical Expert Panel:

A five member diverse Technical Expert Panel provided initial input to the team on the conduct of this report via a telephonic conference at the outset of the study and helped us to identify initial challenges. Subsequently, electronic input from the Technical Expert panel was solicited on the preliminary list of principles and key characteristics. Suggestions from the Technical Expert Panel were also incorporated into the final list of principles for the assessment of benefit and harm.

# Chapter 3. Identification of Challenges

## Background

There are several challenges in conducting an assessment of benefit and harm in systematic reviews of interventions. These challenges occur at several steps of the systematic review process from protocol development to reporting and presentation of results. Systematic reviews usually consider a wide range of outcomes and study designs. There is substantial variability in the reporting of benefit and harm outcomes across studies. Information asymmetry with more reliable and robust data on efficacy and benefit as compared to harms of interventions in comparative effectiveness research may hinder an adequate assessment of the balance of benefits and harms of interventions, ultimately by decision makers. The team sought to examine those challenges which are barriers to an assessment of benefits and harms of interventions in systematic reviews that can optimally inform decision making at a population level regarding the balance of benefits and harms.

## Process of Identification of Challenges

To identify the methodological challenges of assessing the balance of benefits and harms in systematic reviews, the team selected a sample of recent EPC reports of interventions from the effective healthcare website.[3-5 8]In order to improve the generalizability of our report to the wide community of systematic reviewers the team also reviewed a sample of other recent systematic reviews which reported on benefit or harms.[9 10] To inform the challenges the team also wanted to identify quantitative approaches for the assessment of benefits and harms. We identified these quantitative approaches in several ways. We evaluated quantitative approaches for assessment of benefits and harms with key articles garnered from the investigators' reference libraries, including prior work on methods for describing benefits and harms. Several reviews on the topic of benefit and harm assessment have been recently published. [7 11-14]To maximize efficiency and avoid duplication the team evaluated these reviews including prior work on methods for describing benefits and harms. The team did not conduct a systematic review because reviews on the topic of benefit and harm assessment already existed,[7 11-14] and because an electronic search was unlikely to identify all relevant literature in the absence of a reliable standard for indexing relevant articles. The team evaluated articles that were written for quantitative harm-benefit assessment, which included consideration of at least one outcome for both benefit and harm of a medical or public health intervention. The team included approaches that analyzed benefit and harm outcomes entirely separately as well as approaches that provided a benefit-harm comparison metric for the balance of benefit and harm (e.g. ratio of number-needed-to treat [NNT] to number-needed-to harm [NNH]). Articles citing any of the above types of references were evaluated and reviewed to determine whether they described additional relevant concepts. The team screened the reference lists of all included articles for more relevant articles. Each key article was reviewed as a group to inform the challenges of reporting, analysis and assessment of the balance of benefits and harms in systematic reviews until saturation of themes was achieved..

## Results

The principle challenges in benefit harm assessment in systematic reviews are listed below and shown in **Table 1**. Several of these challenges span multiple stages of the systematic review process such as the analytic framework, the eligibility criteria for inclusion and exclusion criteria for articles, quantitative data synthesis, grading the overall strength of the body of evidence, and assessing applicability. The team organized its listing of challenges around the PICOTS framework with some minor adaptations to incorporate the challenges of preferences.

**Populations.** The assessment of benefit and harm in systematic reviews needs to consider whether the evidence on benefit and harm is applicable to the target population of interest. Additionally, the applicability of evidence to particular subgroups of interest (older or comorbid population) on benefit may be different from the applicability of evidence on harm. The ultimate balance of benefits and harms, as determined by decision-makers, and the uncertainty of benefits and harms and their balance may vary for subgroups defined by specific characteristics among the population (e.g. age, or presence of comorbidities) or by estimated risk for outcomes (e.g. <5, 5-10 or >10% risk for 10 year mortality) due to various reasons. Some of these reasons may include the altered pharmacodynamic or pharmacokinetic properties of drugs with age. For example, a drug that is cleared by the kidneys and toxic at higher doses may accumulate in older patients and have a different benefit-harm profile than among younger patients.

**Interventions.** The assessment of benefit and harm in systematic reviews needs to consider that the fidelity of the intervention to a prespecified protocol may vary between studies that report on benefit or harm. For example observational studies on harm may allow for more flexible criteria as compared to the rigid adherence to protocol required in most clinical trials.

**Comparators.** Efficacy of interventions such as drugs is usually established in placebo controlled trials required for approval,[8] while non-pharmacological interventions are commonly compared to usual care. There are two alternatives to address such lack of evidence from direct head-to-head comparisons in the context of a systematic review that may include a benefit harm assessment. First, network meta-analysis based on RCTs provides estimates on the relative effectiveness of health care interventions by considering both direct and indirect evidence. However, for such indirect analyses to be reliable, studies should be comparable in terms of quality, applicability and measurement of outcomes. Numerous network meta-analyses on the relative effects of competing interventions on benefit and harm outcomes have been published recently.[15] Second, observational studies may also evaluate users of the intervention against non-users or alternative users rather than placebo. [16]

**Outcomes.** Since there is no well accepted measure of benefit across therapeutic areas, systematic reviews report on a range of surrogate and patient important outcomes. Some systematic reviews may also report on *composite* outcomes. An assessment of the balance of benefits and harms using composite outcome is challenging especially if the individual endpoints of the composite occur at different frequencies, show different effect sizes, and are of unequal clinical importance.[17]

*Surrogate outcomes* on benefit also pose challenges for assessment of benefit and harm. A recent CER report found evidence of benefit of various oral hypoglycemic agents on the

surrogate outcome of glycated hemoglobin.[8] However there was no conclusive evidence of benefit on hard cardiovascular outcomes or mortality for various oral hypoglycemic agents. To conduct an assessment of benefit and harms, systematic reviewers are challenged to determine if the surrogate has been validated for each intervention for the health outcomes in the analytic framework.[18] Information may be unavailable to systematic reviewers on assessing the linkages between surrogate and health outcomes, and how these linkages may vary by subgroups such as the elderly.

In contrast to benefit outcomes which are prespecified and measured with greater reliability, the harm outcomes are usually unexpected, not always prespecified and have much more variability in definition across randomized controlled trials. Unless the harms of an intervention are known (such as bleeding with anticoagulants) harms are usually recorded as adverse events or serious adverse events across studies. Similarly, in a recent CER the outcome of congestive heart failure was defined variably across studies ranging from those heart failure events that are diagnosed clinically to congestive heart failure requiring hospitalization.[8] Systematic reviewers are challenged to balance the need for sensitivity which increases the power to determine rare harmful events vs specificity of outcome definitions which provide much more confidence in the strength of their association if such an association is present.

**Time.** The time horizon of studies on benefit may be different from those studies on harm. The balance of benefit and harm in the short term may be different from the balance of benefit and harm in the long term. Ultimately, as judged by decision makers, the balance of benefit and harm with interventions may change over time, such as the case with hormone replacement therapy. Systematic reviewers face a challenge on how to present information on benefits and harms in order to inform an assessment of whether the balance of benefits and harms changes over time.

**Study design.** RCTs provide the most valid estimates of treatment effects for both benefits and harms because of the control of confounding and selection bias. However, RCTs are often designed and powered to detect clinically relevant treatments effects for benefit outcomes while harm outcomes receive less attention in terms of the quality of data ascertainment and statistical power. Also, some RCTs exclude patients with certain characteristics (e.g. old age, comorbidities) from trials, who may be at greater risk of experiencing harm, or have duration of follow-up that is sufficient for the benefit but not for harm outcomes. The resulting asymmetry in the quality and quantity of evidence for benefit and harm outcomes from RCTs are the main reason why some systematic reviews may lower the threshold for harm and consider observational studies.[15] Such studies provide important information on harms associated with certain treatments but they are also susceptible to confounding, selection, information and reporting bias. Estimates of harm from randomized controlled trials and observational studies are different due to differences in study quality, applicability (populations, interventions or comparisons), measurement of outcomes, publication bias, outcome reporting or sources of funding.[15] As a result the estimates on harm have more uncertainty compared to estimates of benefit based on RCTs. In the context of a benefit harm assessment, a great challenge is to deal with such asymmetry in the quality and quantity of evidence for benefit and harm outcomes. Recent research suggests that estimates from harm derived from meta-analysis of RCTs are similar to those derived from meta-analysis of observational studies. Thus systematic reviews of harms may not need to be restricted to specific study types.[19]

**Strength of evidence.** The assessment of the balance of benefit and harm for various outcomes in systematic reviews is further challenged by the need to incorporate strength of evidence into the relative judgments of benefit and harm. Systematic reviews grade the strength of evidence separately for each outcome of each Key Question.[4 8 20] A recent comparative effectiveness review on angiotensin converting enzyme inhibitors vs angiotensin receptor blockers [4]presented evidence on benefit and harm with different evidence grades. No formal approaches provide guidance on how to grade the *balance* of benefit and harm of interventions making it difficult to incorporate the various grades of the strength of evidence on benefits and harms into a summary assessment on the balance of benefits and harms. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group does provide guidance on the importance of considering values and preferences when making a recommendation.[21] However, when methods are used to quantitatively assess benefits and harms, with or without the use of a benefit-harm comparison metric, there are not currently methods to grade such approaches. Apart from statistical uncertainty, the need to incorporate lower levels of evidence from heterogeneous data sources and study designs on harms which may not be equally reliable, valid or robust as the data sources on benefit creates an additional level of uncertainty around the assessment of the balance of benefit and harm in systematic reviews.

**Data.** The majority of comparative effectiveness reviews are conducted based on available summary (aggregate) data.[4 5] Thus, systematic reviews usually only have access to *marginal distributions* of benefits and harms (i.e. separate results for each benefit and harm outcome) without access to the information on *joint distribution* (i.e. describing the correlation between benefit and harm outcomes). In some rare cases individual patient data is available to the investigators.[3] This could be either as individual patient from a subset of all studies or sufficient data from all studies reported, or only marginal data made available, or any combination thereof across studies. For example a recent individual patient data systematic review of inhaled budesonide among participants with chronic obstructive pulmonary disease reported no statistically significant increased risk of pneumonia with inhaled budesonide but censored the analysis at one year of follow up despite having access to long term data.[22] In these cases it is unclear if evidence synthesis from the summary data of the entire body of evidence is much more reliable than individual patient data from a partial subset of the body of evidence. The analysis of benefits and harms in systematic reviews is challenged by the differences in scale used to report data on benefits and harms both in the included studies and the meta-analysis. In meta-analysis of harms which are usually relatively rare events, data are usually modeled on an odds or relative scale to leverage their advantageous statistical properties such as stability of estimates. On the other hand benefits could be on an absolute, continuous or relative scale. In contrast most quantitative approaches for the assessment of the balance of benefit and harm model data on an absolute scale.[7] Given the limitations of the data, systematic reviewers are challenged to report an estimate of uncertainty (e.g. 95% confidence interval) around the benefit harm comparison estimate if and when such a benefit harm comparison estimate is reported.

**Preferences.** Systematic reviews attempt to synthesize evidence on multiple outcomes on benefit and harm to serve the needs of a variety of decision makers such as guideline developers, payers, regulatory agencies, and ultimately patients and clinicians. These decision makers may have varying preferences for various outcomes and their values and preferences affect qualitative relative importance of outcomes and quantitative relative importance (or weighting) of outcomes

A systematic review of benefits and harms of strategies to prevent venous thromboembolism, considered prevention of pulmonary embolism as clinically significant outcome, but not prevention of deep vein thrombosis in contrast to previous systematic reviews which consider both as clinical outcomes. The resulting conclusions may reflect differences in preferences for outcomes.[23] The explicit incorporation of preference for various outcomes is challenging because a single comparative effectiveness review may attempt to provide information on benefit and harm for various decision-making contexts. Although these preferences may be implicitly elicited during various stages of the systematic review process by inclusion of a variety of decision makers as key informants or technical experts, systematic reviews are challenged by how to make underlying preferences about the relative importance of these outcomes explicit. Apart from outcome selection, preferences were felt to implicitly inform various stages of evidence synthesis and the assessment of benefits and harms. Thus the role of preferences, including in the choice of outcomes in systematic reviews, in the whole process of translation of evidence is discussed in Chapter 4 in more detail.

| **Table 1. Listing of Challenges of Assessment of Benefit and Harm in Systematic Reviews** | |
|---|---|
| Populations | • Consider how to determine if benefit and harms, or uncertainty about the balance, may vary for subgroups of the population |
| Interventions | • Consider how to determine if the dose of intervention/fidelity to intervention may vary between studies that report on benefit or harm |
| Comparisons | • Consider how to determine if the dose of intervention/fidelity to intervention may vary between studies that report on benefit or harm |
| Outcomes | • Consider how to assess the linkages between surrogates and health outcomes for specific interventions, and their variation for various subgroups.<br>• Consider how to evaluate the effect of interventions on composite outcomes, where individual endpoints have different effect sizes and different gradient of preferences |
| Time Horizon | • Consider how to determine if the time horizon in included studies in the assessment of benefit and harm reflect real life (e.g., early benefit late risks)<br>• Consider how to determine that benefit and harms which may change over time |
| Study Designs | • Consider how to address information asymmetry in systematic reviews with more robust data on benefit and less reliable data on harm<br>• Consider how to synthesize evidence on studies of harm from study designs other than RCTs such as observational studies, or uncontrolled studies such as case reports with evidence on benefit from RCTs |
| Strength of Evidence | • Consider how to convey the uncertainty around balance of benefit and harm outcomes with different strength of evidence grades. |
| Data | • Consider how to report benefits and harms on the relative or absolute scale<br>• Consider how to assess the "joint distribution" of benefit and harm under various scenarios: Individual patient data from all studies, sufficient |

| | |
|---|---|
| | individual patient data from some studies reported, marginal data available |
| Preferences | • Consider how to incorporate values and preferences which affect qualitative relative importance of outcomes and quantitative relative importance (or weighting) of outcomes. |

## Summary

The team identified several challenges that pose barriers to the assessment of benefit and harm in evidence synthesis. Although the team focused on challenges of evidence synthesis, the team also identified several challenges that overlap the stage of evidence generation, evidence synthesis and generation of EBM tools such as clinical practice guidelines (Figure 1). The selection of relevant outcomes for the assessment of benefit and harm may arise from the stage of evidence generation. Additional challenges, such as the need to incorporate preferences may be particularly applicable in the final stages of generation of EBM tools in the appropriate decision making context but are also relevant at various earlier stages, such as the selection of outcomes. These challenges will need to be addressed to facilitate a robust assessment of benefits and harms in systematic reviews. The subsequent sections expand on the role of preferences, and how these challenges need to be considered in various quantitative approaches for the assessment of benefit and harm. The team finally outlines principles that attempt to address these challenges.

# Chapter 4. Values and Preferences

## Background

Decisions, whether at the individual level or policy level, should incorporate best evidence, individual patient information (e.g. disease severity, life-expectancy, comorbidity), and patient preferences.[20 24] This statement, derived from the pioneering work of David Sackett and AHRQ's Effective Health Care Program's methods guide, is well-accepted and makes intuitive sense. However it is difficult to arrive at an assessment of benefit and harm based on "best evidence" without relying on preferences. The team posits that the process of finding and using "best evidence" is fundamentally informed by preferences; both of patients and others in the processes of generating evidence and translating it into practice. Preferences affect every step of the process of generation of evidence to the application of evidence in clinical practice. There are many emerging methods of quantitative assessments of benefits and harms that have been used for individual studies and may be relevant for evidence syntheses such as systematic reviews and modeling tools. These methods are fundamentally dependent on the incorporation of preferences. The team believes that we will arrive at better assessments of benefits and harms if were explicit and transparent about the ways that preferences affect our generation and syntheses of evidence.

For example, when investigators select outcomes they implicitly make a choice that reflects what they think is important for decision making or other purposes. The same applies to EPC systematic reviews even if systematic reviewers may not always be aware of the role of preferences. The key informants help determine important outcomes for key questions; but it is important to recognize that this process will be affected by the preferences of the key informants, and that all key informants may not have the same preferences, as noted in the venous thromboembolism ( deep vein thrombosis and pulmonary embolism ) example from orthopaedics.[23] In order to make end-users of comparative effectiveness reviews of interventions, including policy makers like CMS, guideline developers, aware of the role preferences have played during the generation of a systematic review, it seems preferable to be explicit about the effect preferences have on systematic reviews and explain the rationale for the chosen preferences. In addition, systematic reviewers should consider how end-users make use of systematic reviews, in order to optimally present data on benefits and harms and in order to conduct an assessment of benefits and harms.

In this section, our objective is to describe how preferences play a role in available approaches to assessment of benefits and harms of interventions. The team organizes its findings about the role of preferences around an organizing framework for approaches to benefit harm assessment, in 4 stages: Evidence generation (Randomized Controlled Trial (RCTs) and observational studies), evidence synthesis (systematic reviews and meta-analyses), processes from evidence to generation of Evidence-based Medicine (EBM) tools (modeling or simulation considering decision-making context, and of EBM tools, (clinical practice guidelines, decision aids). (See Figure 1) Ultimately, these evidence-based medicine tools together with individual preferences of patients in clinical practice will influence decision making, but this is not covered in this report.

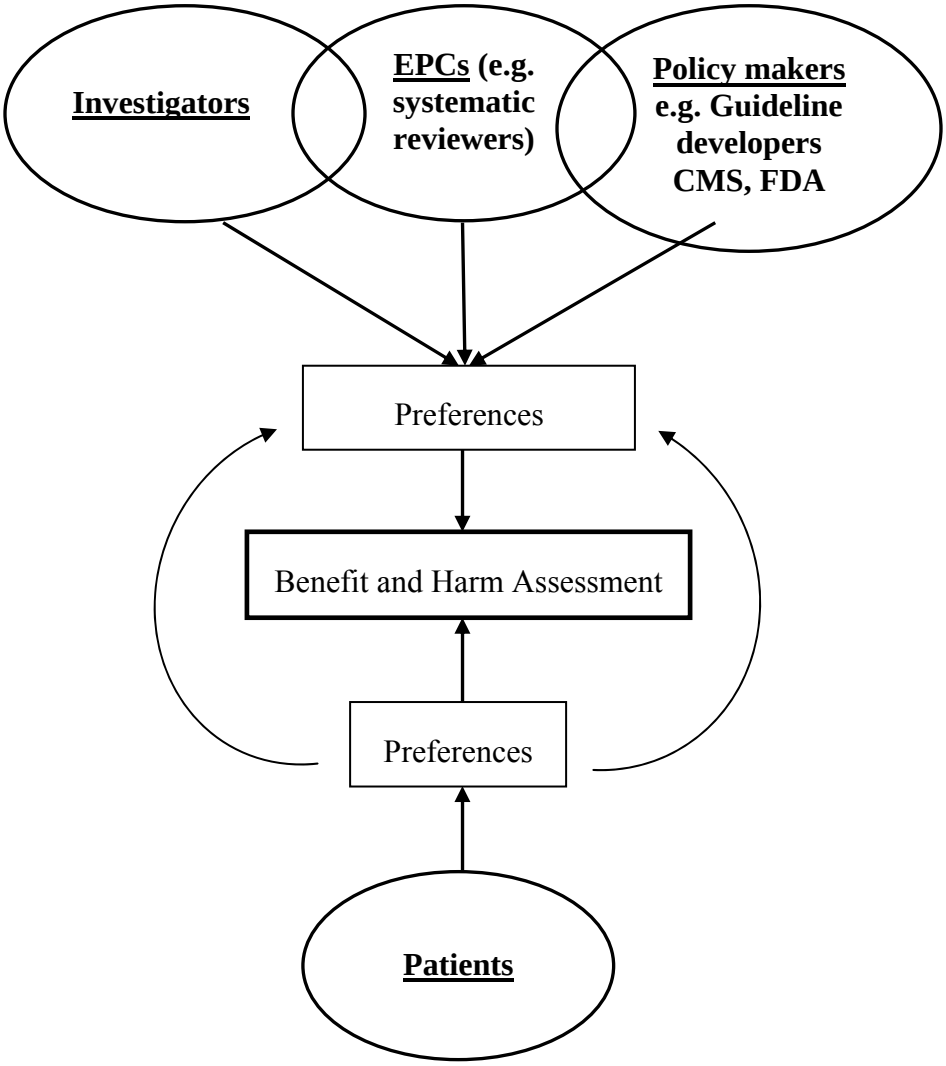
## Process

The team identified the incorporation of preferences as a major challenge to the assessment of benefits and harms in systematic reviews. The team devoted three weekly sessions to evaluate the implicit and explicit role of values and preferences during the steps of evidence generation, evidence synthesis and generation of EBM tools in the assessment of benefit and harms in systematic reviews. The team particularly considered the role of patient preferences for quantitative approaches to assessment of benefits and harms. Apart from patient preferences, the team also evaluated how preferences of investigators, systematic reviewers and policy makers may impact the assessment of benefit and harms in systematic reviews. The team developed operational definitions of terms based on the literature.

## Definitions

The terms "patient preferences" and "patient values" are often used together or interchangeably in the medical literature and the area of benefit harm assessment. Often, these terms are not clearly defined and it is not clear what the difference between values and preferences is.[25] This report refers predominantly to patient preferences and defines the working definitions for patient preferences and values as follows below, as well as defining preferences for decision makers.

**Figure 3: Preferences**




## Patient preferences

Patient preferences express the relative importance that patients or potential patients place on various health outcomes. They refer to the degrees of subjective satisfaction, distress, or desirability that patients or potential patients associate with a particular health outcomes. The various health states can be different severities of one condition (e.g. mild versus moderate versus severe dyspnea) or different conditions (e.g. hospital admission versus severe dyspnea). Preferences are a consequence of values and beliefs and the specific contexts in which patients face decision making. Preferences can be elicited by utility instruments and be expressed as utilities.

## Patient values

Values are a person's beliefs, desires, and expectations of what is "right" or "wrong." Values are based on predispositions, and on family and cultural context and experiences. Values are latent traits that are not directly observable (i.e. measurable) and can only be approximated by their expressions such as preferences. They are not specific to a certain context.

## Decision-maker preferences

The team believes there are multiple decision-makers along the path from I to IV as described in Figure 2, including investigators for clinical trials and observational studies (hereafter referred to as *investigators)*, EPC systematic reviewers (hereafter referred to as *systematic reviewers*), policy makers such as payers, CMS, and guideline developers (hereafter referred to as *policy makers*), and ultimately, in IV, both clinicians and patients.(Figure 3) Importantly, for stages I-III of benefit harm assessment, clinician preferences may also factor in through their roles on guideline panels and regulatory bodies. End-users of ystematic reviews are typically policy makers and other decision-makers, but may also include clinicians.

All of these groups of decision makers make decisions that impact the benefit harm assessment that systematic reviewers undertake themselves or ultimately inform. Ideally, all of these decision makers have patients as their primary focus and care about patient preferences, but may need to make decisions in the absence of perfect information.[26] This is one of the reasons why conflicts of interest are so important for benefit harm assessment. Since patient preferences are often not available, these decision makers must operationalize what is known (or assumed to be true when evidence about patient preferences and their variability is lacking) about patient preferences, and may incorporate other types of information into their preferences as well. Policy makers such as guideline developers are increasingly encouraged to be explicit about the role assumptions about preferences play in making decisions, such as the process of issuing recommendations in clinical practice guidelines. [21 27 28] For example, in addition to wanting to choose a patient-important outcome that may be informed by patient preference data, investigator preferences are also influenced by the financial constraints for a clinical trial which may force decisions about the length of the study, the outcomes followed, and exclusion and inclusion criteria of a trial. For clarity, the report refers to these preferences, informed by both pieces, as "investigator preferences." Similarly, the report refers to "systematic reviewer preferences" and "policy maker preferences" whenever possible to refer to the preferences arrived by consideration of both.

## The role of preferences in benefit and harm assessment

**The Role of Preferences for evidence generation.** Preferences are part of evidence generation in randomized controlled trials and observational studies in several ways, hereafter referred to as investigator preferences.

First, the team considers issues around study design along the Population Intervention Comparator(s) Outcome(s) (PICO) format. Understanding who the target population (P) is relevant for making decisions about study design. For example, the study design decisions about inclusion and exclusion criteria, and how aggressive the study is in recruiting in older, complex, or otherwise vulnerable populations, are affected informed by investigator preferences.

Investigator preferences affect the design and conduct of other aspects of study. For example, what is the intervention to be tested (I), and how long is the study, what are the comparator(s) (C). Next, investigator preferences affect which outcomes of a clinical trial or observational study are chosen (O). For example, what is the primary outcome, and what are secondary outcomes? From the patient's perspective, studies should choose an outcome important to them (mortality or function) as primary outcomes, but in many cases a surrogate outcome is chosen for reasons of feasibility (short-term trials, smaller sample sizes), which reflects the perspective of some investigators and funders. How closely linked surrogate outcomes are to patient important outcomes varies by condition as well as specific drug classes and there is often insufficient evidence to appraise this linkage.[18 26 29] For example, in the case of oral agents for type 2 diabetes, validation of a surrogate outcome such as glycated hemoglobin to cardiovascular outcomes would require that, systematic reviews should appraise the linkage between glycated hemoglobin and cardiovascular outcome separately for each drug class such as metformin, sulfonylureas and thiazolidinediones.[18] The evidence that metformin lowers glycated hemoglobin and potentially reduces cardiovascular risk does not immediately confer validity on the surrogate of glycated hemoglobin in the context of the thiazolidinediones. The thiazolidinediones also effectively lower glycated hemoglobin, similar to metformin, but may have at best neutral or at worse adverse effect on cardiovascular outcomes.[30] Inevitably the decision about outcomes is affected by preferences of patients, as well as the preferences of those conducting the study, its funders, and perceived end-users of the study. Investigator references also affect to what degree study procedures incorporate surveillance for outcomes that may be unintended consequences (such as harms).

The second step is the analysis of the study where, again, a number of decisions need to be made that are influenced by preferences of investigators but sometimes also by policy makers such as the FDA. For example, investigator preferences affect whether heterogeneity of treatment effects is assessed in order to determine whether the effects of an intervention vary across the study population or between subgroups of patients defined by some characteristics or risk profiles, respectively or, investigator preferences may determine whether a quantitative benefit harm assessment is conducted and which outcomes enter such an assessment. Investigator preferences also determine whether the analysis will be intention–to-treat or per-protocol, how rigorously data on harms will be ascertained and with what frequency, and whether data (on harm) will be collected on participants who withdraw from the study. Participants may withdraw because of harms and not counting these participants may result in under ascertainment of harms.[31] Thus investigator preferences in choice of analysis, the rigor and intensity with which harms will be monitored, and whether data will be collected on withdrawn participants impact the adequate ascertainment of harms and the subsequent assessment of benefit and harms.

**The Role of Preferences for evidence Synthesis.** Systematic reviewers may be unaware of all the end-users of their report and their preferences for benefits and harms. Systematic reviews attempt to incorporate different preferences through the presence of different experts on the technical expert panel or by the presence of a diverse team (e.g. diabetologists and internists and often a patient representative). These diverse inputs are implicitly assumed to represent different preferences, but this is often an unverifiable assumption.

"Patient preference" was introduced as a Medical Subject Headings (MeSH) term in PubMed only in the year 2010. Thus the underlying information base for patient preferences is

still evolving. Thus, few EPC reports and/or systematic reviews explicitly search for patient preferences. For example, in contrast to the large number of studies that have reported on the effectives of medications for type 2 diabetes, very few studies are available on the role of patient preferences for various outcomes in type 2 diabetes ( n=24) using simple search terms. This report refers to the preferences that systematic reviewers implement after information gathering through the technical expert panel and other means as systematic reviewer preferences (See Figure 3).

Thus, similarly to evidence generation, in evidence syntheses, preferences of the systematic reviewers and the key informants they engage are an essential component. First, similarly to evidence generation, preferences affect each component of the PICOTS format. Here, these are parallel decisions to those described above under "evidence generation," and these parallel decisions, informed by preferences, are made as systematic reviewers develop their key questions with the input of key informants. For example, preferences affect the choice of outcomes that are considered "important" in the systematic review, and the design of the key questions. Second, if there is any assessment of benefits and harms, there are also often assumptions about the relative weights of benefits and harms. For example, reporting a quantitative ratio of number needed to treat and a number needed to harm assumes that the benefit and harm are equally weighted in terms of the importance of the outcomes. This is an example of an implicit role of preferences. It is not always clear that systematic reviewers are aware that there is an implicit assumption that they are equally weighted. There are extensions of the NNT-NNH ratio approach that explicitly consider patient preferences or that can handle more than just one outcome for benefit and harm (see section on quantitative approaches). We would like to clarify a distinction between reporting NNT and NNH, and reporting a "NNT/NNH ratio". An NNT/NNH ratio, unless explicitly calculated using relative weights, does assume they are equally important. Reporting NNT and NNH separately (no ratio) does not make any assumptions about the relative importance of the outcomes, and that this issue is one of the reviewer putting the results in context and helping the reader interpret them. Quantitative approaches to benefit and harm assessment do need information about preferences. We believe there may be some value to EPCs or EPC users in understanding these methods and using these methods in some circumstances. Quantitative approaches may bring worthwhile information to the commonly "qualitative" process of balancing of benefits and harms by end users of comparative effectiveness reviews. These methods can be used to show the effect of varying preferences on the quantitative assessment of benefit harms, which ultimately could be used by decision-makers to show the effect of varying preferences on the balance of benefits and harms. (Figure 2) Thus, systematic reviews or their end-users would not be forced to decide the preferences and risk getting it wrong, but can describe the effect of preferences on the assessment of benefits and harms.

**The Role of Preferences for processes from evidence to generation of EBM tools: modeling or simulation considering decision-making context.** Modeling and simulation are important tools for the assessment of benefits and harms. Several quantitative approaches to benefit harm assessment include modeling or simulation, and these approaches all require an input about preferences, both in the choices of important outcomes, time horizons and patient profiles In addition, modeling or simulation for a specific decision making context also requires an understanding of the relative weight of benefits and harms, ideally based on patient

preferences, and assumptions about whether the distributions of benefits and harms are independent.

**The Role of Preferences for generation of EBM tools.** The role of preferences in the development of guidelines and decision aids is increasingly explicit.[26 32 33] It is now well recognized that the formulation of recommendations in guidelines depends on the ability of the guideline developers to make decisions about preferences on the behalf of patients. For example, if patient preferences are not thought to vary greatly, and thus the balance of benefits and harms is perceived to favor benefits, then GRADE is more likely to issue a strong recommendation if there is also high quality evidence.[21] Understanding when there is variability in patient preferences is thus important for determining the strength of the recommendation in GRADE. It is important to remember that they often have imperfect information about patient preferences, but nevertheless must arrive at preferences to inform their decision-making process. Thus, decision makers must play an essential role in drawing conclusions about preferences. This role contributes to an emerging consensus about the necessity to minimize conflicts of interest and ensure leadership of guideline panels includes those without conflicts of interest. Similar issues affect decision aids.[34] In order to create useful decision aids, the appropriate clinical situations where preferences affect decision-making must be identified early in the development process. For example, a decision aid needs to recognize that urinary incontinence is an important outcome to patients in order to incorporate preferences regarding this outcome into a decision aid.

**Use of EBM tools in clinical practice.** The preferences of the individual patient must be combined with guideline recommendations and decision aids in clinical decision-making. Population-level decisions such as those made to inform coverage, or guideline-recommendations, will not be identical to every individual's decision, even in settings with optimal shared decision-making using high quality, applicable evidence regarding the likelihood of benefits and harms.[35] This difference is fundamentally due to the variation in how patients would relatively weight the different possible outcomes, including both harms and benefits. Modern (e.g. online interactive) decision aids actually often include an instrument to elicit preferences and to make suggestions to patients on decisions that would reflect their preferences (for example, http://www.predictonline.com/copd/smbs/spm_copd/intro_ntroduction.htm, then see "your values").

Patient preferences may affect decision-making to varying degrees. Most health care decisions, in particular in the context of chronic diseases, are preference-sensitive.[36 37] Preference-sensitive decisions are taken in situations where different people or patients and their health care providers may come to different decisions based on their preferences and what is known about the benefits and harms. For example, Protheroe et al. have shown that patient preferences have a strong effect on whether you recommend guideline-concordant care or not for the prevention of stroke from atrial fibrillation with aspirin or warfarin.[35] Similarly, Sussman et al. have demonstrated that patient preferences have a large impact on whether or not aspirin would be recommended for the primary prevention of myocardial infarction or stroke.[38] On the other hand, there are situations where preferences do not play much of a role because there is unambiguous evidence or consensus that a certain health intervention is virtually necessary. Examples for necessary care are surgical repair of acute bleedings or fractures due to trauma or antibiotic treatment of a young patient with sepsis.

## Conclusion

Preferences affect an assessment of benefit and harms across the path of the translation of evidence. (See Figure 4) In current practice, these preferences are often implicit, and not necessarily recognized and transparently reported by those conducting an assessment of benefits and harms. This is true whether the assessment of benefits and harms is done through reporting of benefits and harms only or whether there are quantitative approaches taken. Preferences also affect decision making at the population-level undertaken by guideline developers and regulatory bodies (see Figure 2) and at the individual level.

# Chapter 5. Quantitative approaches for benefit and harm assessment in the context of systematic reviews: A framework for organizing and selecting quantitative approaches

## Background

There is an ongoing debate about the most appropriate quantitative approaches for benefit and harm assessment.[7 11] Existing quantitative approaches consider treatment benefits and harms in order to estimate the benefit harm balance. They aim at informing medical and public health decisions made by stakeholders that include patients or healthy subjects, health care and public health providers, payers, policy makers or regulatory agencies such as the Food & Drug Administration (FDA) or European Medicines Agency (EMA).

Quantitative benefit and harm approaches differ, as explained below, widely in terms of the data required, the type of benefit-harm comparison metric used that puts benefit and harm outcomes on the same scale (e.g. ratio of number-needed-to treat [NNT] over number-needed-to harm [NNH]) or the way preferences of decision makers are considered (e.g. preference-neutral versus explicit preference weights). Key decisions are whether and how to use a benefit-harm comparison metric and how to jointly consider the multiple treatment effects. The different perspectives of decision makers may explain to some extent why a single quantitative benefit and harm approach is unlikely to fit all the various needs and contexts for benefit and harm assessment. Other reasons for why many quantitative approaches for benefit and harm assessment exist may be that opinions differ on what type of evidence should enter a benefit and harm assessment (e.g. observational versus experimental data), how comprehensive that body of evidence should be (e.g. single trials versus meta-analysis), what assumptions are acceptable (e.g. when different outcomes should be put on one scale) or to what extent preferences of decision makers should be explicitly integrated. Some of these challenges are created by the asymmetry of attention given to or information available on benefit and harm outcomes. Clinical trials are commonly designed to provide high-quality evidence and sufficient power for benefit outcomes whereas harms often receive much less attention in terms of accurate and valid methods of measurement. As a consequence, observational studies become an important source for evidence on harms.[8]

**Past reviews of quantitative approaches for benefit and harm assessment have not organized the approaches according to important characteristics**.[7 11 12] Also none of the reviews have considered the applicability of these methods for systematic reviews and meta-analysis. A framework that recognizes the key characteristics of quantitative approaches and organizes them accordingly could be a step forward to understanding the common and different elements of existing approaches; to guide their further development; and to support investigators conducting a quantitative benefit and harm assessment, in their choice of approach. Such an organizing framework could be particularly attractive for organizations or investigators who conduct systematic reviews because the literature is mostly silent about the use of quantitative benefit and harm assessment in the context of a systematic review.

Most quantitative approaches were developed based on primary data sets where investigators had control over study design, outcome selection and individual patient data. Our aims are two-

fold: (1) to review available quantitative approaches to assess benefit and harm of medical and public health interventions, and (2) to develop a framework for organizing the quantitative approaches to benefit and harm assessment that would help systematic reviewers choose the appropriate method for their problem.

# Processes

## Review of the literature

The team began its search for quantitative benefit and harm approaches with key articles culled from the investigators' reference libraries, including prior work on approaches for assessing benefits and harms. [7 11 12] The team looked for articles that were written for quantitative benefit and harm assessment, which included consideration of at least one outcome for both benefit and harm of a medical or public intervention. The team included approaches that analyzed benefit and harm outcomes entirely separately as well as approaches that provided a benefit and harm comparison metric (e.g. ratio of NNT/NNH). The team screened the reference lists of all included articles for more relevant articles. Each article that seemed potentially relevant was then discussed by the group as described below.

The team did not perform a formal systematic review of the literature because a review on the topic of benefit and harm assessment already existed and because our focus was on organizing available approaches in general and in the context of systematic reviews in particular. [7 11-14]The team capitalized on the work done to create a list of relevant approaches, which allowed us to devote adequate resources to the main focus of organizing available methods for use in systematic reviews.[7 11-14]

## Identification of key characteristics of existing quantitative approaches for benefit and harm assessment and development of a unifying framework

Our team, consisting of clinicians, epidemiologists and statisticians, discussed each of the existing quantitative approaches for benefit and harm assessment in 12 one hour sessions. The discussion served us to develop an understanding of each quantitative approach but also to define properties that characterize quantitative approaches for benefit and harm assessment. To develop an organizing framework the team iteratively defined a list of key characteristics with which existing quantitative approaches for benefit and harm assessment can be described and that allow comparisons across quantitative approaches. These key characteristics were shared with the TEP. For example, the team identified the type of data needed for each quantitative approach, what assumptions underlie the quantitative approaches, what benefit and harm comparison metrics were used, and how the results of the benefit and harm assessment is communicated to decision makers. The team also recognized and recorded limitations inherent to each of the quantitative approaches that may threaten their usefulness. All quantitative approaches were then described using the final list of key characteristics and unified for a comparison using a table.

## Defining desired properties of a quantitative benefit and harm assessment

Starting with the list of key characteristics, through group discussion, the team identified which properties a quantitative approach for benefit and harm assessment should possess in order to be valid and relevant to decision makers. Recognition of such desired properties helps to recognize advantages and disadvantages of existing quantitative approaches and to refine them or to develop new quantitative approaches that fulfill as many desired properties as possible.
In addition, the team defined desired properties of quantitative approaches for benefit and harm assessment that were beyond statistical considerations such as what populations would ideally be considered, the comprehensiveness and quality of data to be collected or what sources of evidence should be considered.

# Results

## Results part I: Key characteristics of quantitative approaches for benefit and harm assessment

Figure 4 provides an overview of the stages of benefit and harm assessment and type of studies, data and analyses underlying such assessments. The team distinguishes between four stages: First, *Evidence generation* refers to primary studies (experimental or observational) where data on benefits and harms associated with medical or public health interventions are collected. Second, *evidence synthesis* refers to systematic reviews defined as "a scientific investigation that focuses on a specific question and uses explicit, pre-planned scientific methods to identify, select, appraise, and summarize similar but separate studies."[39] When appropriate, systematic reviews may include a meta-analysis. The third step (*processes from evidence to generation of EBM tools*), is often, but not always, done before the 4th step (*the generation of EBM tools.)* The development of EBM tools may include processes like simulation that estimate how the evidence applies to real world settings and may take an iterative approach to actually develop such EBM tools. The team defines EBM tools as aids that inform patients and health care provides about benefits and harms of medical interventions and thus support decision making in practice. The most prominent EBM tools are Clinical Practice Guidelines (CPGs) and patient decision aids. CPGs are defined by the IOM as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances".[39 40] Patient decision aids use leaflets, interactive media, or other tools to support patients in their decision making process.[41] The process that leads to the generation of EBM tools typically combines (global) evidence from systematic reviews and preferences of patients and health care providers (physicians, nurses or other health professionals) and resources that define the (local) decision making context.

Benefit and harm assessment can be performed at each stage from evidence generation to the generation and, ultimately, the use of EBM tools. Our report does not cover the use of such EBM tools with individual patients although it describes in the chapter on patient preferences how every step from evidence generation to clinical practice is affected by patient preferences. The purpose of the benefit and harm assessments and at what stage they are performed depends on the decision making context. Benefit and harm assessments can be done based on single trials

and observational studies. A typical example for this is the regulatory decision making context where phase II and III trials, sponsored by a single pharmaceutical company, serve as a basis for decisions on drug approval by the FDA or EMA. When benefit and harm assessment is done as part of systematic reviews or the generation of EBM tools a comprehensive body of evidence is generally considered. These benefit and harm assessments can inform policy decisions by health care agencies and payers, but also decisions at the individual patient level.

The key characteristics of a quantitative benefit harm assessment the team identified are:

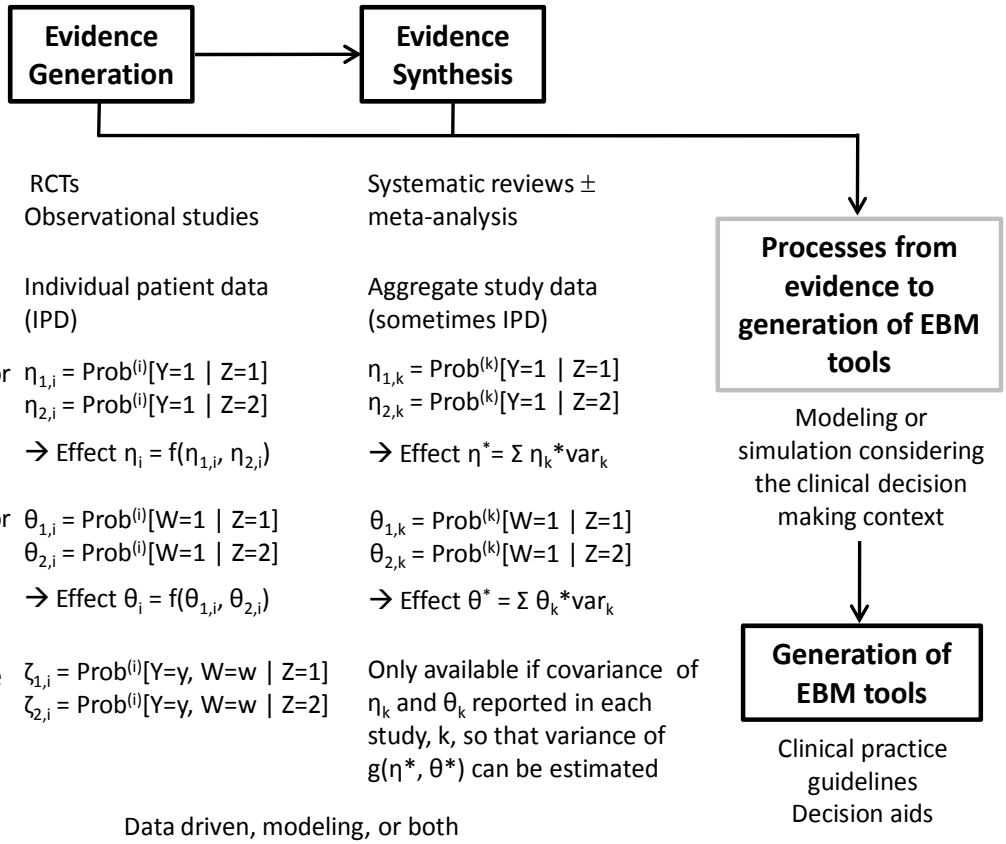
- The studies used for the benefit harm assessment (A in Figure 4)
- The type of data (B)
- The type of analyses (C)
- Assumptions when using a benefit harm comparison metric and assumptions about the joint occurrence of separate outcomes (D)
- The effect measures (E)
- The type of benefit harm comparison metric (F)
- Consideration and incorporation of patient preferences (G)

Figure 4 explains what type of data is typically available at the evidence generation and synthesis stage, which has implications for the type of quantitative benefit and harm assessment that can be done. Availability of individual patient data has some great advantages over aggregate data, which is what is typically available for evidence synthesis. In an individual study, there is, at least theoretically, information on the (co-) occurrence of benefit and harm outcomes for each patient as well as on their temporal sequence. Thus for each patient the cumulative occurrence of benefit and harm outcomes over time can be described as well as their correlation. For example, phosphodiesterase 4 inhibitors are new drugs for patients with COPD that aim at reducing flare ups (exacerbations) but also have some gastrointestinal toxicity. Patients who take these drugs regularly have a lower risk for exacerbations than those who do not take them but they are also at greater risk for nausea, diarrhea or abdominal pain.[42] A patient with greater susceptibility to the gastrointestinal toxicity it is probably more likely to have several and more frequent gastrointestinal symptoms over the course of treatment while a patient with less susceptibility or a lower dose experiences less adverse effects. An important consequence of the availability of such individual data is that the joint probability of benefit and harm outcomes can be considered.

Apart from the availability of individual patient data, the types of analyses, assumptions, effect measures, ways of presenting data and ways to incorporate patient preferences are similar for quantitative benefit and harm assessments that can be performed during evidence generation and synthesis. The following section on existing quantitative approaches to benefit and harm assessment will refer to the key characteristics outlined in Figure 4.

## Results part II: Figure 4

### Key characteristics of benefit and harm assessment



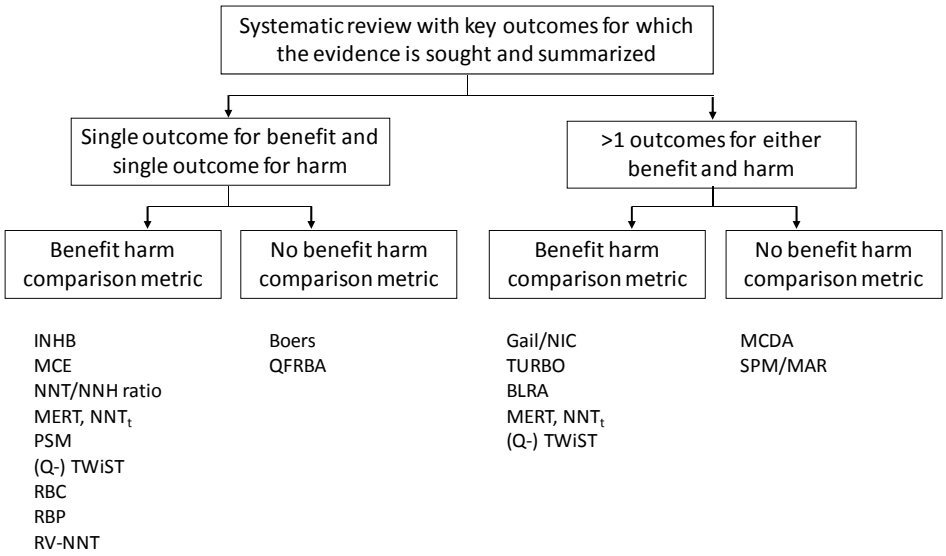

| | | |
|---|---|---|
| **A. Studies/ processes** | RCTs<br>Observational studies | Systematic reviews ± meta-analysis |
| **B. Types of data** | Individual patient data (IPD) | Aggregate study data (sometimes IPD) |
| Outcomes for benefit (B) | $\eta_{1,i}$ = Prob$^{(i)}$[Y=1 \| Z=1]<br>$\eta_{2,i}$ = Prob$^{(i)}$[Y=1 \| Z=2]<br>→ Effect $\eta_i = f(\eta_{1,i}, \eta_{2,i})$ | $\eta_{1,k}$ = Prob$^{(k)}$[Y=1 \| Z=1]<br>$\eta_{2,k}$ = Prob$^{(k)}$[Y=1 \| Z=2]<br>→ Effect $\eta^* = \Sigma\, \eta_k * var_k$ |
| Outcomes for harm (H) | $\theta_{1,i}$ = Prob$^{(i)}$[W=1 \| Z=1]<br>$\theta_{2,i}$ = Prob$^{(i)}$[W=1 \| Z=2]<br>→ Effect $\theta_i = f(\theta_{1,i}, \theta_{2,i})$ | $\theta_{1,k}$ = Prob$^{(k)}$[W=1 \| Z=1]<br>$\theta_{2,k}$ = Prob$^{(k)}$[W=1 \| Z=2]<br>→ Effect $\theta^* = \Sigma\, \theta_k * var_k$ |
| Dependence of B&H outcomes | $\zeta_{1,i}$ = Prob$^{(i)}$[Y=y, W=w \| Z=1]<br>$\zeta_{2,i}$ = Prob$^{(i)}$[Y=y, W=w \| Z=2] | Only available if covariance of $\eta_k$ and $\theta_k$ reported in each study, k, so that variance of $g(\eta^*, \theta^*)$ can be estimated |
| **C. Types of analysis** | Data driven, modeling, or both | |
| **D. Assumptions** | Multiple endpoints or composite outcomes for B&H<br>B&H on same scale (e.g. QALYs, time with symptoms)<br>Distributions of outcomes to estimate uncertainty | |
| **E. Effect measures for B & H outcomes** | Relative measures: RRR, OR, HR,<br>Absolute measures: ARR, NNT, NNH, net event rate<br>QALYs, time with symptoms or other | |
| **F. Metrics or displays to express B&H balance** | Difference ($g_i = \eta_i - \theta_i$) or ratio ($g_i = \eta_i/\theta_i$) of absolute measures (E)<br>Time gained -time lost through treatment<br>B&H graphical displays<br>Decision trees and others | |
| **G. Role of patient preferences** | Selection of population, comparisons and outcomes<br>Selection of time horizon<br>Use of weights for B&H | |

## Review of the literature: Key characteristics of existing quantitative approaches for benefit and harm assessment

The team identified 15 quantitative approaches for benefit and harm assessment. The approaches can be grouped into two broad categories: One category comprises simpler approaches that typically deal with just one outcome for benefit (e.g. prevention of stroke) and one outcome for harm (e.g. gastrointestinal bleeding). Importantly, these approaches can deal with composite outcomes that summarize the occurrence of at least one of two or more outcomes. Also, the approaches considering a single benefit and a single harm outcome can deal with several outcomes but only in separate analyses. The more complex approaches consider multiple benefit and harm outcomes in one analysis.

Figure 5 shows that the use of a benefit and harm comparison metric, that puts all outcomes on a common scale, further distinguishes among approaches. Figure 5 lists all approaches the team identified according to these categories and, thereby, provides an overview of existing approaches and also some preliminary guidance for the selection of approaches for a particular systematic review. It should be noted that Figure 5 provides a preliminary way of categorizing different approaches that the team will further refine in the second part of our project and that will ultimately provide guidance on how to select an approach for a specific systematic review. The team will discuss and test the selection of approaches in our next report but Figure 5 already indicates that for a particular systematic review not all 15 approaches are a sensible option. The number of outcomes, the need for a benefit and harm comparison metric (as by the preference of the decision maker) and the quality and quantitative of available data will probably drive most of the selection of approaches.

**Figure 5. Framework for organizing quantitative approaches**




**Abbreviations:** INHB =Incremental net health benefit; MCE = Minimum clinical efficacy; NNT =Number needed to treat; NNH = Number needed to treat for harm; Q-Twist = (Quality-adjusted) Time without Symptoms and Toxicity: RBC = Risk–benefit contour;;RV-NNT = Relative value adjusted number needed to treat; QFRBA = Quantitative Framework for Risk and Benefit Assessment; TURBO =Transparent Uniform Risk Benefit; BLRA= Benefit-less-risk analysis; PSM = Probabilistic simulation methods; MCDA = Multicriteria decision analysis: RBP = Risk–benefit plane; SPM = Stated preference method; MAR = Maximum acceptable risk;

To facilitate the understanding of and between the approaches, the team selected three approaches for a more detailed description as examples. A more formal evaluation of the approaches in the context of a SR will occur in our next report. The NNT to NNH ratio is an example of an approach where a single benefit and a single harm outcome is of interest. The Gail/National Cancer Institute (NCI) is an example where multiple separate outcomes are put on a benefit harm comparison metric and the multicriteria decision analysis (MCDA) provides an example for an approach where multiple outcomes are considered but where no benefit harm comparison metric is used.

**NNT and NNH.** The NNT and its harm outcome counterpart, the NNH, are perhaps the most widely used measures of risk and benefit when presented separately in systematic reviews and evidence-based medicine. Also, the NNT is the metric most commonly used in clinical practice guidelines when the benefit harm balance is discussed. The NNT or NNH are the number of individuals who need to be treated over a specified period of time for one person to benefit or be harmed, respectively, and will therefore vary as the specified treatment time varies. The NNT and NNH are almost always presented separately, i.e. not on a benefit harm comparison metric such as the **ratio** of NNT and NNH (hereafter referred to as NNT to NNH ratio). For example, the Clinical Practice Guidelines on Antithrombotic Therapy in Atrial Fibrillation of the American College of Chest Physicians present NNTs based on a systematic review of randomized trials of oral anticoagulant therapy vs no antithrombotic therapy: "The efficacy of warfarin was consistent across studies with an overall relative risk reduction of 68% (95% confidence interval [CI], 50 to 79%) analyzed by intention-to-treat.[43] The absolute risk reduction implies that 31 ischemic strokes will be prevented each year for every 1,000 patients treated (or 32 patients needed to treat for 1 year to prevent one stroke, NNT = 32)".

In contrast, the NNT to NNH ratio does not appear to be widely used. One reason for the rare application of this benefit harm comparison metric may be that investigators or guideline developers are reluctant to weigh benefit and harm outcomes equally on the same scale because of uncertainty about their relative clinical importance. To address this dilemma, Guyatt et al. proposed using relative value units to weight the NNT or NNH, which could also be applied to the NNT to NNH ratio.[44]

**MCDA.** The most commonly used multicriteria decision analysis approach is the Analytic Hierarchy Process. The team illustrates this using the comparative effectiveness review of oral hypoglycemic agents for type 2 diabetes. The first step in Analytic Hierarchy Process analysis consists of defining the goal of the decision, the alternatives being considered, and the criteria that determine how well the alternatives can be expected to meet the goal.[45 46] These are organized into a hierarchical decision model with the goal to determine the best treatment of type 2 diabetes. Operationally, we could define two criteria as being necessary for determining the best treatment; the treatment 1) maximizes benefits via glucose reduction, and 2) minimizes harms or medication-related adverse effects. The criteria on maximizing benefits could be divided into three sub-criteria on health-related quality of life, microvascular benefit (such as improvements in incidence of neuropathy, nephropathy and diabetic retinopathy) and potential macrovascular benefit. The criteria on minimizing medication-related adverse events could be sub-divided into six sub-criteria: minimizing the risks of congestive heart failure, fractures in women, macular edema, the risk of bladder cancer, myocardial infarction and hypoglycemia.

In the second step, information about how well the alternatives can be expected to fulfill the decision criteria is obtained from the comparative effectiveness review. The third step consists of two parts: a) comparing the ability of the alternative treatments to fulfill the prespecified criteria (maximizes benefit and minimizes harm) using standard Analytic Hierarchy Process pairwise comparisons and b) assessing the importance of these criteria to the decision goal. In the fourth step, the scales created in step 3 can be combined to create a summary score indicating how well the alternative treatments can be expected to meet the decision goal (Singh, Dolan & Centor, 2006). The fifth step consists of sensitivity analyses to explore the effects of changing the estimates or judgments used in the original analysis. The main advantages of Analytic Hierarchy Process are not use of the summary score *per se* but that it identifies the extent to which every criterion, judgment and weight contributes to that score and also incorporates uncertainty. These visual and graphical results allow one to gain understanding and articulate the divergence between relevant stakeholders (Dolan, 2010).

**Gail/NCI (National Cancer Institute).** Some decision making contexts are more complicated because there are many different treatment outcomes as well as sources of uncertainty. A well known example is the use of tamoxifen for the prevention of breast cancer. Tamoxifen reduces the risk for invasive and in situ breast cancer substantially and also prevents some bone fractures. On the other hand, it increases the risk for endometrial cancer, stroke and pulmonary embolism.

The National Cancer Institute under the leadership of Gail developed an approach to deal with multiple outcomes. Rather than simplifying the benefit harm assessment to single outcomes, as many investigators and guideline developers do, they estimated the probability of various outcomes for women with and without tamoxifen therapy over a period of five years. Based on observational studies, surveillance registries or placebo arms of randomized trials they first estimated the expected number of invasive breast cancers, in situ breast cancers, hip fractures, endometrial cancers, strokes, pulmonary embolisms, deep vein thromboses, colles' fractures, spine fractures and cataracts, each per 10,000 women, in the absence of tamoxifen treatment and over five years. They estimated these numbers overall and stratified for different age and race categories. They then estimated, for each outcome and based on the Breast Cancer Prevention Trial, the expected number of the same outcomes but with tamoxifen treatment, again per 10,000 women and over five years as well as overall and stratified for different age and race categories. In order to put all outcomes on the same scale but to also consider the relative clinical importance of these outcomes, they categorized the outcomes into life-threatening, severe and other outcomes and suggest weighting them with some factor (e.g. 1 for life-threatening, 0.5 for severe and 0.0 for other outcomes). These categories and weights can be modified according to patient or treatment provider preferences.

Ultimately, the results of the benefit harm assessment are presented as the net number of events prevented or in excess per 10,000 women treated with tamoxifen over a period of five years. For example, for a 45 year old woman with a 4% risk of invasive breast cancer over five years and with a uterus, the net number of events (weighted by their clinical importance) prevented is 196 per 10,000 women with this profile (the expected number of prevented invasive and in situ breast cancers was 299/10,000 woman but there were 59 women/10,000 woman with harm such as endometrial cancer, stroke, pulmonary embolism or deep vein thrombosis). The net benefit (benefit minus harm events) varied considerably and was positive for some profiles (as example above) but negative for others (e.g. black woman with age 50-59 years and a five year risk of invasive breast cancer of 4%).

**Description of key characteristics of 15 quantitative approaches for benefit harm assessment**

The 15 approaches identified have been described in detail elsewhere in papers focused on each individual approach. Some have been discussed in a review, although not in the context of systematic reviews.[7] Our focus here is to describe key characteristics that highlight their similarities and differences to enable systematic reviewers to understand the similarities and differences and choose the appropriate approach for their key question and context. The team refers again to the key characteristics as shown in figure 4:

- The studies used for the benefit harm assessment
- The type of data (B)
- The type of analyses (C)
- Assumptions when using a benefit harm comparison metric and assumptions about the joint occurrence of separate outcomes (D)
- The effect measures (E)
- The type of benefit harm comparison metric (F)
- Consideration and incorporation of patient preferences (G)

A summary is followed by a more detailed discussion of each approach.

Thirteen approaches can be pursued using individual or aggregate data while two approaches require individual patient data (B in Figure 4). Twelve of the approaches are data driven without any simulation and 2 approaches do or may also include simulation (C in Figure 4). Eleven of the 15 approaches put benefit and harm outcomes on the same scale to provide a benefit and harm comparison metric (F in Figure 4). Only two approaches provide measures of uncertainty around the benefit and harm comparison metric (C in Figure 4) although considering uncertainty is likely to be of key importance for decision makers and organizations making treatment recommendations. Four out of these 6 approaches could consider the joint distribution of benefit and harm outcomes for the estimation of uncertainty but many examples using these four approaches described in the literature do not consider the dependence between benefit and harm but only consider their marginal distributions (i.e. consider them to be independent, D in Figure 4). Five approaches use composites outcome for benefit and composite outcomes for harm while 11 approaches use multiple outcomes. Eight of the 15 approaches explicitly incorporate patient preferences (G in Figure 4).

Although most quantitative approaches were developed for use at the evidence generation stage, the team found that most of them can also be used at the synthesis stage since most approaches use aggregate data even if used in the context of single studies. In other words, most quantitative approaches do not take advantage of individual patient data but only use aggregate data. This is an important finding since it means that systematic reviewers have a wide range of quantitative approaches for benefit and harm assessment from which the most appropriate and feasible approach for a specific question can be chosen.

**Table 2. Organizing framework of existing quantitative approaches for benefits and harms assessment**

| **Approaches for B&H assessment**<br>**Key characteristics** | BLRA | Boers | Gail | INHB | MCDA | MCE | NNT& NNH | PSM | QFRBA | Q-TWiST | RBC | RBP | RV-NNT | SPM & MAR | TURBO | Number of approaches in each category |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Types of data** | | | | | | | | | | | | | | | | |
| Individual patient versus aggregate data | Indiv. | Either | Either | Either | Either | Either | Either | Either | Either | Either | Either | Either | Either | Indiv. | Either | Indiv: 2<br>Either: 13 |
| **Modeling type** | | | | | | | | | | | | | | | | |
| Data driven (DD) versus simulation (S) | DD | NA | DD/S | DD | DD | DD | DD | DD/S | DD | DD | DD/S | DD | DD | DD | DD | DD: 14<br>S: 3 |
| **Types of B&H metrics** | | | | | | | | | | | | | | | | |
| Absolute versus relative measures versus QALY versus other (O) | Other | A | A | QALY | A / relative | A | A | A | A / relative | A /QALY | A | A | A | A | A / relative | Absolute: 12<br>Relative: 3<br>QALY: 2 |
| **Assumptions** | | | | | | | | | | | | | | | | |
| Put B&H outcomes on same scale | yes | no | yes | yes | no | yes | yes | yes | no | yes | yes | yes | yes | no | yes | Yes: 11<br>No:4 |
| Uncertainty estimates for B&H assessment | no | no | no | no | yes | no | no* | yes | NA | no | no | no | no | NA | no | Yes: 2<br>No: 11 |
| Joint distribution of B&H outcomes considered for uncertainty estimates | no | NA. | P | NA | no | NA | no | P | NA | NA | P | P | no | NA | NA | Possible: 4<br>No: 3<br>n.a.: 8 |
| Multiple endpoints versus composite outcomes for B&H | M | Comp | M | M | M | M | M/ Comp | M | M | M | M | Comp | M | Multiple | Comp | Multiple: 11<br>Comp: 5 |
| **Consideration of preferences** | | | | | | | | | | | | | | | | |
| Explicitly considers preferences for B&H assessment: | yes | no | yes | no | yes | both | no | both | no | yes | no | no | yes | yes | no | Yes: 8<br>No:9 |
| **Types of presenting benefit risk balance** | | | | | | | | | | | | | | | | |
| B&H difference versus B&H ratio versus Time gained/lost versus B&H graphic versus other | D | Graphic | D | D | D, ratio, other | D | Ratio | D | D, ratio, other | Time, D | Graphic | Graphic | Ratio | D | Graphic | Difference: 9<br>Ratio: 4<br>Time: 1<br>Graphic: 3 |
| **Stages of evidence generation and application** | | | | | | | | | | | | | | | | |
| Evidence generation or evidence synthesis or EBM tool generation | Gene-ration, EBM | All | All | All | All | All | All | All | All | All | All | All | All | All | All | All: 14<br>Generation: 1<br>EBM: 1 |

* for some variants of the NNT approach such as NNTt and the minimum target event risk for treatment (MERT) uncertainty estimates exist[47]

NA = Not applicable; BLRA= Benefit-less-risk analysis; B&H = Benefit and harm assessment; INHB =Incremental net health benefit; MCDA = Multicriteria decision analysis: MCE = Minimum clinical efficacy; NNT =Number needed to treat; NNH = Number needed to treat for harm; PSM = Probabilistic simulation methods; QFRBA = Quantitative Framework for Risk and Benefit Assessment; Q-Twist = (Quality-adjusted) Time without Symptoms and Toxicity; RBC = Risk–benefit contour; RBP = Risk–benefit plane; RV-NNT = Relative value adjusted number needed to treat; SPM = Stated preference method; MAR = Maximum acceptable risk; TURBO =Transparent Uniform Risk Benefit Overview; P = Possible; D = Difference; A = Absolute; M = Multiple

**Benefit-less-risk analysis (BLRA).** BLRA combines benefit and harm into a single metric and was primarily developed for clinical trials.[48] For each patient (who is under some treatment or not) of a trial the benefit is recorded (yes=1 or no=0) and the harm is expressed as a value between 0 and 1. The relationship between benefit and risk is presented as risk subtracted from benefit (e.g. 1 (for benefit) – 0.2 (for harm). BLRA thus allows for statistical testing of comparisons between treatment groups. Patient preferences expressing the relative importance of benefit and harm outcomes can be considered. BLRA is the only approach the team identified that takes advantage of individual patient data. Thus if this method was to be applied in a systematic review, individual patient data would need to be gathered from the primary studies.

**Boers' 3x3 table.** This quantitative approach does not require any statistical models but suggests a way of organizing outcome data on the same scale.[49] The outcomes are split into three categories and the number of patients with a certain benefit-harm profile (e.g. major benefit and minimal harm) is displayed in a 3x3 table. Treatment effects are not considered since separate 3x3 tables are constructed for each treatment group. As a consequence, no measures of uncertainty are available. Patient preferences are not considered but the clinicians' view or agreement of what constitutes minimal, moderate or major benefit or harm, respectively. The method is both feasible for single trials or systematic reviews. A disadvantage is that, although each table is simple and easy to read, it requires readers to somehow estimate treatment effects across tables or to provide a benefit and harm comparison metric and thus challenges rather than facilitates conclusions concerning benefits and harms.

**Gail/National Cancer Institute**. This approach is probably the most comprehensive approach for benefit and harm assessment and considers various data sources to balance the benefits and harms of a treatment. As described above, the benefit and harm comparison metric can be calculated as the sum of benefit and harm outcome rates per patient profile. Patient preferences can be incorporated by looking only at one severity grade or by putting weights on outcome rates that reflect how very severe, severe or moderately severe events are perceived. Estimates of uncertainty arising from sampling variation or from combining different data sources of different methodological quality are not provided by this approach, although it should be possible to estimate uncertainty due to sampling variation. However, by looking at benefit and harm comparison estimates across patient profiles one gets an impression of how the net benefit changes, even qualitatively, as the baseline risk changes. This approach is resource intensive because it considers multiple data sources and multiple outcomes. However, an EPC CER report could provide an ideal basis for this approach if additional data, such risk for outcome estimates from observational studies, are also collected. A similar, though simplified approach was used by the United States Preventive Services Taskforce to make recommendations on the use of aspirin for the prevention of myocardial infarction. Similar to the tamoxifen example, the number of benefit (myocardial infarction) and harm (bleedings) events per 1000 men or women was estimated based on observational data and the evidence on treatment benefits and harms was then combined with these outcome estimates. The benefit harm comparison metric provided the number of net events (benefit minus harm) prevented or in excess when aspirin is used.[50]

**Incremental net health benefit (INHB).** INHB provides a benefit and harm comparison metric, using QALYs, to place one or more benefits and harms on the same scale and calculates the difference between benefits and harm between treatments (thus a result >0 is favorable).[51 52]

A key requirement for this approach is the valid measurement of utilities or the sometimes inaccurate transformation of quality of life scores into utilities. Also, it is often difficult to distinguish between the effects of benefits and harms when using utilities.

**Multicriteria decision analysis (MCDA).** A multi-criteria decision analysis allows for a systematic decision making in complex situations involving tradeoffs by considering various harms and benefits associated with treatments.[53 54] A decision tree model is developed to incorporate benefits from clinical trials and harms such as adverse effects. It allows for input from various stakeholders who may assign different preference weights to the risks and benefits. MCDA represents an approach to reduce the multidimensionality of benefit harm assessment in a systematic way and makes judgments explicit and transparent.[55] It allows for decision making in the presence of uncertainty and can incorporate data from multiple sources including systematic reviews.[56] The challenges of its application to systematic reviews include getting reliable information on various preferences, agreement on all relevant important benefits and harms and the relative importance and weighting of these outcomes, and the need to specify a decision context since systematic reviews are usually conducted to meet the needs of multiple decision makers. The flexibility of MCDA also poses challenges for benefit harm assessment as systematic reviews are unable to inform on all inputs, especially less tangible inputs (e.g. societal values, opportunity costs) that may alter harm and benefit balance in a particular decision context. The team provided an example for MCDA further above.

**Minimum clinical efficacy (MCE).** MCE incorporates harm and benefit into a benefit and harm comparison metric. The benefit is the difference in efficacy and harm both of which are expressed on a probability scale by applying relative risks reductions (treatment benefits) and increases (harms) to absolute probabilities as observed in untreated groups.[55 57] The intervention is said to have minimal clinically efficacy if the difference between benefit and harm is positive or above a minimally acceptably threshold. MCE can consider relative utilities. A limitation includes the inability to provide uncertainty estimates for the benefit and harm comparison metric.

**Number needed to treat (NNT) and number needed to treat for harm (NNH)**. The number needed to treat for benefit or harm is the number of individuals who needed to be treated over a specified period of time with the intervention for one person to benefit and experience the harm.[58 59] NNT and NNH depend on baseline risk (and are thus sensitive to different patient profiles) and the degree of relative risk reduction provided by the intervention, which is often assumed to be constant across the disease spectrum but may actually vary. NNT and NNH cannot be calculated for continuous outcomes unless such outcomes are dichotomized. NNT and NNH is perhaps the most widely used measure of risk and benefit reported in systematic reviews and evidence-based medicine. NNT and NNH can be calculated for single outcomes (e.g. NNT for exacerbations vs. NNH for fractures) or for composite outcomes for both benefit and harm. But since the concept of NNT is one of frequency and not of importance, NNT to NNH ratios or NNT to NNH differences should only be calculated for outcomes of similar importance unless they are weighed.[60] When a NNT to NNH ratio or difference between NNT and NNH is calculated as a benefit and harm comparison metric, their independence is assumed and they may need to be extrapolated so that they refer to the same time period. Extensions of the NNT to NNH ratio approach are the threshold NNTt and the minimum target event risk for treatment

(MERT).[47] The NNTt reflects the point at which the risks and costs of a clinical intervention balance the benefit and the MERT defines the minimum target event risk at which the intervention is justified. Methods for providing uncertainty for these benefit and harm comparison metrics are available.[47] The NNT to NNH ratio, NNTt and MERT all seem feasible within a systematic review context.

**Probabilistic simulation methods (PSM).** The PSM uses probabilistic simulations for benefit and harm comparison estimates using Monte Carlo methods. PSM estimates the incremental benefit versus the incremental harm for only one benefit and one harm outcome in a single model. Multiple outcomes need to be dealth with different models (similar to the NNT to NNH ratio). PSM can incorporate parameters from multiple data sources (systematic reviews of RCTs and observational studies), patient preferences (e.g. from conjoint analysis) and different patient profiles.[61-63] This method estimates uncertainty around the benefit and harm comparison estimate with or without consideration of the joint distribution of benefits and harms (depending on the availability of individual-level data or reporting of covariances). PSM thus may provide a comprehensive approaches to benefit and harm assessment but it requires methodologists with special expertise in PSM

**Quantitative Framework for Risk and Benefit Assessment (QFRBA).** QFRBA reports on benefit and harm separately. It does not provide a benefit and harm comparison metric and uncertainty estimates are only available for the separate treatment effects for benefit or harm outcomes.[7] The major advantage of this method is that keeps benefit and harm separate, leaves room for incorporation of preferences by decision makers and consideration of multiple outcomes. Also, QFRBA is probably the way most systematic reviews that contain a meta-analysis currently report or discuss the benefit and harm assessment.

**(Quality-adjusted) Time without Symptoms and Toxicity (TWiST or Q-TWiST).** TWiST compares treatments in terms of the time without symptoms gained versus the time lost due to the experience of adverse effects.[64 65] It thus puts the benefit and harm on the same scale (time). Q-TWiST is a further development where time is converted into QALYs.[66] Here the benefit and harm comparison metric is the difference between the drug associated gain in QALYs and the loss in QALYs associated with the treatment due to adverse effects. Q-TWiST has been widely used in oncology. The major advantage of this method is the ability to incorporate patient preferences which may change over time. The method depends heavily on the availability of measurements that allow estimating the length of time periods without symptoms and of time periods where adverse effects are experienced. Also, measurement instruments need to distinguish benefit and harm. For example, quality of life and some preference-based instruments often provide a composite score that already synthesizes the overall experience of a patient. QALYs value health states rather than changes in health states, and lack of a measure of uncertainty around these measurements may limit the usefulness of Q-TWiST. In a systematic review this method may be difficult to apply since QALYs associated with benefit and harm are unlikely to be reported in reports of primary studies.

**Risk–benefit contour (RBC).** The RBC plot portrays the probability of harm is higher for a new treatment, compared to another course of treatment, on the y-axis, and the probability that benefit is higher for a treatment, compared to another treatment, on the x-axis.[67] Contour lines

portray the shape of this relationship for a number of different probabilities and confidence levels. The dependence of the probability of benefit and harm are not considered at the individual level, rather the plot conveys study-level relationships. Though the method does not incorporate weights for each type of outcome it could easily be adapted to do so. RBC should probably be seen as a way to present data whereas the underlying analyses that yield the probability estimates can be based on different statistical approaches such as various forms of PSM.

**Risk–benefit plane (RBP) and Risk–benefit acceptability threshold (RBAT).** The RBP and RBAT display both separate estimates of benefit and harm and a benefit and harm comparison metric in a simple figure.[62 63] The individual-level dependence between benefit and harm is not considered. Using an absolute scale, the probability of benefit (from a comparison between two treatments) is plotted against the probability of harm. The slope created by a line between the origin and the two-dimensional result is called the risk-benefit acceptability threshold. This method does not consider outcome weights that would reflect patient preferences.

**Relative value adjusted number needed to treat (RV-NNT).** The major advantage of RV –NNT over NNT and NNH is that it allows for incorporation of preferences into the assessment of benefit and harm.[55 57] Otherwise it offers the same advantages as the NNT to NNH ratio approach and suffers from some the same limitations. Systematic reviews would need information on preferences to incorporate this method.

**Stated preference method (SPM) or Maximum acceptable risk (MAR).** SPM and MAR are approaches to elicit patient preferences for various tradeoffs and their acceptability for treatment. As such, these approaches do not attempt to balance benefits and risks. SPM and MAR are used to survey patients on how much burden from adverse effects or serious adverse events they are willing to accept to experience the benefit from treatment.[68-72] The typical method to elicit preferences is discrete choice or conjoint analysis where respondents have to pick their preferred treatment from two treatment scenarios that characterize the benefit and harm of these treatments. These approaches assume that the attractiveness of a particular treatment is a function of the benefit and harm attributes, which are combined in various ways in the different vignettes of the survey.[73]

The team is not aware of any systematic reviews where the stated preference or maximum acceptable risk approaches were applied as part of a systematic review because it requires a new primary study. However, it is imaginable that a systematic review searches for studies reporting on discrete choice experiments in addition to the evidence on treatment benefits and harms. Or, systematic reviews could provide the basis for designing discrete choice experiments so that vignettes of a survey provide realistic scenarios in terms of the benefits and harms of the treatments of interest.

The key characteristics described in Figure 4 do not directly apply to the stated preference or maximum acceptable risk approach since they do not provide a quantitative benefit harm assessment. But the preferences elicited by these methods can be incorporated into quantitative approaches that consider patient preferences and that are described in this document.

**Transparent Uniform Risk Benefit Overview (TURBO).** The TURBO diagram displays the factors "R" and "B". "R" is the sum of the most serious adverse effect (scored from 1-5) and the second most serious adverse effect (scored from 1-2).[14] The scores are based on the frequency and severity of the harm outcome. Similarly, factor "B" is the sum of the primary benefit (1-5), and the ancillary benefit (1-2) and the scores are based on the probability and extent of the benefit outcome. The "T" score represents the benefit and harm comparison metric and ranges from 1 (high "R" and low "B" score) to 7 (high "B" and low "R" score).
The TURBO diagram is typically used in a regulatory context (e.g. EMA) and therefore, based on single trials, but can easily be extended to systematic reviews. The factors "R" and "B" can be based on absolute or relative measures of treatment effects for which uncertainty estimates are available. But there is no uncertainty estimates for the benefit and harm comparison metric, i.e. the "T" score. Unlike other approaches, the TURBO diagram explicitly considers not only one but two outcomes for both benefit and harm that are weighted differently (up to 2 or 5 points). Challenges to applying the TURBO method include arbitrary selection of the two benefit and harm outcomes from a comprehensive list of outcomes and they way scores (combining frequency and importance of outcomes) are assigned.

## Results part III: Desired properties of quantitative benefit and harm assessment

**Study population.** The study population should cover the range of subjects for which a quantitative benefit and harm assessment is relevant. Consideration needs to be given to the health care setting in which particular decisions are taken, to indicators of disease severity, socio-demographics and comorbidities in complex patients (≥2 chronic conditions). Thus the study population does not need to be particularly broad but the key is that the study population reflects that target population.

**Selection of outcomes.** Ideally, evidence is available on all benefit and harm outcomes that should influence decision making. The selection of these outcomes depends on the decision makers, which could be patients, health care providers, policy makers or payers. For many treatment decisions, multiple benefit and harm outcomes will be necessary.

**Quality of outcome measurement.** Clinical trials are commonly designed to provide high-quality evidence and sufficient power for benefit outcomes. Harms often receive much less attention in terms of accurate and valid methods of measurement. Such asymmetry in the quality of outcome ascertainment affects the validity of a quantitative benefit and harm assessment and, as a result, the quality of evidence and strength of recommendation for or against a treatment. Therefore, a desired property of a quantitative benefit and harm assessment is the availability of high quality data from trials and observational studies for both benefits and harm outcomes.

**Outcome assessment across studies (does not apply to evidence generation).** The ascertainment of benefit and harm outcomes should be accurate and identical or at least similar across studies (e.g. how a cardiovascular event, an exacerbation or pain is defined). Whether this is the case depends much on the disease area. While in some areas both benefit and harm outcome measurement is harmonized (e.g. Outcome Measures in Rheumatology [OMERACT[74]), the selection of outcomes and their measurement varies widely in other disease areas.

**Study duration.** Patient outcomes should be collected for the entire study period (including treatment exposure and follow-up) if possible. This means that patients with an outcome (benefit or harm) should be followed until death, if possible. This is especially a commonly occurring problem where the follow-up is stopped after a benefit-related outcome happens, since this could result in an underestimation of harm-related effects of the intervention.

**Selection of evidence (does not apply to evidence generation).** The quantitative benefit and harm assessment should be based on comprehensive and the best available evidence. A systematic review should underlie quantitative benefit and harm assessments at the evidence synthesis and EBM tool development stage. Thereby, all studies that potentially contribute to the quantitative benefit and harm assessment are identified. The quality of the evidence that is finally considered depends on the underlying study designs but also on the desired properties for the quantitative benefit and harm assessment. As a consequence, a tradeoff between experimental and observational data is often necessary. While trials provide a higher quality of treatment effects (for both benefit and harm outcomes) observational studies are often the only source for harm outcomes in relevant populations and over relevant time horizons.[19]

**Treatment information.** Ideally, time-dependent information on treatment exposures is available because patients may start and stop treatment during a study. Realistically, such information is only available at the evidence generation stage and it is unlikely that dependent information on treatment exposures can be considered in the context of a systematic review.

**Provision of a benefit and harm comparison metric that considers multiple relevant benefit and harm outcomes.** Ideally, a quantitative benefit and harm assessment considers all relevant outcomes for benefit and harm and use a benefit and harm comparison metric such as QALYs, NNT to NNH ratio or an event rate (e.g. rate of events prevented or caused by some treatment).[13]

**Consideration of time of occurrence of benefit and harm outcomes and variability over time.** A quantitative benefit and harm assessment should consider that time of occurrence of benefit and harm outcomes may be different. Sometimes, e.g. for preventive treatments, harm outcomes are likely to precede benefit outcomes whereas for other treatments, benefit might be experience earlier than harm. In addition, a quantitative benefit and harm assessment should ideally be able to incorporate variability of benefit and harm in relation to time.

**Handling of different types of data.** Quantitative benefit and harm assessment should be able to handle different types of data (binary, recurrent, continuous, time to event). Many quantitative approaches identified here focus on binary outcomes with or without consideration of time to event. While this is appropriate for event data, patient-important outcomes such as quality of life or symptoms cannot be expressed appropriately as binary outcomes without substantial loss of information.

**Uncertainty estimates for the benefit and harm comparison metric.** Uncertainty for the benefit and harm comparison metric is likely to be of key importance for decision makers and organizations making treatment recommendations. Typically, uncertainty is represented by 95% confidence interval but there are also graphical displays to visualize the extent of uncertainty

such as RBC. As explained above, benefit and harm outcomes are likely to be correlated and such dependence of benefit and harm outcomes should ideally be taken into consideration. It is important to note that uncertainty may also arise from evidence of low quality. While it is probably not sensible to express such uncertainty statistically (e.g. 95% CI) there are other ways to express the quality of evidence (e.g. evidence grading according to the United States Preventive Services Task Force or the GRADE working group).

**Incorporation of preferences.** A quantitative benefit and harm assessment should explicitly state if and how preferences are incorporated into the estimate. Most of the quantitative approaches identified here do not incorporate preferences, at least, not explicitly. In the context of a systematic review explicit data on patient preferences can be gathered through identification of studies that used conjoint analysis or other methods to elicit patient preferences.

**Consideration of different patient profiles.** A quantitative benefit and harm assessment should allow taking different patient profiles into consideration.

**Communication of benefit and harm assessment to decision makers.** A quantitative benefit and harm assessment needs to be communicated effectively to decision makers so that they can take informed decisions that are line with their preferences. It is currently unknown which of the presentations formats are most effective. Important aspects that need to be considered are (1) parsimony of information as a result of a transparent and reproducible reduction of multidimensionality, (2) an explicit statement if preferences are already incorporated into the estimate or how individual preferences can be incorporated, (3) graphical displays to accurately convey quantitative information including uncertainty.[75] A recent paper highlights that in contrast to most systematic reviews that cite benefit and harm in different locations a conceptual simple visualization of benefit and harm in a single visual image could enhance communication of the benefits and harms to decision makers. [75]

# Discussion

The main findings of our review of the available quantitative approaches for benefit and harm assessment using a unifying framework are: (1) existing quantitative approaches can be broadly categorized into approaches that consider single or multiple benefit and harm outcomes and, in addition, into approaches that use a benefit harm comparison metric or not; (2) although none of the approaches seemed to be developed specifically for systematic reviews, all but two quantitative approaches (BLRA and SPM&MAR, which require individual patient data), can be used in the context of a systematic review where aggregate data are typically available; (3) most of the advantages and limitations of available quantitative approaches apply irrespective of whether they are used at the evidence generation or synthesis stage or for processes that lead to EBM tools; (4) none of the quantitative approaches for benefit and harm assessment explicitly consider the asymmetry in the quality and quantity of evidence, which is generally higher and larger for benefit as compared to harm outcomes. Such asymmetry is particularly important in the context of a systematic review because it has implications for the search strategy, selection of the evidence and the overall work load of reviewers; (5) there is no evidence that a single quantitative approach that can be favored clearly over the others in the context of every systematic review. The selection of a quantitative approach depends on the number of key

outcomes, the need or desire for a benefit harm comparison metric, the balance between offering as many of the outlined desired properties as possible and feasibility in terms of ascertaining a comprehensive evidence base, and the availability of epidemiological and statistical expertise.

The team identified a number of assumptions that are (implicitly) made when applying almost any quantitative approach for benefit and harm assessment: First, for some approaches it is assumed that one or more benefit and harm outcomes can be put on the same scale to calculate a benefit and harm comparison metric. Challenges for putting different outcomes on the same scale include their relative importance to decision makers (which may require different weighting), simplification of the outcomes (e.g. dichotomizing continuous outcomes, which may lead to substantial loss of information) or different methods and timing in the ascertainment of different outcomes. However, the advantages of a benefit harm comparison metric may be substantial, in particular in the context of situations where multiple outcomes are important and where patient, provider and policy maker preferences vary. A judgment on the balance between benefit and harm is made by decision makers irrespective of whether a benefit harm comparison metric is used or not. Decision makers need to come to a decision. The major advantage of using a benefit harm comparison metric (over using an approach without such a common metric) is that the assumptions about the relative importance of outcomes is made explicit and that sensitivity analyses provide evidence as to how the benefit harm balance changes if different assumptions are made. Also, a single number provide some advantages for the communication of net treatment benefits to patients and avoids overwhelming the patients with data on multiple different outcomes.

Second, an assumption for some approaches is that composite outcomes can be built that provide useful information for decisions makers. This may be seen as a similar process to putting different outcomes on the same scale, and it is associated with similar challenges. In addition, composite outcomes are seen increasingly critically because of some examples where composite outcomes led to highly misleading results.[17 76] In the context of systematic reviews the use of quantitative approaches for benefit and harm assessment that consider multiple endpoints seem more feasible than the use of composite outcomes although the amount and quality of evidence for different outcomes could differ quite a bit. For example, a different number of trials might contribute to pooled estimates for different benefit and harm outcomes because some outcomes were not measured or reported consistently across trials. However, composite outcomes, as an alternative, may often not be a good option in systematic reviews because composite outcomes for benefit or harm outcomes are often defined differently across trials and because consistent composite outcomes could only be constructed if individual patient data were available.

Third, the team was surprised to see that the joint distribution of benefit and harm outcomes was almost never considered or even discussed in any quantitative approach, even when individual patient data were available. Uncertainty estimates for the benefit and harm comparison metric may be different if joint distributions were considered but the team is not aware of methodological studies that quantified such a difference between uncertainty estimates with and without consideration of the joint distribution. Nonetheless in context of a systematic review it is unrealistic that information on the joint distribution of benefit and harm outcomes would ever be reported in any of the primary studies, let alone to find consistent reporting of the joint distribution of benefit and harm outcomes across studies. But systematic reviewers should keep in mind the limitation of not considering the joint distribution when interpreting results from a quantitative benefit and harm assessment.

If systematic reviewers decide to conduct quantitative benefit and harm assessment they have the choice among the approaches presented in this report and it is likely that additional quantitative approaches are currently being developed or that existing approaches are being modified. Feasibility is obviously one of the factors that will drive the decision on which approach to choose. Quantitative approaches such as NNT to NNH ratio and its modifications, MCE, Boers und TURBO are relatively simple approaches that might be feasible in many systematic reviews. Other quantitative approaches such as INHB, PSM, MCDA or Gail/NCI require information that is not collected automatically in a systematic reviews and epidemiological and statistical expertise that may not be available in the majority of systematic review teams. Beside feasibility it is important to clearly define the decision making context of a systematic review. This includes a characterization of the decision makers that need to be informed and of their needs to solve decisional conflicts. Defining the decision making context helps defining the exact research question, the in- and exclusion criteria for studies (PICO and aspects of study design) and the quantitative approach for benefit and harm assessment that will fulfill as many of the desired properties as possible. As a result, it may be that for some decision making contexts available randomized trials provide all evidence needed and that one of the simpler approaches for quantitative benefit and harm assessment fulfills important desired properties as outlined above. If the number of benefit and harm outcomes becomes larger, the time horizon longer and the patient population more heterogeneous it is likely that only one of the more complex approaches (e.g. PSM, MCDA or Gail/NCI) provide a quantitative benefit and harm assessment that fulfills the described properties and, as a consequence, of meets the decision makers' needs.

The next steps that seem necessary are: (1) to build upon the preliminary overview that the team developed to categorize quantitative approaches (Figure 5) and the key characteristics (Table 2) in order to develop a finer algorithm for making a decision about the quantitative approach for benefit and harm assessment that is most appropriate for a given example; (2) to pilot test such an algorithm; and (3) to finally use the selected quantitative approach for a concrete benefit and harm assessment. The team will follow these steps in the second part of this project (Aim 3), not reported here.

# Chapter 6. Development of Principles

## Background

Most systematic reviews of interventions, including those conducted by EPCs (Evidence based Practice Centers) *inform* the assessment of the balance of benefit and harm, which is Stage III on Figure 2). The assessment of the balance of benefit and harm usually requires specification of the particular decision making context. However systematic reviews may attempt to provide results for various decision making contexts or be unaware of the specific context in which systematic review might be utilized. Evidence synthesis, including both reporting of benefits and harms and quantitative approaches to the assessment of benefits and harms, is usually insufficient in itself to inform the assessment of the *balance* of benefit and harm. This assessment of the balance of benefit and harm is conducted at the population level or the individual level. This assessment of the balance of benefits and harms is usually an act of subjective judgment. This judgment requires qualitative integration of available quantitative data from reporting of benefits and harms, and includes additional consideration such as biological mechanisms, surrogate outcomes and their linkages to clinical outcomes, risk of bias, strength of evidence and applicability some of which may not be quantifiable. Additional information is needed on values and preferences. The assessment of the balance of benefits and harms is thus often the role of decision-makers, or end-users of systematic reviews. Incorporating information on the underlying decision context and the explicit preferences of various decision makers are usually beyond the scope of most systematic reviews. However some systematic reviewers may be asked by a guideline panel to conduct a systematic review that could result in *an assessment of the balance of benefit and harm* by the guideline panel.[77] In certain situations, the quantitative approaches may aid, inform, support and explicate qualitative judgements about the assessment of balance of benefit and harm by decision makers.

## Process

A principle assists when making a decision or considering a matter. These principles below are meant to inform the conduct, reporting and analysis of systematic reviews, and may be useful for systematic reviews. The team aimed to bridge the interface between evidence synthesis and the subsequent decision making process to facilitate a rigorous assessment of the balance of benefits and harms of interventions. To develop principles for the assessment of the balance of benefit and harms in systematic reviews the team reviewed the above list of challenges, the role of values and preferences, the previously defined key characteristics and the desired properties of the various quantitative approaches and developed preliminary principles. These principles addressed various stages of the systematic review process including protocol development, conduct, reporting, analysis, presentation and the role of preferences in systematic reviews. These preliminary principles were informed and revised after electronic input was received from a diverse 5 member Technical Expert panel.

The team developed the following principles to guide our work. Some of these principles outlined below, such as identifying key outcomes and presenting information fundamental to the assessment of benefit and harm are part of the Methods Guide on Comparative Effectiveness reviews of Interventions, [17] and are thus noted. Others, such as the principles on incorporation of

preferences, choice of various quantitative approaches, their underlying assumptions, and conveying uncertainity are specific to the charge of this project, and may merit consideration by other systematic reviewers who wish to conduct an assessment of the balance of benefit and harm.

# Principles for Benefit and Harm Assessment in Systematic Reviews

## Overarching Principles

**Present benefits and harms in such a way as to enable the assessment of the balance of the benefits and harms in different decision-making contexts.**

***Explanatory text:*** The assessment of the balance of benefit and harm assessment should consider the underlying decision context, such as the approval of therapies by regulatory bodies or the development of clinical practice guidelines and quality standards, and the needs of the decision-makers in different scenarios. For example the harm-benefit profile of chloramphenicol to treat enteric fever which is associated with the rare occurrence of aplastic anemia, may be different in the developed world, where many treatment alternatives are available, compared to the developing world where such alternatives do not exist.[78] The choice of the appropriate quantitative approach for the assessment of the balance of benefits and harms will depend on the needs of the decision makers, such as ease of interpretation vs. the need for uncertainty bounds around the estimates. For example, if the balance of benefits and harms suggests that there is high quality evidence that for most people the benefits clearly outweigh harms, this may translate into a strong recommendation, and ultimately this recommendation may be appropriate to turn into a quality standard which would incentivize providers and health delivery systems to increase the use of the therapy or service for their patients. However, there may be situations where combination of the quality of evidence and the balance of benefits and harms is more tenuous. Factors that might affect this are the degree that there is known heterogeneity in the balance of benefits and harms in the population that are expected will use the therapy or service, or the degree that there is uncertainty about the known balance of benefits and harms or how well it applies to some parts of the populations. Similarly, the degree to which preferences about the relative importance of benefits and harms vary is more crucial for some decision makers, and this variability may often be unknown. This would suggest the need for a method which would transparently allow an end-user see how variability in the preferences affects the ultimate balance of benefits and harms. This overarching principle relates to several of those below.

## Systematic Review Protocol Development

**Identify the key potential harms in the assessment <u>and</u> the key (usually clinical) outcomes that will be considered as benefits.**

***Explanatory text:*** Systematic reviews should attempt to inform the harm and benefit assessment by identifying the key harms and benefits, usually clinical outcomes in the analytic framework. This is already standard guidance for systematic reviews and is outlined in the Methods Guide for Comparative Effectiveness Reviews. The team includes this statement here as identification of both key harms and benefits is essential for an assessment of the balance of benefits and harms, and there are implications of this statement for the assessment of the balance

of benefits and harms. Systematic reviewers should carefully consider how treatment effects on multiple benefit outcomes and multiple harm outcomes will be combined, and how patient preferences for these different outcomes will be incorporated. When data on clinical outcomes are unavailable, reviewers should report information on benefits on potential surrogate outcomes, but also assess the strength of their linkage to clinical outcomes. In circumstances where only benefits on potential surrogate outcomes are available and their linkages to clinical outcomes are not validated, the potential value of a benefit harm assessment on surrogate outcomes should be discussed with end-users.

Systematic reviews should be explicit about whether evidence on patient and decision maker preferences were systematically evaluated or implicitly informed the systematic review process via the representation of a diverse panel of experts. To the extent that SRs report on a portion (i.e. those that are determined by the Systematic Reviewers to be key) of all the outcomes (both harms and benefits) in the literature, they may not be preference neutral. There is much more reliable and robust data on benefits compared to harms. To identify key harms systematic reviewers may have a lower threshold for harms and include study designs other than randomized clinical trials[15], and conduct systematic searches for unpublished or "gray literature" information such as contacting authors of included studies to gather relevant data (e.g. for unpublished outcomes) to make assessment of the balance of benefits and harms as comprehensive as possible. This is already standard guidance for systematic reviews and is outlined in the Methods Guide for Comparative Effectiveness Reviews.[17] However, the need for comprehensiveness should be balanced by the need to consider the reliability and validity of this evidence and the impact of this on the uncertainty of the estimates of the balance of benefit and harm.

**Report the different characteristics and assumptions of the various quantitative approaches to assess the balance of benefit and harm to determine the appropriate approach for a specific decision making context.**

***Explanatory text:*** This principle only applies if the systematic review conducts a quantitative approach for the assessment of benefits and harms or collaborates with another group that will do such an analysis based on the systematic review. The various quantitative approaches differ in key characteristics: While most quantitative approaches for benefit harm assessment are feasible using aggregate or individual patient data, some methods (e.g. "Benefit less risk analysis") require individual patient data. Some are typically based on randomized trials only (ratio of number needed to treat [NNT] over number needed to harm [NNH]), while others can consider both experimental and observational data (e.g. probabilistic simulation modeling, Gail/National Cancer Institute approach or Multi Criteria Decision Analysis). Another key characteristics of a quantitative approaches for benefit harm assessment is the number and diversity of benefit and harm outcomes that are incorporated. While some approaches typically focus on a single or few benefits and harm outcomes of similar severity (e.g. NNT/NNH ratio, TURBO) others incorporate a potentially large number of outcomes of different importance (Gail/National Cancer Institute approach or Multi Criteria Decision Analysis). Since there is little empirical data to suggest that one particular quantitative approach to assessment of benefits and harms is superior to another, systematic reviews should consider which quantitative approach is most appropriate for the decision making context, the data that are available through the systematic review and the methodological expertise of the review team

**State whose preferences were considered, how these were ascertained, and how variation in preferences would affect the results of an assessment of the balance of benefits and harms.**

***Explanatory Text:*** Preferences affect an assessment of benefit and harms across the path of the translation of evidence. (see figure 4) In current practice, these preferences are often implicit, and not necessarily recognized and transparently reported by those conducting an assessment of benefits and harms. This is true whether the assessment of benefits and harms is done through reporting of benefits and harms only, or whether there are quantitative approaches taken that provide a benefit harm comparison estimate. Preferences also affect decision-making on the population-level undertaken by guideline developers and regulatory bodies. (see figure 2) Systematic reviews should explicate and justify the rationale for their choice of various study designs or various outcomes that inform the assessment of benefits and harm and any other steps in the process of evidence synthesis where preferences play a key role. They could consider performing systematic searches for studies on patient preferences for relevant outcomes, or consider performing sensitivity analysis to determine the impact of varying preferences on the assessment of benefit and harm.

**If an assessment of B&H balance is undertaken, define whether (yes/no) a general assessment or a quantitative approach for benefit-harm assessment will be done. If a quantitative approach will be done, systematic reviews should define which specific quantitative approach and benefit-harm comparison metric will be used.**

***Explanatory text:*** A qualitative assessment of the results of a systematic review for benefit and harm outcomes is a general assessment. These results may or may not be summarized using meta-analysis. A general assessment is currently the most common approach for the assessment of the balance of benefits and harms by systematic reviewers (e.g. EPCs, Cochrane Collaboration) but also by policy makers. As indicated in Figure 3, a quantitative assessment of benefits and harms using a benefit-harm comparison metric may be informative for the third step, of decision-making. Systematic reviews should be explicit whether they endorse such a general assessment or choose to use a quantitative approach that will provide a benefit harm comparison metric (e.g. NNT/NNH ratio, QALYs etc).If a quantitative approach is chosen the systematic review protocol should outline which approach is chosen, and their justification for the choice of that particular approach. The choice of a particular approach has implications for (1) the type of evidence (e.g. randomized trials +/- observational studies) that needs to be identified, for the (2) electronic and non-electronic search strategy, for (3) the involvement of specific stakeholders if their preferences need to be considered, for (4) the type of outcome data to be extracted or asked for from primary studies or for (5) the methodological-statistical expertise needed to carry out the handling of multiple data sources and the statistical analyses. Finally, the quantitative assessment of benefit and harm should be done on an absolute and not relative scale.

**Convey all convey all major sources of uncertainty, uncertainty that arises from sampling as well as that due to quality of the evidence, in the assessment of the balance of benefit and harm.**

***Explanatory text:*** The various sources of uncertainty should be described qualitatively along with the balance of benefit and harm assessment. Systematic reviews should be as explicit as possible concerning the applicability of the evidence on harm to particular subgroups such as

older adults or people with important comorbidity. Previous work has noted that uncertainty can arise from at least 5 different types of issues: inherent uncertainty regarding the future; uncertainty regarding validity; uncertainty regarding significance for the individual; uncertainty related to complexity, e.g. the fact that the risks for several outcomes each change with time; or uncertainty related to what is not known.[79]

## Conduct and Reporting of Systematic Reviews

**Report information fundamental to the assessment of the balance of benefits and harms ("information preserving").**

***Explanatory text:*** This has the advantage of allowing the users of SR to calculate any metric that compares the event rates in the two treatment groups, if they desire. For example, presenting event rate (e.g., # of events per 1000 person-year) of an outcome for each treatment group allows calculation of both risk difference and relative risk and is an information preserving approach, whereas if only the relative risk were to be presented it would not be possible to work backwards to obtain the risk difference. The AHRQ methods guide advises that absolute risks should be reported along with relative risks.[17] Another example of this principle is how the outcomes are combined. An information preserving approach would report the event rates for each individual outcome in each treatment group. Combining all the benefit outcomes into a single outcome (e.g., a composite outcome) is not information preserving,

**State how decisions about comparisons, outcomes, baseline risks, and time horizons were made to increase transparency and traceability.**

***Explanatory text:*** To enable end-users to replicate these methods, systematic reviews should state how decisions about comparisons, outcomes, baseline risks and time-horizons were made to increase transparency and traceability. They need to provide the sources of data and assumptions used in conducting a benefit and harm assessment. For example in the case of NNT/NNH ratio, systematic reviews should not only provide information on the underlying meta-analysis for the intervention which informed the Relative or absolute risk estimates but also whether the baseline rate used was the control event rate in the trials or some other population-based study and over what period of time these estimates apply. They should also clarify that relative benefits and risks are assumed to be constant over time and applicable to different study populations and emphasize that a NNT/NNH ratio implies equal weights to treatment benefit and harm. Some of these assumptions are unverifiable and inherent to the respective quantitative approach. For example, some methods require the assumption that effect estimates do not vary across populations. On the other hand, assumptions about constant risks over time, or variations in the contour of benefit and harm over time, could be verified with either individual patient data or closer examination of summary data on benefit and harm. Traceability means that both the effect sizes from the systematic review and meta-analysis and the estimates of baseline risks from the trials or observational studies that inform the quantitative approaches should be clearly identified and sourced. There should be an adequate justification for the choice of these estimates to enable end-users to determine whether these assumptions fit their needs.

**Limitations:** There are limitations to our report. Since we did not conduct a new systematic review to identify the relevant approaches, we may have also missed a small number of benefit and harm approaches. However we believe this is likely to be minimal. Additionally, several of

the identified approaches share similarities as well as subtle differences and we have reviewed the most commonly used approaches for a single study in the context of systematic reviews. It is possible that we did not identify all the relevant challenges for benefit harm assessment in the context of a systematic review since we reviewed a small sample of EPC reports. However the sample was illustrative of the major challenges facing systematic reviews in conducting a benefit and harm assessment. There may be other key characteristics that we may have not recognized. Our principles are not meant to be prescriptive. We hope that these principles will be informative in future methodological efforts at conducting benefit and harm assessment in systematic reviews.

# Chapter 7. Conclusion

The team outlined several challenges in the conduct, reporting, analysis, and presentation of benefits and harms in systematic reviews of interventions. Systematic reviews should explicate the role of preferences of various decisionmakers during the conduct of the systematic review. Careful attention to the common as well as the unique characteristics of the outlined quantitative approaches for the assessment of the balance of benefit and harm should inform the second part of this project (called Aim 3) where the team implements these guiding principles through choosing a quantitative approach for an existing systematic review. Ultimately, similar processes will enable systematic reviewers to choose one or more suitable approaches that meet the needs of their end-users. The above outlined guiding principles should aid systematic reviewers with bridging the gap between evidence synthesis and its translation to evidence based public policy.

# Reference List


1 Bennett WL, Wilson LM, Bolen S *et al*. Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update [internet]. Agency for Healthcare Research and Quality (US); 2011 Mar. Report No.: 11-EHC038-EF 2011.

2 Matchar DB, McCrory DC, Orlando LA *et al*. Comparative Effectiveness of Angiotensin-Converting Enzyme Inhibitors (ACEIs) and Angiotensin II Receptor Antagonists (ARBs) for Treating Essential Hypertension [Internet]. Agency for Healthcare Research and Quality (US); 2007 Nov. Report No.: 08-EHC003-EF. 2007.

3 Bravata DM, McDonald KM, Gienger AL *et al*. Comparative Effectiveness of Percutaneous Coronary Interventions and Coronary Artery Bypass Grafting for Coronary Artery Disease. Agency for Healthcare Research and Quality (US); 2007 Oct. Report No.: 08-EHC002-EF 2007.

4 Coleman CI, Baker WL, Kluger J *et al*. Comparative Effectiveness of Angiotensin Converting Enzyme Inhibitors or Angiotensin II Receptor Blockers Added to Standard Medical Therapy for Treating Stable Ischemic Heart Disease. Agency for Healthcare Research and Quality (US); 2009 Oct. Report No.: 10-EHC002-EF 2009.

5 Bolen S, Wilson L, Vassy J *et al*. Comparative Effectiveness and Safety of Oral Diabetes Medications for Adults With Type 2 Diabetes . Agency for Healthcare Research and Quality (US); 2007 Jul. Report No.: 07-EHC010-EF 2007.

6 Hochman M, McCormick D. Characteristics of published comparative effectiveness studies of medications. JAMA 2010; 303(10):951-8.

7 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; 13(5):657-66.

8 Bennett WL, Maruthur NM, Singh S *et al*. Comparative effectiveness and safety of medications for type 2 diabetes: an update including new drugs and 2-drug combinations. Ann Intern Med 2011; 154(9):602-13.

9 Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007; 356(24):2457-71.

10 Bangalore S, Kumar S, Messerli FH. Angiotensin-converting enzyme inhibitor associated cough: deceptive information from the Physicians' Desk Reference. Am J Med 2010; 123(11):1016-30.

11 O'Neill RT. A Perspective on Characterizing Benefits and Risks Derived From Clinical Trials: Can We Do More? Drug Information Journal. 2008; 42, no. 3,pp235-45.

12 Hughes DA, Bayoumi AM, Pirmohamed M. Current assessment of risk-benefit by regulators: is it time to introduce decision analyses? Clin Pharmacol Ther 2007; 82(2):123-7.

13 Gail MH, Costantino JP, Bryant J *et al*. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 1999; 91(21):1829-46.

14 Report of the Committee for Medicinal Products Working Group on Benefit-Risk Assessment Models and Methods. European Medicine Agency 2007 Doc. Ref. EMEA/CHMP/15404/2007 Copyright @ EMEA 2007.

15 Chou R, Aronson N, Atkins D *et al*. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010; 63(5):502-12.

16 Singh JA, Christensen R, Wells GA *et al*. A network meta-analysis of randomized controlled trials of biologics for rheumatoid arthritis: a Cochrane overview. CMAJ 2009; 181(11):787-96.

17 Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D *et al*. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007; 334(7597):786.

18 Christine M. Micheel and John R. Ball e. Evaluation of biomarkers and surrogate endpoints in chronic disease /Committee on Qualification of Biomarkers and Surrogate Endpoints in Chronic Disease, Board on Health Care Services, Board on Health Sciences Policy, Food and Nutrition Board, Institute of Medicine . The National Academies Press, 2010.

19 Golder S, Loke YK, Bland M. Meta-analyses of Adverse Effects Data Derived from Randomised Controlled Trials as Compared to Observational Studies: Methodological Overview. PLoS Med

2011; 8(5):e1001026.

20 Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. March 2011. Chapters available at: www.effectivehealthcare.ahrq.gov .

21 Guyatt GH, Oxman AD, Vist GE *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336(7650):924-6.

22 Sin DD, Tashkin D, Zhang X *et al*. Budesonide and the risk of pneumonia: a meta-analysis of individual patient data. Lancet 2009; 374(9691):712-9.

23 American Academy of Orthopedic Surgeons Clinical Guideline On Prevention of Pulmonary Embolism in Patients Undergoing Total Hip or Knee Arthroplasty. May 2007.

24 Sackett D.L., Richardson WS, Rosenberg W, Haynes RB. Evidence Based Medicine. How to Practice and Teach EBM. Churchill Livingstone, Edinburgh. 1997.

25 JAMAevidence Glossary [Web Page]. Available at http://jamaevidence.com/glossary/V.

26 Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. JAMA 2008; 300(4):436-8.

27 Guyatt G, Cook D, Haynes B. Evidence based medicine has come a long way. BMJ 2004; 329(7473):990-1.

28 Guyatt G, Montori V, Devereaux PJ, Sch++nemann H, Bhandari M. Patients at the centre: in our practice, and in our use of language. Evidence Based Medicine 2004; 9(1):6-7.

29 D'Agostino RBJ . Debate: The slippery slope of surrogate outcomes. Curr Control Trials Cardiovasc Med 2000; 1(2):76-8.

30 Singh S, Loke YK, Furberg CD. Long-term risk of cardiovascular events with rosiglitazone: a meta-analysis. JAMA 2007; 298(10):1189-95.

31 Loke YK, Price D, Herxheimer A. Systematic reviews of adverse effects: framework for a structured approach. BMC Med Res Methodol 2007; 7:32.

32 Guyatt GH, Oxman AD, Kunz R *et al*. Going from evidence to recommendations. BMJ 2008; 336(7652):1049-51.

33 Methods and Processes. U.S. Preventive Services Task Force. http://www.uspreventiveservicestaskforce.org/methods.htm [Web Page].

34 Guyatt G, Akl EA, Hirsh J *et al*. The vexing problem of guidelines and conflict of interest: a potential solution. Ann Intern Med 2010; 152(11):738-41.

35 Protheroe J, Fahey T, Montgomery AA, Peters TJ. Effects of patients' preferences on the treatment of atrial fibrillation: observational study of patient-based decision analysis. West J Med 2001; 174(5):311-5.

36 John E. Wennberg. Tracking Medicine: A Researcher's Quest to Understand Health Care. Oxford University Press, 2010.

37 Fisher ES, Wennberg JE. Health care quality, geographic variations, and the challenge of supply-sensitive care. Perspect Biol Med 2003; 46(1):69-79.

38 Sussman JB, Vijan S, Choi H, Hayward RA. Individual and population benefits of daily aspirin therapy: a proposal for personalizing national guidelines. Circ Cardiovasc Qual Outcomes 2011; 4(3):268-75.

39 Institute of Medicine (IOM). Knowing what works in health care: A roadmap for the Nation. Washington, DC: The National Academies Press, 2008.

40 . IOM(Institute of Medicine) 2011. Clinical Practice Guidelines We can Trust. Washington, DC: The National Academies Press.

41 Elwyn G, O'Connor A, Stacey D *et al*. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ 2006; 333(7565):417.

42 Chong J PPLBBP . Phosphodiesterase 4 inhibitors for chronic obstructive pulmonary disease. Cochrane Database of Systematic Reviews 2011, Issue 5. Art. No.: CD002309. DOI: 10.1002/14651858.CD002309.pub3. 2011.

43 Singer DE, Albers GW, Dalen JE *et al*. Antithrombotic therapy in atrial fibrillation: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest 2008; 133(6 Suppl):546S-92S.

44 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; 281(19):1836-43.

45 Dolan JG. Multi-criteria clinical decision support: A primer on the use of multiple criteria decision making methods to promote evidence-based, patient-centered healthcare. Patient 2010; 3(4):229-48.

46 Singh S, Dolan JG, Centor RM. Optimal management of adults with pharyngitis--a multi-criteria decision analysis. BMC Med Inform Decis Mak 2006; 6:14.

47 Walter SD, Sinclair JC. Uncertainty in the minimum event risk to justify treatment was evaluated. J Clin Epidemiol 2009; 62(8):816-24.

48 Chuang-Stein C. A new proposal for benefit-less-risk analysis in clinical trials. Control Clin Trials 1994; 15(1):30-43.

49 Boers M, Brooks P, Fries JF, Simon LS, Strand V, Tugwell P. A first step to assess harm and benefit in clinical trials in one scale. J Clin Epidemiol 2010; 63(6):627-32.

50 Aspirin for the prevention of cardiovascular disease: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2009; 150(6):396-404.

51 Ponce RA, Bartell SM, Wong EY *et al*. Use of quality-adjusted life year weights with dose-response models for public health decisions: a case study of the risks and benefits of fish consumption. Risk Anal 2000; 20(4):529-42.

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

53 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; 16 Suppl 1:S2-S15.

54 Mussen F, Salek S, Walker S, Phillips L. A quantitative approach to benefit-risk assessment of medicines - part 2: the practical application of a new model. Pharmacoepidemiol Drug Saf 2007; 16 Suppl 1:S16-41.

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

56 Dolan JG. Shared decision-making--transferring research into practice: the Analytic Hierarchy Process (AHP). Patient Educ Couns 2008; 73(3):418-25.

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

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

59 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995; 310(6977):452-4.

60 Osiri M, Suarez-Almazor ME, Wells GA, Robinson V, Tugwell P. Number needed to treat (NNT): implication in rheumatology clinical practice. Ann Rheum Dis 2003; 62(4):316-21.

61 Lynd LD, Najafzadeh M, Colley L *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; 13(4):411-7.

62 Shaffer ML, Watterberg KL. Joint distribution approaches to simultaneously quantifying benefit and risk. BMC Med Res Methodol 2006; 6:48.

63 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; 57(8):795-803.

64 Gelber RD, Goldhirsch A, Cole BF. Evaluation of effectiveness: Q-TWiST. The International Breast Cancer Study Group. Cancer Treat Rev 1993; 19 Suppl A:73-84.

65 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; 88(15):1039-45.

66 Sherrill B, Amonkar MM, Stein S, Walker M, Geyer C, Cameron D. Q-TWiST analysis of lapatinib combined with capecitabine for the treatment of metastatic breast cancer. Br J Cancer 2008; 99(5):711-5.

67 Shakespeare TP, Gebski VJ, Veness MJ, Simes J. Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. Lancet 2001;

357(9265):1349-53.

68 Ryan M, McIntosh E, Shackley P. Methodological issues in the application of conjoint analysis in health care. Health Econ 1998; 7(4):373-8.

69 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; 29(1):121-36.

70 Hauber AB, Mohamed AF, Johnson FR, Falvey H. Treatment preferences and medication adherence of people with Type 2 diabetes using oral glucose-lowering agents. Diabet Med 2009; 26(4):416-24.

71 Johnson FR, Van Houtven G, Ozdemir S *et al*. Multiple sclerosis patients' benefit-risk preferences: serious adverse event risks versus treatment efficacy. J Neurol 2009; 256(4):554-62.

72 Bachmann LM, Muhleisen A, Bock A, ter Riet G, Held U, Kessels AG. Vignette studies of medical choice and judgement to study caregivers' medical decision behaviour: systematic review. BMC Med Res Methodol 2008; 8:50.

73 Johnson FR, Ozdemir S, Mansfield C *et al*. Crohn's disease patients' risk-benefit preferences: serious adverse event risks versus treatment efficacy. Gastroenterology 2007; 133(3):769-79.

74 Tugwell P, Boers M, Strand V *et al*. OMERACT 9 - 9th International Consensus Conference on outcome measures in rheumatology clinical trials. J Rheumatol 2009; 36(8):1765-8.

75 Levitan B. A concise display of multiple end points for benefit-risk assessment. Clin Pharmacol Ther 2011; 89(1):56-9.

76 Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I *et al*. Validity of composite end points in clinical trials. BMJ 2005; 330(7491):594-6.

77 Wolff T, Miller T, Ko S. Aspirin for the primary prevention of cardiovascular events: an update of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2009; 150(6):405-10.

78 Arjyal A, Basnyat B, Koirala S *et al*. Gatifloxacin versus chloramphenicol for uncomplicated enteric fever: an open-label, randomised, controlled trial. Lancet Infect Dis 2011; 11(6):445-54.

79 Politi MC, Han PK, Col NF. Communicating the uncertainty of harms and benefits of medical interventions. Med Decis Making 2007; 27(5):681-95.

# List of Acronyms/Abbreviations

| Acronym | Definition |
|---|---|
| ACE | Angiotensin-converting Enzyme |
| AHRQ | Agency for Healthcare Research and Quality |
| ARB | Angiotensin II Receptor Blocker |
| ARR | Absolute Risk Reduction |
| B&H | Benefit and Harm |
| BLRA | Benefit-less-risk Analysis |
| CER | Comparative Effectiveness Review |
| CMS | Center for Medicare and Medicaid Services |
| COPD | Chronic Obstructive Pulmonary Disease |
| CPGs | Clinical Practice Guidelines |
| CV | Cardiovascular |
| EBM | Evidence-based Medicine |
| EMA | European Medicines Agency |
| EPC | Evidence-based Practice Centers |
| FDA | Food and Drug Administration |
| GRADE | Group for Analysis and Development |
| HR | Hazard Ratio |
| INHB | Incremental Net Health Benefit |
| IOM | Institute of Medicine |
| IPD | Individual Participant Data |
| KQ | Key Question |
| MAR | Maximum Acceptable Risk |
| MCDA | Multicriteria Decision Analysis |
| MCE | Minimum Clinical Efficacy |
| MERT | Minimum Target Event Risk for Treatment |
| MeSH | Medical Subject Headings |
| NA | Not Applicable |
| NCI | National Cancer Institute |
| NNH | Number-needed-to-harm |
| NNT | Number-needed-to-treat |
| NNT-NNH | Ratio of Number-needed-to-treat over Number-needed-to-harm |
| OMERACT | Outcome Measures in Rheumatology Clinical Trials |
| OR | Odds Ratio |
| PICO | Population Intervention Comparator(s) Outcome(s) |
| PSM | Probabilistic Simulation Methods |
| QALYs | Quality-adjusted Life-years |
| QFRBA | Quantitative Framework for Risk and Benefit Assessment |

| | |
|---|---|
| Q-Twist | (Quality-adjusted) Time Without Symptoms and Toxicity |
| RBAT | Risk-benefit Acceptability Threshold |
| RBC | Risk-benefit Contour |
| RBP | Risk-benefit plane |
| RCTs | Randomized Controlled Trials |
| RR | Relative Risk |
| RRR | Relative Risk Reduction |
| RV-NNT | Relative Value Adjusted Number-needed-to-treat |
| SPM | Stated Preference Method |
| SR | Systematic review |
| TURBO | Transparent Uniform Risk Benefit Overview |