eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

**GUIDELINE**

# LEVELS OF EVIDENCE

## Applicability of evidence for the context of a relative effectiveness assessment

**Adapted version (2015)**

**based on**

**“LEVELS OF EVIDENCE: - Applicability of evidence for the context of a relative effectiveness assessment” - February 2013**

The primary objective of EUnetHTA JA1 WP5 methodology guidelines was to focus on methodological challenges that are encountered by HTA assessors while performing a rapid relative effectiveness assessment of pharmaceuticals.

The guideline “Levels of evidence: applicability of evidence for the context of a relative effectiveness assessment of pharmaceuticals” has been elaborated during Joint Action 1 by experts from ZIN (former CVZ), reviewed and validated by all members of WP5 of the EUnetHTA network; the whole process was coordinated by HAS.

During Joint Action 2 the wording in this document has been revised by WP7 in order to extend the scope of the text and recommendations from pharmaceuticals only to the assessment of all health technologies. Content and recommendations remained unchanged.

This guideline represents a consolidated view of non-binding recommendations of EUnetHTA network members and in no case an official opinion of the participating institutions or individuals.

**Disclaimer:** EUnetHTA Joint Action 2 is supported by a grant from the European Commission. The sole responsibility for the content of this document lies with the authors and neither the European Commission nor EUnetHTA are responsible for any use that may be made of the information contained therein.

# Table of contents



# Acronyms – Abbreviations

| | |
|---|---|
| AHRQ | Agency for Healthcare Research and Quality |
| AHTAPol | Agency for Health Technology Assessment in Poland |
| CADTH | Canadian Agency for Drugs and Technologies in Health |
| DACEHTA | Danish Centre for Health Technology Assessment |
| HIQA | Health Information & Quality Authority (Ireland) |
| HTA | health technology assessment |
| IQWIG | Institute for Quality and Efficiency in Health Care (Germany) |
| INAHTA | International Network of Agencies for Health Technology Assessment |
| NICE | National Institute for Health and Clinical Excellence (United Kingdom) |
| PBAC | Pharmaceutical Benefits Advisory Committee (Australia) |
| PHARMAC | The New Zealand Pharmaceutical Management Agency |
| PICOS | patient intervention comparison outcome setting |
| RCT | randomised controlled trial |
| TLV | Dental and Pharmaceutical Benefits Agency (Sweden) |
| ZIN | Zorginstituut Nederland |

# Summary and recommendations

## *Summary*

Applicability, also known as external validity/ generalisability/ or transposability, is the extent to which the effects observed in clinical studies are likely to reflect the expected results when a specific intervention is applied to the population of interest. In case of a relative effectiveness assessment (REA), the population of interest refers to the patient population that is being assessed as part of the REA. Internal validity is the extent to which the design, conduct, analysis and reporting of a randomised controlled trial (RCT) eliminate the possibility of bias. Bias is defined as the systematic distortion of the estimated intervention effect away from the "truth". ‘Internal validity’ is discussed in more detail in the EUnetHTA guideline on internal validity. The aim of this guideline on applicability is to assess whether there is a relevant effect modification when a specific intervention is applied to the population of interest.

To assess the relative effectiveness of interventions, trials with a pragmatic approach which have more ‘noise of practice’, are more suitable than trials with an explanatory approach that are conducted within a strict trials setting. No trial is completely pragmatic or explanatory, rather every trial can be positioned somewhere between the extremes and has its pragmatic and explanatory elements. In practice, especially at the time of a rapid assessment, trials with a pragmatic approach may not be available. In this instance it is even more important to consider the applicability of the data that are available. A useful instrument to test the applicability is through statistical modelling with, for example, meta-analysis. However, even this type of evidence may not be commonly available, especially in case of a rapid assessment. Moreover, time and resources to do such analysis as part of the assessment may be scarce.

Regardless of the availability of trials with a pragmatic approach or meta-analysis that address applicability, the assessor of a relative effectiveness assessment should always indicate the likeliness that the available evidence is applicable to the decision problems. In order to address the applicability in an assessment report a 4-step process is recommended, after carefully defining the target population. This process was developed by Atkins et al. (2011) and it is based on Patient, Intervention, Comparator, Outcome, Setting (PICOS):
Step 1. Determine the most important factors that may affect applicability (the table in Annexe 1 can be helpful);
Step 2. Systematically abstract and report key characteristics that may affect applicability in evidence tables (highlight studies with a pragmatic approach and data on size of effect modification);
Step 3. Make and report judgements about major limitations to applicability of individual studies;
Step 4. Consider and summarize the applicability of a body of evidence (use format of table 2 in section 2.2.2).
Due to the limited timeframe of a Rapid assessment (e.g. 90 days) the 4-steps process may be considered too labour intensive. However, a summary table of the applicability of the evidence based on the PICOS framework should at least be presented in each relative effectiveness assessment in order to envisage potential applicability problems.

In conclusion, to assess applicability of clinical data for the population of interest, this guideline recommends usage of data from trials with a pragmatic approach. If available, statistical analysis that addresses effect modification of results to a specific/general patient population/setting should be included in the assessment. In addition, to address the applicability of the evidence in each relative effectiveness assessment systematically and in a transparent manner a summary table of the applicability of the evidence based on the PICOS framework should always be presented. Preferably this should be based on the 4-step process that is proposed by Atkins et al. (2011). Evaluating the applicability of the evidence cannot be based on a pre-defined formula. Depending on the topic, interpretation of the applicability may vary. Finally, it should always be considered

whether the relevant elements of applicability are context dependent, and as such should be considered in a local context, or can be addressed in general.

## *Recommendations*

**Recommendation 1**

Applicability is defined as the extent to which the effects observed in clinical studies are likely to reflect the expected results when a specific intervention is applied to the population of interest. Applicability should be considered in each assessment of relative effectiveness. The aim of assessing applicability is to consider whether a relevant effect modification is likely in the population of interest as compared to the results in the clinical studies.

(section 2.1)

**Recommendation 2**

Prior to assessing the applicability, causality between treatment and outcome should be established (internal validity is a pre-requisite of applicability).

(section 2.1)

**Recommendation 3**

To assess the relative effectiveness of interventions, trials with a pragmatic approach are more suitable than trials with an explanatory approach as the results are more likely to occur in clinical practice. If available, data from trials with a pragmatic approach should always be included in the assessment (if the trial has been performed in the population of interest).

(section 2.1)

**Recommendation 4**

If available, analysis that addresses effect modification of results to a specific/general patient population/setting (e.g. effect model, meta-analysis) should be included in the assessment.

(section 2.2.1)

**Recommendation 5**

Assessors should describe differences between available evidence and the ideal evidence to address the question being asked. They should offer a qualitative judgement about the importance and potential effect of those differences.

a) First, the authors should carefully identify and describe the target population

b) It should be noted that the size of the effect modifications (the numerical value of the effect) can only be addressed by statistical methods.

c) The most applicable evidence may differ when considering benefits or harms since these often depend on distinct physiological processes. Therefore applicability should be judged separately for different outcomes.

d) To address the applicability in a report the 4-step process developed by Atkins et al (2011) is recommended:

**Step 1.** Determine the most important factors that may affect applicability (the table in Annexe 1 can be helpful)

**Step 2.** Systematically abstract and report key characteristics that may affect applicability in evidence tables (highlight studies with a pragmatic approach and data on effect size of effect modification).

**Step 3.** Make and report judgements about major limitations to applicability of individual studies.

**Step 4.** Consider and summarize the applicability of a body of evidence (use format of table below)

For details we refer to the guideline by Atkins et al.(2011)

e) For a rapid assessment (limited timeframe) the 4-step process described above may not be feasible. In any case, it is recommended to at least fill in the summary table which will help envisage potential applicability issues.

f) The following aspects are important to include in the description:

- It is likely that not all data are available to complete the table. In case of missing data this should be described as well.
- The section on outcomes should include a comment regarding which effect measure is less likely to be subject to effect modification (e.g. which effect measure is more/less likely to be different in the population of interest in a particular setting than in the available trials).
- It should always be considered and addressed whether a specific element that is relevant for the applicability can be assessed in general or whether this should be done in the local (national) context.

(section 2.2.2)

**Recommendation 6**

It should be noted that evaluating the applicability of the evidence is not a pre-defined formula. Depending on the topic interpretation of the applicability may vary. For example, for a rare disease other considerations and requirements may be relevant compared to a non-rare disease. Regardless of the topic it is very relevant that ***the considerations are transparently reported in the assessment report***.

**Table 1. Elements to be included in a summary table characterising the applicability of a body of studies**

| Domain | Description of applicability of evidence |
|---|---|
| Population | *[Describe general characteristics of enrolled populations, how this might differ from target population, and effects on baseline risk for benefits or harms. Where possible, describe the proportion with characteristics potentially affecting applicability (e.g. % over age 65) rather than the range or average.]* |
| Intervention | *[Describe general characteristics and range of interventions and how they compare to those in routine use, and how this might affect benefits or harms from the intervention.]* |
| Comparators | *[Describe comparators used. Describe whether they reflect best alternative treatment and how this may influence treatment effect size.]* |
| Outcomes | *[Describe what outcomes are most frequently reported and over what time period. Describe whether the measured outcomes and timing reflect the most important clinical benefits and harms.]* |
| Setting | *[Describe geographic and clinical setting of studies. Describe whether or not they reflect the settings in which the intervention will be typically used and how this may influence the assessment of intervention effect.]* |

Source: Atkins et al. 2011

# 1. Introduction

## 1.1. Definitions

- **Applicability:** The extent to which the effects observed in clinical studies are likely to reflect the expected results when a specific intervention is applied to the population of interest. The aim of assessing applicability is to assess whether a relevant effect modification is likely in the population of interest.
- **Relative effectiveness:** can be defined as the extent to which an intervention does more good than harm compared to one or more intervention alternatives for achieving the desired results when provided under the usual circumstances of health care practice (Pharmaceutical Forum 2008).
- **Health technology assessment**: The systematic evaluation of properties, effects, and/or impacts of health care technology. It may address the direct, intended consequences of technologies as well as their indirect, unintended consequences. Its main purpose is to inform technology-related policymaking in health care. Health technology assessment is conducted by interdisciplinary groups using explicit analytical frameworks drawing from a variety of methods (INAHTA).
- (**Single) Rapid assessment of relative effectiveness**: defined as rapid assessment of a new technology at the time of introduction to the market and comparing the new technology to standard of care. This will be referred to hereafter as the ***rapid assessment****;*
- **(Multiple) Full assessment of relative effectiveness**: defined as full assessment (non-rapid) of (all) available technolog(y)(ies) for a particular step in a treatment pathway for a specific condition. This will be referred to hereafter as the ***full assessment****.*
- **Effect modification:** when characteristics of the patient, intervention, or setting modify the relative effect of the intervention on the main outcome (Atkins et al. 2011)

## 1.2. Context

### 1.2.1. Problem statement

Clinical studies must be internally valid. But to be clinically useful, the result must also be applicable to a definable group of patients, in a particular clinical setting, for which the health technology assessment (HTA) is required. This is especially relevant for assessing the relative effectiveness of an intervention, which focuses on results when provided under the usual circumstances of health care practice. Assessing whether the results of a clinical trial are also relevant to a definable group of patients in a particular clinical setting is the concept of ‘applicability’.

## 1.3. Scope/Objective(s) of the guideline

This guideline addresses the following question:
How to assess whether there is a relevant modification of the effect of the results in the clinical studies (e.g. a RCT) if the intervention is applied to the population of interest in clinical setting?

The guideline is intended to provide recommendations to the assessor of the relative effectiveness of an intervention in the context of a reimbursement request (a rapid assessment soon after market authorisation). The recommendations in this guideline are based on a systematic review of literature in combination with expert involvement from national health technology assessment (HTA) agencies.

**The following is excluded from the scope:**
The current guideline focuses on evidence from controlled trials as – especially for pharmaceuticals, on which the first version of this guideline was focussed - this type of evidence is most commonly available soon after market authorisation. Hence, ***this guideline does not address evidence from observational studies***.
Whereas the use of modelling techniques is common for pharmacoeconomic analysis, it is not common to use them for relative effectiveness assessments (this resulted from the literature review for the preceding version of this guideline as well as the JA1 WP5 background review). Therefore ***modelling techniques to address applicability*** are not discussed in this version of the guideline. Interpretation of statistical methods to address effect modification is discussed in Annexe 2.

## 1.4. Relevant EUnetHTA documents

This document should be read in conjunction with the following document:

- EUnetHTA guideline on levels of evidence: internal validity

# 2. Summary of the literature

## 2.1. Introduction

### The concept of applicability

There is broad consensus that RCTs should be the basis for developing clinical guidelines and for decisions about individual patient management. They should also inform public health policy (Seale et al. 2004). However, their capacity to fulfil these roles will depend on how closely the trial results reflect the results observed in the intended population when provided under the usual circumstances of health care practice. The observation that effectiveness of an intervention varies in different populations or settings is known as heterogeneity of treatment effect. One cause of heterogeneity is true effect modification, defined when characteristics of the patient, intervention, or setting modify the relative effect of the intervention on the main outcome (Atkins et al. 2011).

RCTs must be internally valid, i.e. the design and conduct must reduce the possibility of bias (for more details on the concept of internal validity see the EUnetHTA guideline ‘Internal validity’). But to be clinically useful, the result must also be relevant to a definable group of patients in a particular clinical setting (i.e., they must be externally valid) (Rothwell et al. 2006).

There is not a single definition of applicability that is widely used or accepted and the terms applicability, external validity, generalisability and transposability are used interchangeably in the literature. Various definitions were identified, as presented in Box 1.

**Box 1. Definitions commonly referred to in literature**

| |
|---|
| The National Institute for Health and Clinical Excellence (NICE) uses the term external validity and has defined it as the degree to which the results of an observation, study or review are likely to hold true in a population or clinical practice setting outside of the study population/setting (NICE, 2008). |
| The CONSORT group uses the term external validity and defines it as the extent to which the results of a trial provide a correct basis for generalisations to other circumstances. Also called "generalisability’ or "applicability" (CONSORT glossary). |
| Dekkers et al. (2009) have made a distinction between ‘external validity’ and ‘applicability’. External validity refers to the question of whether the study results are valid for patients, other than those in the original study population, in a treatment setting that is in all respects equal to the treatment setting of the original study. External validity therefore involves patient and disease characteristics. In contrast, applicability is referred to as the question of whether study results are valid for patients to whom results are generalisable but who are in a different treatment setting than the original study population. Consequently, applicability involves characteristics of the treatment setting. |
| The Agency for Healthcare Research and Quality (AHRQ) uses the term applicability, which is defined as the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under “real-world” conditions (Atkins et al. 2011). |

The aim of this guideline is to provide guidance on how to assess whether there is a relevant effect modification in the population of interest. Therefore we choose to use the term applicability, which we define as the extent to which the effects observed in clinical studies are likely to reflect the expected results when a specific intervention is applied to the population of interest. In case of a REA, the population of interest refers to the patient population that is being assessed as part of the REA.

### Explanatory vs pragmatic approach

For regulatory purposes a distinction is made between exploratory trials and confirmatory trials. The first type of trial aims at, for example, exploring the use for the targeted indication or estimating the dosage of a pharmaceutical for subsequent studies. The latter aims at, for example, demonstrating/confirming the efficacy or establishing a safety profile (EMA, 1998).

In the context of HTA, researchers commonly refer to explanatory approach vs pragmatic approach which was first introduced by Schwartz et al. in 1967 (Schwartz et al.1967&2009). They proposed a distinction between trials that aim at confirming a physiological hypothesis, precisely specified as a causal relationship between administration of an intervention and some physiological outcome (which they called an '***explanatory'*** approach) and trials that aim at informing a clinical, health service or policy decision, where this decision involves the choice between two or more interventions (called a '***pragmatic***' approach). These explanations may be a bit confusing as answers that help users choose between options of care also address questions of causal relationship. It may be rather the ‘noise of practice’ that differs pragmatic from explanatory and not the general aim of identifying and quantifying causal effects (Windeler 2010). The ‘noise of practice’ refers to a trial setting that corresponds to usual circumstances of healthcare instead of a strict protocol driven setting that is used in trials of explanatory nature.

It should be noted that the difference between explanatory and pragmatic approach is a continuum rather than a dichotomy between trials (Treweek et al. 2009). There is no such thing as a pragmatic trial or an explanatory trial, rather every trial can be positioned somewhere between the extremes and has its pragmatic and explanatory elements (Windeler 2010). For example, in a trial with an otherwise explanatory approach, there may be some aspects of the intervention that are beyond the investigator’s control. Similarly, the act of conducting an otherwise pragmatic approach may impose some control resulting in the setting being atypical. For example, the very act of collecting data required for a trial that would not otherwise be collected in usual practice could be a sufficient trigger to modify participant behaviour in unanticipated ways (Thorpe et al. 2009).

For relative effectiveness assessments, trials with a pragmatic attitude can be of great value as the results may be more applicable to the population of interest in clinical setting. However, these trials may be affected by the local clinical practices resulting in limited transferability and generalisability to other (local) settings. In addition, there should be a balance between making eligibility criteria pragmatic and broad which rely heavily on the clinical judgement of investigators, and making them very detailed to avoid any ambiguity as internal validity is a prerequisite for the applicability (Flather et al. 2006, Dekkers et al. 2009). Study results that deviate from the true effect due to systematic error (e.g. are not internally valid) lack basis for applicability (Dekkers et al. 2009).

## *2.2. How to address the applicability of trial data?*

Methods to address the applicability are not well developed yet, although there is an increasing interest. The following sections summarise currently available methods that can be used to address the applicability for relative effectiveness assessment. In section 2.2.1 we will discuss how to interpret data that try to quantify the applicability (statistical methods) such as effect model and meta-analysis. In section 2.2.2 tools will be discussed that can help to explore the applicability of data in a qualitative manner, and how this can be presented in an assessment report.

One should always keep in mind that as applicability depends on a target population, the first step in the assessment of the applicability is to define this target population (Romijn et al 2010).

### 2.2.1. Statistical methods

Effect modification may ideally be estimated through statistical modelling. Here the influence of one or more features of a trial, such as the selection of participants, is investigated using statistical techniques to see how sensitive the trial result is to the feature or features being varied (Treweek et al. 2009). However, these are mostly limited to assessment of one aspect of applicability. It is unlikely that within the timeframe of an HTA (especially a rapid assessment) assessors will have

the opportunity to do these types of analysis/modelling. Annexe 2 focuses on how assessors can interpret analysis/models that are already published.

### 2.2.2. Summary table to address applicability

Apart from statistical methods to assess the applicability, determinants of the applicability of an RCT requires clinical rather than statistical expertise, and often depends on a detailed understanding of the particular clinical condition under study and its management in routine clinical practice (Rothwell et al. 2006). For the assessment of internal validity of a trial, many widely-used checklists exist such as the CONSORT statement, the Jadad Scale, the CLAR-NPT checklist and the PEDro checklist. However none of these checklists and scales put emphasis on the applicability of the trial results. The CONSORT statement for example attributes only one out of 25 items to the applicability (Zwarenstein et al. 2008). This is explainable as in contrast to the accumulating body of empirical data on factors affecting the risk of bias, or internal validity, there has been less empiric data to determine which factors affect applicability (Atkins et al. 2011). In addition, applicability is a matter of a certain situation (are the results of the trial applicable to the patient population you want to treat in clinical practice) and not a matter of a certain trial (a trial can not be ‘applicable’ in general).

Several authors have discussed the relevance of applicability and listed a number of criteria that are relevant to determine the applicability of the trial data (Dekkers et al. 2009; Green et al. 2006; Flather et al. 2006; Julian et al. 1997; Rothwell et al. 2006; Seale et a. 2004). These lists with criteria are summarised in Annexe 4. There is variance in level of detail of the lists/criteria. Some focus only on the patient population/participants whereas other lists also take into account (some of the) following subjects: the study design, treatment setting, the treatment, outcome measures and follow-up, outcomes for decision making, and conclusion.

It should be noted that most of these lists are intended for checking good clinical practice and are not developed from the viewpoint of the decision maker. There is no study which has tested the value for usage of these lists for health technology assessment doers. A usable checklist should be comprehensive but also feasible to assess on multiple trials within the limited timeframe of a health technology assessment. This is especially true for a rapid assessment[1].

In HTA methodology guidelines the concept of applicability is frequently mentioned (DACEHTA 2007; HIQA 2010; Hungary 2002; IQWIG 2008; NICE 2008; PBAC 2008), which is confirmed by the findings of the background survey of WP5 during JA1, in which all 28 countries surveyed indicated to at least sometimes consider the generalisability of trial data for a relative effectiveness assessment (Kleijnen et al. 2011). However the guidelines and the agencies generally do not refer to or recommend a specific instrument to be used to assess the applicability of a trial.

Only the New Zealand Pharmaceutical Management Agency (PHARMAC) has concretely phrased three questions to assess the applicability (PHARMAC, 2010). The authors of this EUnetHTA guideline consider that these questions do not directly address the most important element of applicability: whether these items result in a different effect when the treatment is provided to the patient population of interest in usual practice.

Recently an article was published with more detailed guidance on how to assess applicability (Atkins et al. 2011). This guidance document was specifically developed because of the unmet need for detailed guidance for assessing applicability of evidence in producing systematic reviews. This is especially relevant for comparative effectiveness reviews that aim to assess the effect of an intervention in the real world. The key points are summarised below. For details, we refer to the

[1] (single) rapid assessment of relative effectiveness of pharmaceuticals is defined in WP5 as a rapid assessment of a new technology at the time of introduction to the market and comparing the new technology to standard care

original document (Atkins D, Chang S, Gartlehner G, Buckley DI, Whitlock EP, Berliner E, Matchar D. Assessing applicability when comparing medical interventions: Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol. 2011 Apr 2):

- Because applicability depends on the specific questions and needs of the users, it is difficult to devise a valid uniform scale for rating the overall applicability of individual studies or body of evidence.
- The Patient, Intervention, Comparator, Outcome, Setting (PICOS) framework is a useful way of organising the review and presentation of factors that affect applicability.
- Input from clinical experts and stakeholders can help identify specific study elements that should be routinely abstracted to examine applicability.
- Population-based surveys, pharmacoepidemiologic studies, and large case series or registries of devices or surgical procedures can be used to determine whether the populations, interventions, and comparisons in existing studies are representative of current practice.
- Reviewers should assess whether benefits or harms vary along with differences in patient or intervention characteristics (i.e., effect modification) or with differences in underlying risk.
    - The most applicable evidence may differ when considering benefits or harms since these often depend on distinct physiologic processes. Therefore applicability should be judged separately for different outcomes. This is illustrated by the following example in the AHRQ guideline. Evidence of the benefits of aspirin for prevention of cardiovascular events from patients with heart disease cannot be readily applied to healthy populations. However, studies of patients with and without heart disease may be useful for estimating the gastrointestinal risks of aspirin which act through different mechanisms and do not vary with underlying cardiac risk;
- Reports should clearly highlight important issues relevant to applicability of individual studies in a ‘‘Comments’’ or ‘‘Limitations’’ section of evidence tables and in text.
- Metaregression, subgroup analysis, and/or separate applicability summary tables may help reviewers, and those using the reports see how well the body of evidence applies to the question at hand.
- Judgments about applicability of the evidence should consider the entire body of studies.
- Important limitations of the applicability of the evidence should be described within each summary conclusion.

To address the applicability in a report a 4-step process is recommended:
Step 1. Determine the most important factors that may affect applicability (the table in Annexe 1 can be helpful);
Step 2. Systematically abstract and report key characteristics that may affect applicability in evidence tables (highlight studies with a pragmatic approach and data on effect size of effect modification);
Step 3. Make and report judgements about major limitations to applicability of individual studies;
Step 4. Consider and summarise the applicability of a body of evidence (Table 2).

Atkins et al. 2011 developed a table that is useful to summarise the important limitations of the applicability of the evidence (Table 2).

**Table 2. Elements to be included in a summary table characterising the applicability of a body of studies (Atkins et al. 2011)**

| Domain | Description of applicability of evidence |
|---|---|
| Population | *[Describe general characteristics of enrolled populations, how this might differ from target population, and effects on baseline risk for benefits or harms. Where possible, describe the proportion with characteristics potentially affecting applicability (e.g. % over age 65) rather than the range or average.]* |
| Intervention | *[Describe general characteristics and range of interventions and how they compare to those in routine use and how this might affect benefits or harms from the intervention.]* |
| Comparators | *[Describe comparators used. Describe whether they reflect best alternative treatment and how this may influence treatment effect size.]* |
| Outcomes | *[Describe what outcomes are most frequently reported and over what time period. Describe whether the measured outcomes and timing reflect the most important clinical benefits and harms.]* |
| Setting | *[Describe geographic and clinical setting of studies. Describe whether or not they reflect the settings in which the intervention will be typically used and how this may influence the assessment of intervention effect.]* |

It is likely that not all data are available to complete the table. Missing data should be described as well. In addition, the section on outcomes should include a comment regarding which effect measure is less likely to be subject to effect modification.

Atkins et al. 2011 have also summarised examples of characteristics of studies that may affect the applicability (see Annexe 1).

It should be noted that the *size* of the effect modification can only be addressed by statistical methods.

# 3. Discussion and conclusion

There is no single definition of applicability that is widely used/accepted and the terms external validity/generalisability/applicability/transposability are used interchangeably. The aim of this guideline is to provide guidance on how to assess whether there is a relevant effect modification in the population of interest. Therefore we choose to use the term applicability, which we define as the extent to which the effects observed in clinical studies are likely to reflect the expected results when a specific intervention is applied to the population of interest.

To assess the relative effectiveness of interventions, trials with a pragmatic approach which have more ‘noise of practice’, are more suitable than trials with an explanatory approach that are conducted within a strict trial setting. However, there is a balance between making eligibility criteria pragmatic and broad, which relies heavily on the clinical judgement of investigators, and making them very detailed to avoid any ambiguity as internal validity is a prerequisite for the applicability. It should be noted that there is no such thing as a pragmatic trial or an explanatory trial, rather every trial can be positioned somewhere between the extremes and has its pragmatic and explanatory elements.

In practice, especially at the time of a rapid assessment, trials with a pragmatic approach may not be published. If such data are not published it is even more important to consider the applicability of the data that are available. A useful method to test the applicability is through statistical modelling with for example meta-analysis. However, also for this type of evidence applies that these data are not that commonly available, especially in case of a rapid assessment. In addition, time and resources to do such analysis as part of the assessment may be scarce.

Regardless of whether trial with a pragmatic approach or meta-analysis that address the applicability are available, or if they address only specific aspects of the applicability problem, the assessor of a relative effectiveness assessment should always indicate whether it likely that the available evidence is applicable to the questions at hand. In practice, this is a relevant aspect of a relative effectiveness assessment, especially at the time of a Rapid assessment, as relatively few clinical trials are designed with applicability in mind and clinical studies typically report only a few of the factors needed to fully assess applicability.

All countries consider the applicability of the trials at least sometimes for their assessment; however, currently the applicability is not addressed systematically. This is partly, because in contrast to the accumulating body of empiric data on factors affecting the risk of bias, or internal validity, there has been much less empiric data to determine which factors affect applicability. Several authors have made suggestions of lists with criteria; however these lists are intended for checking good clinical practice and are not developed from the viewpoint of the decision maker. None of these lists is widely used and in addition they have not been tested for usage by health technology assessment doers. A usable checklist should be comprehensive but also feasible to apply on multiple trials within the limited timeframe of a health technology assessment.

Because of the unmet need of any detailed guidance for assessing applicability, Atkins et al. (2011) recently published more detailed guidance. The guidance was specifically developed for producing systematic reviews. In this guideline we adopt many of their recommendations and put them in the context of a relative effectiveness assessment. They use the Patient, Intervention, Comparator, Outcome, Setting framework to present factors that (may) affect applicability. In addition, it is stated that because the applicability depends on the specific questions and needs of users it is not possible to devise a valid uniform scale for rating the overall applicability. The concept is to summarise factors that (may) affect applicability in a summary table. By doing this systematically, assessors and readers of the assessment are stimulated to consider the applicability as an important element in the relative effectiveness assessment. This also includes the awareness of data that are not available but should be in order to consider the applicability. One should however not forget that this type of exercise cannot determine the *effect size* of the

effect modification. This can only be addressed by statistical methods. In addition, the 4-step process proposed by the authors may not be feasible in the limited time frame of a Rapid relative effectiveness assessment. If this is the case, at least the summary table of the applicability of the evidence based on the PICOS framework should be included in the assessment report. The summary report should mainly focus on the relative comparison between the intervention and comparator. For example, males may have a score that is 10 units higher than females, but if this is true for both interventions, then the comparative difference between intervention and comparator is the same for males and females.

It should be noted that evaluating the applicability of the evidence cannot be based on a pre-defined formula. Depending on the topic interpretation of the applicability may vary. For example, for a rare disease other considerations and requirements may be relevant compared to a non-rare disease. Regardless of the topic it is very relevant that the considerations are transparently reported.

Finally, it may very well be that specific elements depend on the local context (for example the standards of care may differ per country or region). Hence, one should always consider and address whether a specific element that is relevant for the applicability can be assessed in general or whether this should be done in the local (national) context.

# Bibliography

**References used from search:**

Atkins D, Chang S, Gartlehner G, Buckley DI, Whitlock EP, et al. Assessing the Applicability of Studies When Comparing Medical Interventions. Agency for Healthcare Research and Quality; 2008-2010 Dec 30. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC019-EF. Available at http://effectivehealthcare.ahrq.gov/.

Atkins D, Chang S, Gartlehner G, Buckley DI, Whitlock EP, Berliner E, Matchar D. Assessing applicability when comparing medical interventions: Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol. 2011 Apr 2.

Dekkers OM, von Elm E, Algra A, Romijn JA, Vandenbroucke JP. Int J Epidemiol. 2010 Feb;39(1):89-94. Epub 2009 Apr 17. How to assess the external validity of therapeutic trials: a conceptual approach.

Flather M, Delahunty N, Collinson J. Generalizing results of randomized trials to clinical practice: reliability and cautions. Clin Trials 2006; 3: 508-12.

Glasgow RE, Green LW, Klesges LM, Abrams DB, Fisher EB, Goldstein MG, Hayman LL, Ockene JK, Orleans CT. External validity: we need to do more. Ann Behav Med. 2006 Apr;31(2):105-8.

Green LW, Glasgow RE. Eval Health Prof. 2006 Mar;29(1):126-53. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology.

Julian DG, Pocock SJ. Interpreting a trial report.) in het volgende boek: In: Pitt B, Julian D, Pocock S, editors. Clinical trials in cardiology. London: Saunders; 1997.

Revicki DA and Frank L. Pharmacoeconomic evaluation in the real world. Effectiveness versus efficacy studies. Pharmacoeconomics 1999; 15: 423-34

Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: a conceptual approach. Int J Epidemiol. 2010 Feb;39(1):89-94.

Rothwell PM. Commentary: External validity of results of randomized trials: disentangling a complex concept. Int J Epidemiol. 2010 Feb;39(1):94-6. Epub 2009 Sep 23. No abstract available.

Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005 Jan 1-7;365(9453):82-93.

Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials. 2006 May;1(1):e9.

Seale JP, Gebski VJ, Keech AC. Generalising the results of trials to clinical practice. Med J Aust 2004; 181: 558-60.

Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Obstet Gynecol. 2010 May; 115(5):1063-70.

Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967, 20:637-648.

Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, Tunis S, Bergel E, Harvey I, Magid DJ, Chalkidou K. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ. 2009 May 12;180(10):E47-57

Tonkin AM. Issues in extrapolating from clinical trials to clinical practice and outcomes. Aust N Z J Med 1998; 28: 574-8.

Treweek S, Zwarenstein M. Trials. Making trials matter: pragmatic and explanatory trials and the problem of applicability. 2009 Jun 3;10:37

Walach H, Falkenberg T, Fonnebo V, et al. Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol 2006; 6: 29.

Windeler J. It is "the noise of practice". J Clin Epidemiol. 2010 Jun;63(6):694.

Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D; CONSORT group; Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008 Nov 11;337:a2390.

**Guidelines:**

General Methods Version 3.0. Cologne: Institute for Quality and Efficiency in Health Care (IQWIG); 2008.

Guidelines for Funding Applications to PHARMAC. Wellington: Pharmaceutical Management Agency (PHARMAC); 2010.

Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee. Version 4.3. Canberra: Pharmaceutical Benefits Advisory Committee (PBAC); 2008.

Guidelines for the Economics of Health Technologies in Ireland. Dublin: Health Information and Quality Authority (HIQA); 2010

Guidelines for the Economic Evaluation of Health Technologies: Canada. 3rd ed. Ottawa: Canadian Agency for Drugs and Technologies in Health (CADTH); 2006.

Guide to the Methods of Technology Appraisal. London: National Institute for Health and Clinical Excellence (NICE); 2008.

Health Technology Assessment Handbook: Copenhagen: Kristensen FB & Sigmund H, Danish Centre for Health Technology Assessment (DACEHTA ); 2007

Methodological guidelines for conducting economic evaluation of healthcare interventions. a Hungarian proposal for methodology standards: Szende, Mogyorósy, Muszbek, Nagy, Pallos, Dózsa; 2002.

Working guidelines for the pharmaceutical reimbursement review. The Dental and Pharmaceutical Benefits Agency; 2008

**Additional references used:**

European Medicines Agency. ICH Topic E 8 General Considerations for Clinical Trials. London, 1998. Available at URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002877.pdf (accessed May 2011)

Kleijnen S, Goettsch W, d’Andon A, Vitre P, George E et al. EUnetHTA JA WP5: Relative Effectiveness Assessment (REA) of Pharmaceuticals. Background review. Diemen, 2011.

Pharmaceutical Forum. Core principles on relative effectiveness. Available at URL: http://ec.europa.eu/pharmaforum/docs/rea_principles_en.pdf (accessed December 2010)

**References for effect model:**

Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. Stat Med. 2000; 19(24):3497-518

Boissel JP, Collet JP, Lievre M, Girard P. An effect model for the assessment of drug benefit: example of antiarrhythmic drugs in postmyocardial infarction patients. J Cardiovasc Pharmacol. 1993; 22(3):356-63

Boissel JP, Cucherat M, Nony P, Chabaud S, Gueyffier F, Wright JM, Lievre M, Leizorovicz A. . New insights on the relation between untreated and treated outcomes for a given therapy effect model is not necessarily linear. J Clin Epidemiol. 2008; 61(3):301-7

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000; 19(13):1707-28

Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. Stat Med. 2000; 19(23):3251-74

Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med. 1997; 16(23):2741-58

Wang H, Boissel JP, Nony P. Revisiting the relationship between baseline risk and risk under treatment. Emerg Themes Epidemiol. 2009; 6:1

**References for meta-analysis:**

Wesberg HI, Hayden VC, Pontes VP. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? Clin Trials. 2009 Apr;6(2):109-18.

Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Threats to applicability of randomised trials: exclusions and selective participation. J Health Serv Res Policy. 1999 Apr;4(2):112-21.

Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. Psychol Methods. 2001 Dec;6(4):413-29.

Lipsey MW, Wilson DB. The way in which intervention studies have "personality" and why it is important to meta-analysis. Eval Health Prof. 2001 Sep;24(3):236-54.

Cochrane guidelines

PRISMA recommendations

# Annexe 1. Characteristics of individual studies that may affect external validity (Atkins et al. 2011)

| | Condition that may limit applicability | Example | Feature that should be abstracted into evidence tables |
|---|---|---|---|
| **Population** | Narrow eligibility criteria and exclusion of those with comorbidities | In the FIT trial, the trial randomized only 4000 of 54,000 originally screened. Participants were healthier, younger, thinner, and more adherent than typical women with osteoporosis. | Eligibility criteria and proportion of screened patients enrolled; presence of comorbidities |
| | Large differences between demographics of study population and community patients | Cardiovascular clinical trials used to inform Medicare coverage enrolled patients who were significantly younger (60.1 vs. 74.7 years) and more likely to be male (75% vs. 42%) than Medicare patients with cardiovascular disease.[2] | Demographic characteristics: age, sex, race and ethnicity |
| | Narrow or unrepresentative severity, stage of illness, or comorbidities | Two-thirds of patients treated for congestive heart failure (CHF) would have been ineligible for major trials. Community patients had less severe CHF, more comorbidities and were more likely to have had a recent cardiac event or procedure.[2] | **Severity or stage of illness; comorbidities; referral or primary care population; volunteers vs. population-based recruitment strategies**. |
| | Run in period with high-exclusion rate for nonadherence or side effects | Trial of etanercept for juvenile arthritis used an active run in phase and excluded children who had side-effects, resulting in study with low rate of side-effects.[3] | Run in period; include attrition before randomization and reasons (nonadherence, side-effects, nonresponse).[2, 4] |
| | Event rates much higher or lower than observed in population-based studies | In the Women's Health Initiative trial of post-menopausal hormone therapy, the relatively healthy volunteer participants had a lower rate of heart disease (by up to 50%) than expected for a similar population in the community.[5] | Event rates in treatment and control groups |
| **Intervention** | Doses or schedules not reflected in current practice | Duloxetine is usually prescribed at 40-60mg/d. Most published trials, however, used up to 120 mg/d.[6] | Dose, schedule, and duration of medication |
| | Intensity and delivery of behavioral interventions that may not be feasible for | Studies of behavioral interventions to promote healthy diet employed high number and longer duration of visits | Hours, frequency, delivery mechanisms (group vs. individual) and duration. |

[2] Dhruva SS, Redberg RF. Variations between clinical trial participants and Medicare beneficiaries in evidence used for Medicare National Coverage Decisions. Arch Intern Med 2008 Jan; 169(2):136-140

[3] Cummings SR, Black DM, Thompson DE, et al. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from the fracture intervention trial. JAMA 1998;280(24):2077-2082

[4] Bravata DM, McDonald KM, Gienger AL, et al. Comparative Effectiveness of Percutaneous Coronary Interventions and Coronary Artery Bypass Grafting for Coronary Artery Disease. Comparative Effectiveness Review No. 9. (Prepared by Stanford-UCSF Evidence-based Practice Center under Contract No. 290-02-0017.) Rockville, MD: Agency for Healthcare Research and Quality; October 2007.

[5] Anderson GL, Limacher M, Assaf AR, et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. JAMA 2004 Apr 14;291(14):1701-1712.

[6] Gartlehner G, Hansen RA, Thieda P, et al. Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression. Comparative Effectiveness Review No. 7. (Prepared by RTI International-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality; January 2007.

| | Condition that may limit applicability | Example | Feature that should be abstracted into evidence tables |
|---|---|---|---|
| | routine use | than is available to most community patients.[7] | |
| | Monitoring practices or visit frequency not used in typical practice | Efficacy studies with strict pill counts and monitoring for antiretroviral treatment does not always translate to effectiveness in real world practice.[8] | Interventions to promote adherence (e.g., monitoring, frequent contact). Incentives given to study participants. |
| | Older versions of an intervention no longer in common use | Only one of 23 trials comparing coronary artery bypass surgery with percutaneous coronary angioplasty used the type of drug eluting stent that is currently used in practice.[4] | Specific product and features for rapidly changing technology |
| | Co-interventions that are likely to modify effectiveness of therapy | Supplementing zinc with iron reduces the effectiveness of iron alone on hemoglobin outcomes.[9] Recommendations for iron are based on studies examining iron alone, but patients most often take vitamins in a multivitamin form. | Co-interventions |
| | Highly selected intervention team or level of training/proficiency not widely available | Trials of carotid endarterectomy selected surgeons based on operative experience and low complication rates and are not representative of community experience of vascular surgeons.[10] | Selection process, training and skill of intervention team. |
| **Comparator** | Inadequate dose of comparison therapy | A fixed dose study by the makers of duloxetine compared 80 and 120 mg/d of duloxetine (high dose) with 20 mg of paroxetine (low dose).[11] | Dose and schedule of comparator, if applicable |
| | Use of substandard alternative therapy | In early trials of magnesium in acute myocardial infarction, standard of treatment did not include many current practices including thrombolysis and beta-blockade.[12] | Relative comparability to the treatment option. |
| **Outcomes** | Composite outcomes that mix outcomes of different significance | Cardiovascular trials frequently use composite outcomes that mix outcomes of varying importance to patients.[13] | Effects of intervention on most important benefits and harms, and how they are defined |
| | Short-term or surrogate outcomes | Trials of biologics for rheumatoid arthritis used radiographic progression rather than symptoms.[14] | How outcome defined and at what time |

[7] Whitlock EP, O'Connor EA, Williams SB, et al. Effectiveness of Weight Management Programs in Children and Adolescents. Evidence Report/Technology Assessment No. 170 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). AHRQ Publication No. 08-E014. Rockville, MD: Agency for Healthcare Research and Quality; September 2008.

[8] Fletcher CV. Translating efficacy into effectiveness in antiretroviral therapy: beyond the pill count. Drugs 2007;67(14):1969-1979.

[9] Walker, CF, Kordas K, Stoltzfus, RJ, et al. Interactive effects of iron and zinc on biochemical and functional outcomes in supplementation trials. Am J Clin Nutr 2005 82: 5-12.

[10] Wennberg D, Lucas F, Birkmeyer J, et al. Variation in carotid endarterectomy mortality in the Medicare population. JAMA 1998;279:1278-1281.

[11] Detke MJ, Wiltse CG, Mallinckrodt CH, et al. Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. Eur Neuropsychopharmacol 2004 Dec;14(6):457-470.

[12] Li J, Zhang Q, ZhangM, et al. Intravenous magnesium for acute myocardial infarction. Cochrane Database of Systematic Reviews 2007, Issue 2. Art. No.: CD002755. DOI: 10.1002/14651858.CD002755.pub2.

[13] Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007;334;786; originally published online 2 Apr 2007

[14] Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol 1997 Oct;50(10):1089-1098.

| | Condition that may limit applicability | Example | Feature that should be abstracted into evidence tables |
|---|---|---|---|
| | | Trials of Alzheimer's disease drugs primarily looked at changes in scales of cognitive function over 6 months which may not reflect their ability to produce clinically important changes such as institutionalization rates.[15] | |
| **Setting** | Standards of care differ markedly from setting of interest | Studies conducted in China and Russia examined the effectiveness of self breast exams on reducing breast cancer mortality, but these countries do not routinely have concurrent mammogram screening as is available in the United States.[16] | Geographic setting |
| | Specialty population or level of care differs from that seen in community | Early studies of open surgical repair for abdominal aortic aneurysms found an inverse relationship between hospital volume and short-term mortality.[17] | Clinical setting (e.g. referral center vs. community) |

[15] Hansen RA, Gartlehner G, Kaufer D, et al. Drug class review of Alzheimer's drugs. Final report. 2006. Available at: http://www.ohsu.edu/drugeffectiveness/reports/final.cfm.

[16] Humphrey L, Chan BKS, Detlefsen S, et al. Screening for Breast Cancer. Prepared by Oregon Health Sciences University under Contract No. 290-97-0018. Rockville, MD. Agency for Healthcare Research and Quality; August 2002.

[17] Wilt TJ, Lederle FA, MacDonald R, et al. Comparison of Endovascular and Open Surgical Repairs for Abdominal Aortic Aneurysm. Evidence Report/Technology Assessment No. 144. (Prepared by the University of Minnesota Evidence-based Practice Center under Contract No. 290-02-0009.) AHRQ Publication No. 06-E017. Rockville, MD: Agency for Healthcare Research and Quality; August 2006.

# Annexe 2. Statistical methods

As it is unlikely that within the timeframe of an HTA (especially a rapid assessment) assessors will have the opportunity to do these type of analyses/models, this section below will focus on how assessors can interpret analysis/models that are already published. Meta-analysis and estimate of the effect model are not concurrent approach. Rather, whenever it is possible (i.e. a batch of clinical trials is available or predefined, not overlapping sub-groups), a preferred strategy is to perform first a meta-analysis (with at least two types of criteria, e.g. relative risk and rate difference,), and then to estimate the effect model with the (pooled) individual data.

***Effect model***

The expected effect of a treatment is a decrease of incidence of an event caused by the illness, i.e. mortality and/or morbidity. The effect model of a treatment (effect) is the relation between the incidence '*Rc'* of the event in the patients who do not receive the treatment and the incidence '*Rt*' of this event in the same population of treated patients. This relationship may be written: *Rt* = f (*Rc* , $\theta$) where $\theta$ indicates it is treatment dependent.

When RCTs are available on a treatment, this relationship should be explored to help identify good responders to this treatment. When a single trial has been achieved, exploring the effect model requires access to individual data.

Often, this relationship is assumed to be linear, most of the time multiplicative (*Rt* = *Rc* $\times$ $\theta$) – when there is no natural threshold. Treatment effect $\theta$ is estimated in this first case by a relative risk. The effect of a treatment that has both favourable and iatrogenic effects can be modelled with use of a mixed linear model (*Rt* = k *Rc* + b) that combines multiplicative and additive effects (Boissel *et al.* 1993).

Effect model can be graphically illustrated with *Rc* and *Rt* respectively on the x and y axes (see Figure 1 and Figure 2). Trial results are plotted by dots of coordinates (x=*Rc* , y=*Rt*). Dots represent either a trial or a patient. Figures are divided in two areas, one below the *Rt* = *Rc* line corresponding to a beneficial treatment effect and one above the *Rt* = *Rc* line where treatment is harmful. One approach still used in medical literature to assess effect model is to compute the weighted least squares regression line.

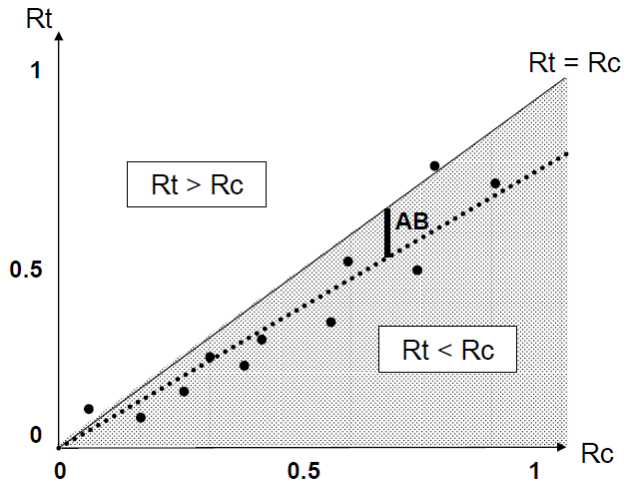



Figure 1 : Linear multiplicative effect model, *Rt* = kRc

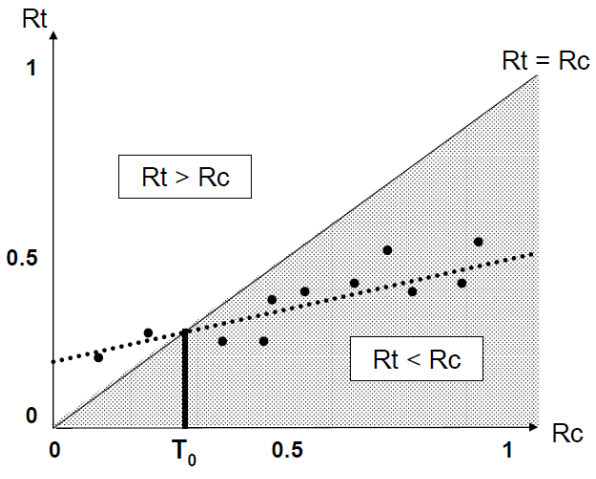



Figure 2: Mixed linear effect model , *R t*= k*Rc* + b

Each dot represents a trial result. The relation between *Rt* and *Rc* is represented by the dotted line. Absolute benefit (AB) increases with *Rc*. Figure 2 shows the deleterious effect of treatment when *Rc* lies below a threshold value $T_0$.

Linear models are simple to use but may not always be plausible given the complexity of the biological mechanisms involved in a treatment effect. It has therefore been proposed to integrate in the model other covariates such as the characteristics of patients other than those included in *Rc*, like phenotype- or genotype-derived variables (Boissel et al*.* 2008). Effect model may thus be written: *Rt* = f (*Rc* , θ , X) with X, a vector of characteristics of patients. Relevant individual patient data at baseline are therefore needed to explore the effect model. When the effect model is estimated from individual data (from a single trials or pooled from all trials), the *Rc* and *Rt* are obtained as prediction with appropriate statistical techniques rather than as frequency when aggregated data from all trials are used.

Boissel et al*.* identified three mechanisms of action of a therapy they believed to cover the whole possible modes of action of a treatment: 1) alteration of the circumstances leading to the disease occurrence, 2) alteration of a causal risk factor and 3) alteration of the intimate mechanism of the disease. For each mechanism of action, a therapeutic effect model was developed and a simulation study performed. A linear relationship between *Rt* and *Rc* was found only for the first mechanism of action. In the two other cases, the predicted effect model was curvilinear suggesting specific interactions between treatment and individuals (see Figure 3).

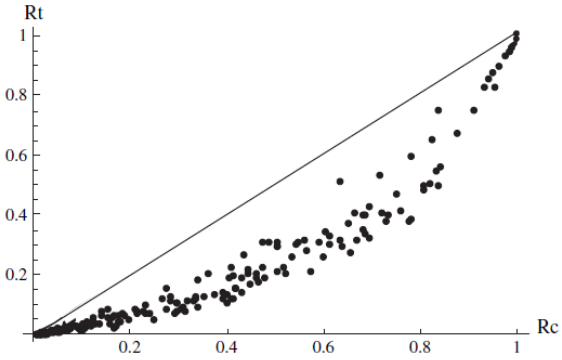



Figure 3. Simulated effect model for a mechanism of action that is alteration of a causal risk (the treatment acts on a causal risk factor). The value of *Rc* is directly affected through the mechanism that causes the risk. Each dot represents a subject. For a given value of *Rc* ordinates of dots differ by the values of *X*. The absolute benefit predictable for a patient is the length between the corresponding dot ordinate and the line 0,0-1,1 [extract from Boissel *et al*. 2008]

The above examples illustrate the importance of exploring the effect model, i.e. the relation between baseline risk and treatment effect, to select patients to be treated. This relationship is often investigated in meta-analyses as it provides a possible explanation of between-study heterogeneity (Sharp et al*.* 2000, Thompson et al*.* 1997). Indeed, one major criticism of these statistical methods is that they combine results from trials with very different patient characteristics and designs (Engels et al. 2000). Thus, exploring sources of heterogeneity is a key issue to assess whether observed differences in treatment effects can be explained by trial-level characteristic. For example patient age or other patient characteristics may influence the baseline risk so that treatment effect may be over- or underestimated when applying the results of a trial to other patients.

Techniques and outcomes are not the same when the effect model estimate is obtained from aggregated or from individual data. In the latter case, pooling data from all the trials is even better than working on a single data set. It allows more precise estimates and more relevant model validation. As a third option, the effect model can be obtained from simulation with a mathematical model of the disease and drug interactions with the body. This allows to taking into account patient descriptors that are potential treatment effect modulators that have not been measured in clinical trials.

Although statistical tools are the same as for multivariate statistical analysis of data, in effect model estimate they are not used with the same approach (adapted analytical strategy).

In conclusion, an effect model is a quantitative method that allows knowing how the treatment benefit varies according to certain patient characteristics, in order to assess treatment effect in a specific population of patients. For a health technology assessor, information on effect model may

be available from meta-analyses, in particular in the section that explores sources of heterogeneity between included trials. Though the approach consisting in computing the weighted least squares regression line of the *Rt-Rc* plot is still used in medical literature, more appropriate approaches that take into account either biological interactions between baseline characteristics of patients and treatment effect (Boissel et al*.* 2008, Wang et al. 2009) or measurement errors of treatment effect and baseline risk estimations (Arends et al*.* 2000) have been developed and should be preferred.

***Meta-analysis***

Meta-analysis allows to examine the relationship between cross-study variability in effects and study characteristics that represent dimensions of potential applicability such as subject characteristics, organizational and geographic setting, research context etc. Two statistical approaches explore the applicability in meta-analysis: the heterogeneity between studies and multivariate modelling to determine relevant features that influence treatment effects. The later requires individual data. Seeking for treatment size modulators is more effective with meta-analysis on individual data, either from a single trial or from pooled trials.

- o Heterogeneity

Although the RCT is regarded as the 'gold standard' in terms of evaluating the efficacy of interventions, it is susceptible to challenges to its external validity if those participating are unrepresentative of the reference population for whom the intervention in question is intended. But in most RCTs, subjects are selected randomly and representativity is not assessed. Design, intervention and setting characteristics are also specifically selected. With several different RCTs on the same intervention, the key question is whether effects vary between RCTs and, if so, how it is related to study characteristics. When variation of treatment effects across studies is plausible, lack of applicability should be a serious concern.

Meta-analysis of multiple studies with good internal validity can thus be characterized as an empirical study of the applicability of intervention. A significant heterogeneity indicates an interaction between treatment effects and characteristics of studies (e.g. population, treatment, design characteristics). The usual way of assessing whether a set of single studies are heterogeneous is by means of the *Q* test. However, the *Q* test only informs about the presence *versus* the absence of heterogeneity, but it does not report on the extent of such heterogeneity. Recently, the *I*2 index has been proposed to quantify the degree of heterogeneity in a meta-analysis. In practice, power of such test is often limited by a small number of studies. In case of heterogeneity, it should be investigated further to reveal its sources which could be related to population, treatment or setting and affects applicability. And with significant heterogeneity, estimation of effect size with random effect is preferable than fixed model. Heterogeneity can also be explored within a trial. The same techniques as above are used. The issue is sub-grouping of patients. “Natural” sub-groupings are centres, countries, scores of disease severity, age groups…

- o Multivariate Modelling

Identification of any features of population or intervention that modify treatment effects could be done by multivariate models allowing prediction of treatment effect for given scenarios of characteristics within each individual study (or by subgroup analysis). In meta-analysis, a study level meta-regression adjusted for potential confounding by study features could be done to identify study characteristics related to effect size. But design differences are often confounded with the substantive variables most relevant for applicability. Thus, to address applicability, construction of multivariate models allowing to predict treatment effect for a given scenario of characteristics of population, intervention, setting, etc. requires a relatively large number of diverse studies providing adequate information on the relevant study features. Alternatively, if meta-analysis based on individual-patient data is generally more powerful in terms of statistical unit, it rarely allows completeness and is often limited by lack of relevant variables and unavailability of individual data from some trials. Thus, this method may reduce the number of diverse studies compare to meta-analysis on summarized data and this diversity is particularly important to address applicability, unless cooperation across academic teams and with companies permits to bring in all the available data.

Both method and practice of meta-analysis should be improved with greater attention to the applicability of study results and the systematic multivariate relationships between study characteristics and the effect sizes reported in those studies. But meta-analysis with RCT does not allow to explore all determinants of the applicability and its power is often limited by a small number of studies.

# Annexe 3. Methods and results of literature search

A literature review has been conducted for the original guideline version in JA 1 with the aim of locating studies that provide recommendations of assessing the applicability of trial results.

## *Keywords*

The following keywords have been included in the search:

- effect model
- extrapolation
- real world
- real life
- effectiveness
- external validity
- generalizability/generalisability
- applicability
- transposability
- technology assessment
- relative effectiveness
- comparative effectiveness
- clinical Trials as Topic

## *Search engines and sources of information*

The following databases and websites have been searched:

Databases:

- Embase
- Medline
- Centre for Reviews and Dissemination, University of York

Websites:

- Agency for Healthcare Research and Quality (AHRQ)
- National Guideline Clearinghouse
- International Society For Pharmacoeconomics and Outcomes Research (ISPOR)
- HTAi
- EMA
- Google and Google Scholar
- ScienceDirect
- The Cochrane Collaboration
- The Cochrane Methodology Register
- Wiley-Interscience

The following guidelines (in English) of health technology assessment/reimbursement agencies have been included in the search[18]:

- Agency for Health Technology Assessment in Poland (AHTAPol)
- Canadian Agency for Drugs and Technologies in Health (CADTH)
- Danish Centre for Health Technology Assessment (DACEHTA)
- Health Information & Quality Authority (HIQA)
- Institute for Quality and Efficiency in Health Care (IQWIG)
- National Institute for Health and Clinical Excellence (NICE)
- Pharmaceutical Benefits Advisory Committee (PBAC)
- The New Zealand Pharmaceutical Management Agency (PHARMAC)
- Dental and Pharmaceutical Benefits Agency (TLV)

[18] The guidelines were identified in the Background review on Relative Effectiveness Assessment of Pharmaceuticals of WP5 (Kleijnen et al. 2011)

In addition we have hand searched references cited in relevant documents.

## *Inclusion and non-inclusion criteria*

The following inclusion criteria were applied to:

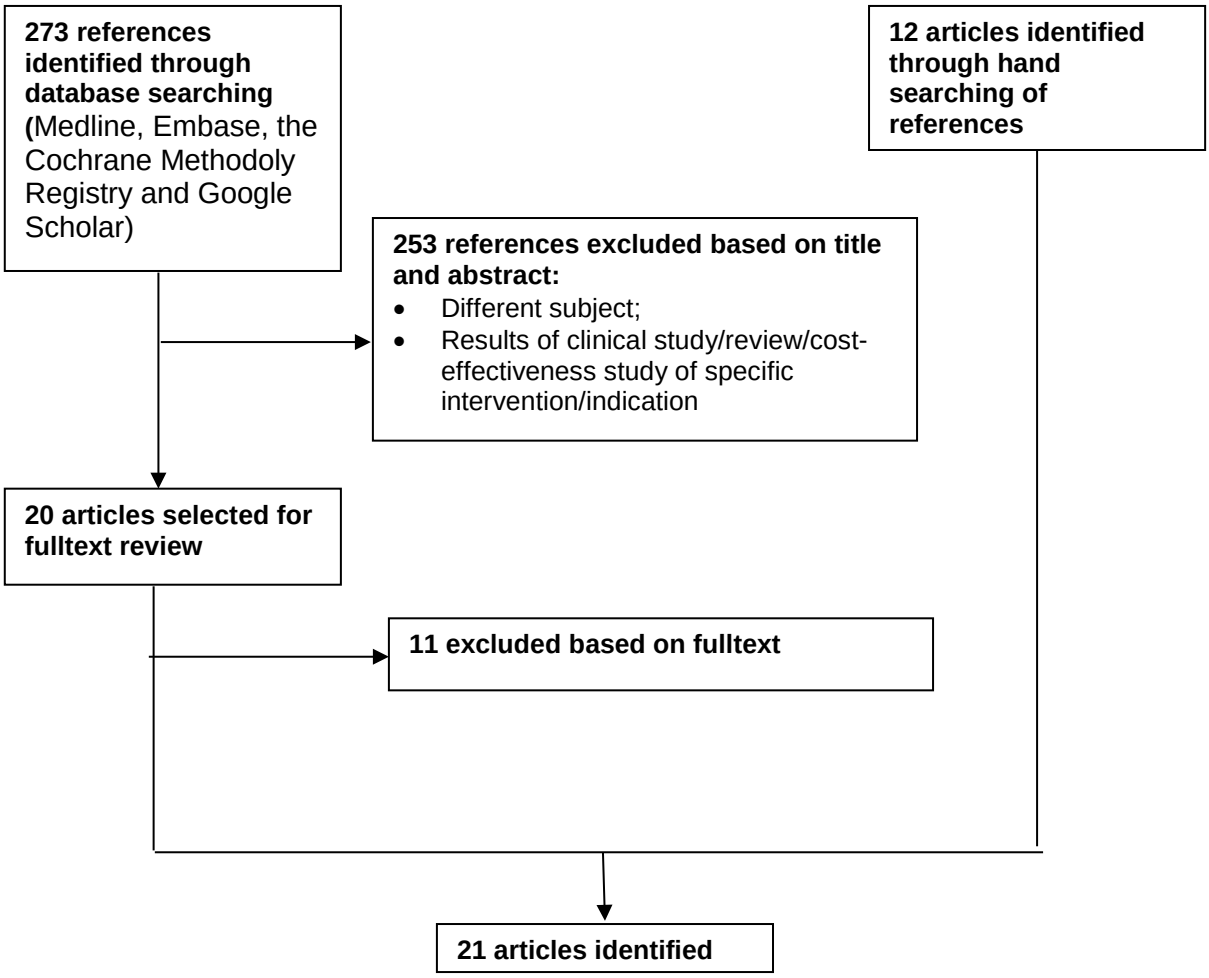
- Where time limits could be specified (e.g. PubMed) the database searches were limited to the period 1995/01/01 to 2011/05/04.
- Publication written in English
- Critical analysis of methods to determine the applicability
- General reflections and theoretical considerations

The following exclusion criteria were applied:

- Letters
- Studies on specific interventions and/or indications

## *Results of search*

The search in Medline, Embase, the Cochrane Methodology Registry and Google Scholar resulted in 273 references. Based on a first selection, based on title and abstract, we retained 20 references of which the full text was obtained. Selection based on full texts reduced the number of relevant papers to 11. In addition, 12 articles we identified through hand searching of references.



# Annexe 4. Overview of lists with criteria to determine the applicability

| Seale et al. 2004 | Flather et al. 2006 | Rothwell et al. 2006 | Dekkers et al. 2009 | Julian et al. 1997 | Green et al. 2006 |
|---|---|---|---|---|---|
| **Is the patient population representative of the broad target group?** (inclusion and exclusion criteria and baseline data)<br><br>**Participant flow diagram** (analysis of patients that were eligible, but not included)<br><br>**Screening logs** (analysis of criteria for excluding patients)<br><br>**Comorbidities** (comparison of comorbities in trials groups vs target patient population)<br><br>**Subgroup analysis** (in large clinical trials it is possible to have reliable subgroup analyses which may help prescribers to relate the trial's findings more closely to patients for whom they are trying to select appropriate therapies) | **Patient selection** (Differences relating to baseline characteristics such as age, gender, and severity of disease are likely to occur)<br><br>**Study design and validity of results** (low compliance rate, unacceptable rate of serious unwanted side effects, flaws in the randomisation process, 'Per protocol' or 'on treatment' analyses and fraud or scientific misconduct)<br><br>**Application of the treatment** (competence and experience of clinicians as well as the health care setting should be taken into account) | **Setting of the trial** (health-care system, country, recruitment from primary, secondary, or tertiary care, selection of participating centres, selection of participating clinicians)<br><br>**Selection of patients** (method of prerandomisation diagnosis and investigation, eligibility criteria, exclusion criteria, placebo run-in period, treatment run-in period, 'enrichment' strategies, ratio of randomised patients to eligible nonrandomised patients in participating centres, proportion of patients who declined randomisation);<br><br>**Characteristics of randomised patients** (baseline clinical characteristics, racial group, uniformity of underlying pathology, stage in the natural history of their disease, severity of disease, comorbidity, absolute risks of a poor outcome in the control group)<br><br>**Difference between trial protocol and routine practice** (trial intervention, timing of treatment, appropriateness/relevance of control intervention, adequacy of nontrial treatment –both intended and actual, prohibition of certain nontrial treatments, therapeutic or diagnostic advances since was performed)<br><br>**Outcome measures and follow-up** (clinical relevance of surrogate outcomes, clinical relevance validity and reproducibility of | **Are the eligibility criteria a proper reflection of the study population?** (selection of study population, run-in period, participating centres)<br><br>**Do temporal, ethnical and geographical differences between study population and target populations translate in to a limited generalisability ?** (temporal aspects, ethnical aspects, geographical and socio-economic aspects)<br><br>**Can study results be generalized beyond the eligibility criteria?** (age, co-morbidities)<br><br>**Do differences in treatment setting translate into possible differences in treatment effects?** (treatment physicians, treatment setting, administrative policy) | **Patients studied**<br>Where the patients included in the trial adequately representative of the patients to be encountered in normal clinical practice?<br>Where the eligibility criteria too narrow or too broad?<br>*In/exclusion criteria should clearly be stated. Are women, elderly, comorbitities, risks well represented?*<br><br>Where adequate steps taken to ensure a high proportion of eligible patients was randomised?<br>*How do randomised patients with those who are not randomised (eligible vs noneligible?)*<br><br>Was the setting of the trial and the manner of patient selection appropriate?<br>Were the eligibility criteria too narrow or too broad? Have the authors inappropriately extrapolated their findings | **Reach and representativeness**<br>A. Participation: Are there analyses of the participation rate among potential (a) settings, (b) delivery staff, and (c) patients (consumers)?<br>B. Target audience: Is the intended target audience stated for adoption (at the intended settings such as worksites, medical offices, etc.) and application (at the individual level)?<br>C. Representativeness—Settings: Are comparisons made of the similarity of settings in study to the intended target audience of program settings—or to those settings that decline to participate?<br>D. Representativeness—Individuals: Are analyses conducted of the similarity and differences between patients, consumers, or other subjects who participate versus either those who decline, or the intended target audience?<br><br>**Program or policy implementation and adaptation**<br>A. Consistent implementation: Are data presented on level and quality of implementation of different program components?<br>B. Staff expertise: Are data presented on the level of |

| Seale et al. 2004 | Flather et al. 2006 | Rothwell et al. 2006 | Dekkers et al. 2009 | Julian et al. 1997 | Green et al. 2006 |
|---|---|---|---|---|---|
| | | complex scales, effects of intervention on most treatment components, Who measured outcome, Use of patient-centred outcomes, Frequency of follow-up, Adequacy of the length of follow-up)<br><br>**Adverse effects of treatment**<br>(Completeness of reporting of relevant adverse effects, Rates of discontinuation of treatment, Selection of trial centres and/or clinicians on the basis of skill or experience, Exclusion of patients at risk of complications, Exclusion of patients who experienced adverse effects during a run-in period, Intensity of trial safety procedures) | | to types of patients that were not adequately presented?<br><br>**Treatments**<br>Were the treatments under comparison, including dose schedule, duration of treatment, noncomplianc e, and the control group regimens (placebo or standard treatment) appropriate for normal clinical practice and determining future treatment policy in such patients.<br>*e.g. if treatment duration in the study was relatively short then extrapolation to longer duration (e.g chronic conditions) may not be justified. If details are lacking on actual drug use (departures from scheduled dose) then applicability to future patients will be unclear.*<br><br>Were all aspects of current good clinical practice adequately taken into account?<br>*Ancillary treatment should be clearly* | training or experience required to deliver the program or quality of implementation by different types of staff?<br>C. Program adaptation: Is information reported on the extent to which different settings modified or adapted the program to fit their setting?<br>D. Mechanisms: Are data reported on the process(es) or mediating variables through which the program or policy achieved its effects?<br><br>**Outcomes for decision making**<br>A. Significance: Are outcomes reported in a way that can be compared to either clinical guidelines or public health goals?<br>B. Adverse consequences: Do the outcomes reported include quality of life or potential negative outcomes?<br>C. Moderators: Are there any analyses of moderator effects—including of different subgroups of participants and types of intervention staff—to assess robustness versus specificity of effects?<br>D. Sensitivity: Are there any sensitivity analyses to assess dose-response effects, threshold level, or point of diminishing returns on the resources expended?<br>E. Costs: Are data on the costs presented? If so, are standard economic or accounting methods used to fully account for costs? |

| Seale et al. 2004 | Flather et al. 2006 | Rothwell et al. 2006 | Dekkers et al. 2009 | Julian et al. 1997 | Green et al. 2006 |
|---|---|---|---|---|---|
| | | | | *specified (in in/exclusion criteria but also the actual number of patients receiving it). The aim is to make clear the role of a new treatment in the context of other existing treatments.*<br><br>**Outcome measures and follow-up**<br>Were the outcome measure (endpoints, indicators of patient response) appropriate for reaching overall conclusions about the treatment(s) under investigation. Consistency of measured outcomes with conclusions drawn.<br>Was too much evidence given to surrogate markers of response (e.g. physical indicators) rather than the more major indicators of overall prognosis (e.g. mortality, major clinical events)?<br>Appropriate balance of surrogate and clinical outcomes<br><br>Was the treatment duration and length of | **Maintenance and institutionalization**<br>A. Long-term effects: Are data reported on longer term effects, at least 12 months following treatment?<br>B. Institutionalization: Are data reported on the sustainability (or reinvention or evolution) of program implementation at least 12 months after the formal evaluation?<br>C. Attrition: Are data on attrition by condition reported, and are analyses conducted of the representativeness of those who drop out? |

| Seale et al. 2004 | Flather et al. 2006 | Rothwell et al. 2006 | Dekkers et al. 2009 | Julian et al. 1997 | Green et al. 2006 |
|---|---|---|---|---|---|
| | | | | patient follow-up sufficiently reliable to assess the efficacy and safety of treatment?<br><br>Where adequate steps taken to elicit all relevant adverse events and side-effects of treatment?<br>Coverage of all relevant outcomes (adverse events and side effects).<br><br>**Conclusion**<br>Consideration of the study findings in the context of other available evidence. | |

# Annexe 5. Questions developed by PHARMAC to address the applicability of evidence

PHARMAC has phrased in its’ guideline the following questions to assess the applicability of the clinical trial data (PHARMAC, 2010):

- Patient population: Was the patient population in the trial similar to those considered for funding?
- Comparator: Was the comparator consistent with current clinical practice in New Zealand?

Dose, formulation and administration regimen: Were these consistent with recommended treatment regimes in New Zealand?