eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

GUIDELINE

Endpoints used for Relative Effectiveness Assessment:

Clinical Endpoints

Adapted version (2015)

based on

"Endpoints used for Relative Effectiveness Assessment of pharmaceuticals: Clinical Endpoints" - 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 (REA) of pharmaceuticals.

The guideline “Endpoints used in REA of pharmaceuticals: clinical endpoints” has been elaborated during JA1 by experts from HIQA, 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

REA – relative effectiveness assessment
HRQoL – Health-Related Quality of Life
HIV - human immunodeficiency virus
HbA1C – haemoglobin A1C also known as glycated haemoglobin, glycohaemoglobin,
PRO – patient reported outcomes
EEG - electroencephalograph
FEV1 – forced expiratory volume in one second
PFS – progression-free survival
OS – overall survival
EPACs - Endpoint adjudication committees

Summary and Recommendations

Summary

This guideline provides a set of recommendations for the selection and assessment of clinical endpoints when completing a Relative Effectiveness Assessment (REA). Clinical endpoints are regarded as a means to measure the impact of a treatment on how a patient feels, functions and survives. That impact is usually in the form of improved health status (e.g. survival, cure, remission), but it may also be worsened health status (e.g. adverse reactions, hospitalisations, death). Clinical endpoints can be broadly categorised into three domains: mortality, morbidity and Health Related Quality of Life (HRQoL) measures. The endpoints reported for an assessment should be clearly relevant to the disease, condition or process of interest. Clinical endpoints should be: a main symptom or sign of a disease; a valid measure of clinical benefit due to treatment; clinically relevant; sensitive (responsive to change); and recognised/used by physicians. Clinical endpoints should be reproducible and valid. A reproducible endpoint facilitates comparisons across studies and jurisdictions. A valid endpoint measures what was intended to be measured. Validity may be hampered by selection bias, information bias and residual confounding. Endpoints should be shown to be fit for purpose in the context that they are being used. Issues regarding the precision of study results may be reflected in the statistical significance of the treatment effect. A clinically relevant effect meets some standard or consensus about the magnitude and quality of the study result that is considered meaningful by independent clinicians and/or patients. The clear definition, reproducibility, validity, and the statistical and clinical relevance of an endpoint should all be made evident.

In any REA a hierarchy of endpoints should be established (e.g. primary endpoints, secondary endpoints) even if all endpoints will be simultaneously assessed. The choice of clinical (primary) endpoint will depend upon the target population, main characteristics of the disease of interest (non life-threatening versus life-threatening) and the aim of treatment. For a life-threatening disease, a mortality or survival endpoint is generally preferred as the primary endpoint, whereas morbidity and/or HRQoL may be preferred as secondary endpoints. In non life-threatening diseases, morbidity and HRQoL endpoints will be preferred for the primary endpoints. The clinical endpoints used should be measurable for all or most patients within a reasonable time frame.

Clinical endpoints may be reported by a patient (Patient-Reported Outcomes, PRO), clinician, caregiver or an observer (e.g. paediatrics). Changes in HRQoL may be linked to factors other than the treatment effect and as such they can be susceptible to bias, unless the HRQoL instrument used has been specifically developed to capture the direct impact of a given pathology or its treatment.

Clinical endpoints can be intermediate or final (see Section VI and EUnetHTA guideline on Surrogate endpoints). Validated intermediate endpoints may be used when it is not feasible to measure long term or final endpoints. Caution must be exercised in extrapolating from intermediate to final outcomes. Final endpoints are often defined by survival and, whenever possible, should be used in preference to intermediate outcomes. If final outcomes are unavailable, intermediate outcomes may be acceptable where there is compelling evidence of a clear and consistent correlation with the final outcome of interest. Composite endpoints combine multiple single endpoints into one endpoint showing the overall treatment effect (see EUnetHTA guideline on Composite endpoints). Caution must be exercised when interpreting the composite endpoints as all of the component endpoints should meet the criteria of validity, reproducibility and clinical relevance.

Outcome data may be continuous, binary, ordinal, categorical or count. Where an outcome in continuous form is converted to a categorical or binary outcome, care must be taken to use either unbiased cut-points or widely accepted cut-points that were chosen *a priori*.

Outcomes can be summarised and presented in absolute or relative terms. Absolute measures are useful to clinicians as they provide a quantification of treatment effect that is meaningful for treatment evaluation and prognosis. However, due to the dependence of absolute measures on baseline risk, relative measures are more generalisable across studies. The manner in which clinical outcomes are presented leaves significant scope for misleading conclusions to be supported. Every attempt should be made to provide both absolute and relative measures in tandem. Data from survival analysis are common and should ideally contain overall survival. Censoring is an issue as is failure to follow-up patients after the first non-fatal event.

Recommendations

1. All clinical endpoints should be comprehensively defined and justified in the study protocol(s) and report. They should be clinically relevant to the disease being treated.
2. Endpoint estimates should be presented to show both statistical significance and clinical relevance.
3. Where appropriate, endpoints should be expressed in natural units (e.g. post-operative infections prevented).
4. The implications of the observed treatment effect on clinical endpoint should be easy to interpret.
5. Clinical endpoints should be sensitive to treatment differences.
6. Measurement of clinical endpoints:
    a. Clinical endpoints should be measurable within a reasonable period of time for all or a high proportion of patients.
    b. Both relative and absolute measures should be presented. Responder analysis may be presented when appropriate.
    c. A clinical endpoint should be measured with minimal measurement or assessment error.
7. Where a continuous or ordinal endpoint is converted to dichotomous, there should be a clear justification for the choice of cut-point.
8. Clinical endpoint estimates should come from unbiased studies, especially with respect to detection bias (e.g. appropriate blinding).
9. An endpoint should be independent of jurisdiction or region to maximise comparability.
10. The analysis of endpoint data should explicitly state the handling of missing data.
11. Clinical endpoints should be long-term or final endpoints where possible, although short-term endpoints are acceptable for acute conditions with no long-term consequences. All-cause mortality should be used where relevant as it is the most unbiased endpoint. Overall survival is the preferred clinical endpoint in a survival analysis.
12. Any extrapolation from intermediate to final endpoints should be underpinned by a clear biological or medical rational or a strong or validated link.
13. Multiple endpoints can be presented, including adverse event endpoints. It might be helpful to determine a hierarchy of endpoints.
14. Appropriate adjustment should be considered for multiple hypothesis testing.
15. Composite endpoints should be presented in disaggregated form, be based on endpoints of clinical importance for REA and ideally show a homogenous response across all components.

1. Introduction

1.1. Definitions and general information

A **clinical endpoint** is an aspect of a patient’s clinical or health status that is measured to assess the benefit or harm of a treatment. A clinical endpoint describes a valid measure of clinical benefit due to treatment: the impact of treatment on how a patient feels, functions and survives.(1) It is clinically relevant, sensitive (responsive to change) and is both accepted and used by physicians and patients. Clinical endpoints may be a clinical event (e.g. mortality,) a composite of several events, a measure of clinical status, or health related quality of life (HRQoL). In the literature, the terms 'endpoint' and 'outcome’ are generally used interchangeably. In this document the term ‘endpoint’ will be used. In addition, it is proposed to use the term “clinical endpoint” instead of the term “patient-relevant endpoint” (defined in most cases as mortality, morbidity and/or HRQoL).

Patient reported outcome (PRO): the term PRO covers a whole range of measurement types, encompassing simple symptom measures (such as pain measured by Likert scale), more complex measures (such as activities of daily living or function), multidimensional measures (such as health-related quality of life) and satisfaction with treatment.

A **surrogate endpoint** is an endpoint that is intended to replace clinical endpoint of interest that cannot be observed in a trial - it is a variable that provides an indirect measurement of an effect in situations where direct measurement of clinical effect is not feasible in a reasonable timeframe. A surrogate endpoint is expected to predict the effect of therapy (either benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence. In many cases, an effect on a surrogate endpoint will not per se be of any benefit to the patient (biomarkers are typical examples). A surrogate endpoint may be a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint may also be a clinical endpoint that is used to replace the endpoint of interest, such as an intermediate clinical endpoint.

A **composite endpoint** combines two or more of single events (e.g. mortality, non-fatal myocardial infarction, stroke, hospitalisation and revascularisation procedures) in one endpoint showing the overall and clinically relevant treatment effect. Patients who have experienced any of the components of a composite endpoint are considered to have experienced the composite endpoint. Composite endpoint usually refers to combined morbidity and mortality endpoints; it may also be a combination of patient-reported, observer reported or clinician reported measures. Composite endpoints are often used where statistical power is poor to increase event rates and decrease sample size and to avoid the issue of multiple testing.

1.2. Context

1.2.1. Problem statement

“Which clinical endpoints are accepted for the assessment? How are (absolute, incremental, relative) differences between treatments assessed; what is the role of absolute differences (e.g. 2 months survival gain) and relative differences (e.g. hazard ratio's)?”

1.2.2. Discussion (on the problem statement)

Clinical endpoints are measured in clinical trials or other types of studies assessing the effect of a treatment. For any given disease there may be a variety of possible endpoints that might reasonably be impacted by the treatment under consideration. A relative effectiveness assessment (REA) needs to convey whether the treatment has a clinically and statistically significant effect on a relevant endpoint compared to some alternative under real-world conditions. What defines a relevant endpoint and how might that endpoint be best presented to convey the information in an unbiased and objective manner?

1.3. Scope/Objective(s) of the guideline

The guideline is intended to describe the common characteristics of clinical endpoints, issues relating to their measurement and presentation, and to briefly outline some of the problems arising when comparing or pooling clinical endpoint data from a number of studies. Finally, this guideline will provide a set of recommendations for the selection and the interpretation of clinical endpoints when completing an REA.

1.4. Relevant EunetHTA documents

This guideline should be read in conjunction with the following documents:

1. EUnetHTA guideline on Endpoints used in REA: Surrogate endpoints
2. EUnetHTA guideline on Endpoints used in REA: Composite endpoints
3. EUnetHTA guideline on Endpoints used in REA: HRQoL
4. EUnetHTA guideline on Endpoints used in REA: Safety.

2. Analysis and synthesis of the literature

2.1. Characteristics of clinical endpoints

An REA compares the effects on clinical endpoints obtained from a health care intervention being evaluated against those obtained from standard therapy or placebo. A clinical endpoint describes the impact of treatment on how a patient feels, functions and survives.(1) Clinical endpoints are often measured in terms of changes in health status (e.g. number of symptoms, disease progression rates, cure rates). In general, long-term or final endpoints are preferred for REA, and the choice of endpoint should be justified and relevant for REA purposes. All-cause mortality is considered to be the most unbiased clinical endpoint.(2)

The choice of endpoint will depend on the target population and main characteristics of a disease (e.g. non life-threatening versus life-threatening disease) as well as on the aim of therapy (e.g. curative versus palliative therapy). Final endpoints will typically measure mortality or survival, whereas non-final endpoints measure morbidity and function. Depending on the context, final endpoints (e.g. survival in curative therapy of a life-threatening disease) are preferred, whereas non-final endpoints may be more suitable to assess treatment benefit in other situations (e.g. HRQoL in palliative therapy or symptoms in non-life-threatening symptomatic diseases).(3) There might also be a hierarchy of endpoints: a non-final endpoint may add important information as a secondary endpoint (e.g. symptom, function, HRQoL) to better explain the significance observed on the primary endpoint (e.g. survival).

2.1.1. Endpoint domains

The following figure broadly summarises main categories of endpoints.

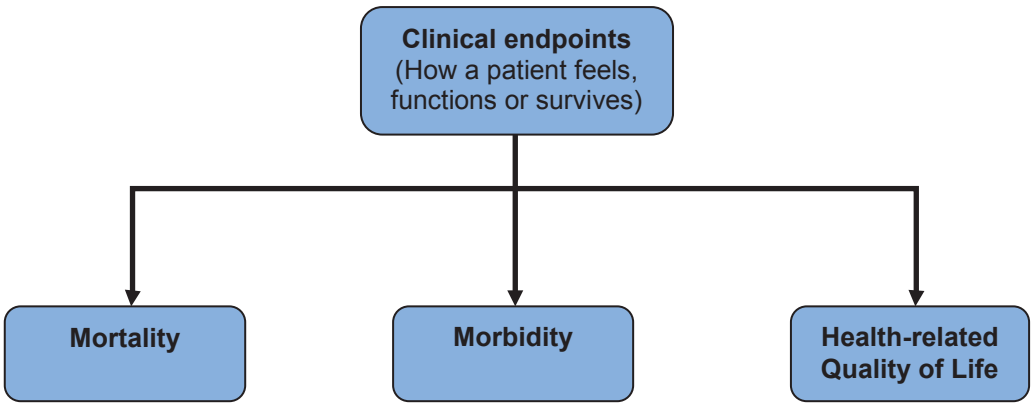

Figure 1. Endpoint domains.

The relevance of the different endpoints will depend on the research question and on the disease and treatment investigated.

The choice of endpoints will be influenced by the purpose for which they are measured.(4) Efficacy studies tend to favour condition-specific endpoints with strong links to the mechanism of action. They also tend to be collected in a short-term horizon. Effectiveness studies tend to collect more comprehensive endpoint measures that reflect the range of benefits expected from the treatment that are relevant to the patient and to the payer, including improvement in ability to function and quality of life. These measures often have a weaker link to the mechanism of action. Both short- and longer-term horizons are considered in effectiveness studies. The distinction between efficacy and effectiveness may be more pronounced for some endpoints, particularly endpoints that are sensitive to individual-level factors.

Common types of endpoints include:

- mortality, either as dichotomous outcome or as time to death
- morbidity events (e.g. myocardial infarction, stroke)
- clinical status (e.g. cholesterol, blood pressure)
- symptoms (e.g. pain, itching)
- function (e.g. Health Assessment Questionnaire Disability Index)
- health-related quality of life (e.g. SF-36)

It is important to note that endpoints may include adverse reactions that reflect on the safety of the treatment (e.g. hospitalisation due to reaction to a drug). Adverse reactions are often collected as secondary endpoints and there is likely to be variation across studies in how adverse events are reported in terms of both detail and terminology. As serious adverse reactions can be anticipated to be relatively rare, studies are not generally powered to detect differences in their occurrence. Furthermore, events such as elevated blood pressure will be of less importance in studies where the primary objective is a reduction in mortality.

Quality of life measures constitute clinical endpoints, given the broad definition of what constitutes a clinical endpoint (how a patients feels, functions and/or survives).(6) However, as HRQoL measures result from assessment of multiple dimensions related to patients' disease and its treatment, they may be susceptible to changes due to a variety of external factors (e.g. life circumstances unrelated to the illness being treated). Therefore, HRQoL is not adequate to be assessed as primary or the only relevant endpoint; it should be better assessed simultaneously with other morbidity or mortality endpoints. The exception may be HRQoL questionnaires that have been specifically developed to capture the specific impact of a given pathology. Specific detail on quality of life endpoints is included in section 2.1.3. See also EUnetHTA guideline on HRQoL.

2.1.2. Intermediate, surrogate and final endpoints

An intermediate endpoint is a clinical endpoint associated with the use of an intervention such as measure of a function or of symptoms (disease-free survival, angina frequency, exercise tolerance) that is expected to correlate with changes observed on final endpoints. It is not the ultimate endpoint of the disease, such as survival or the rate of irreversible morbid events (stroke, myocardial infarction). Improvement in an intermediate endpoint due to treatment is well perceived and can be of value to the patient even if it does not lead to the improvement of morbidity or mortality. Intermediate endpoints may be considered as surrogates (see EUnetHTA guideline on surrogate endpoints). A surrogate endpoint is one that is intended to substitute for a clinical endpoint of interest for REA. A surrogate endpoint is expected to predict clinical benefit or harm based on epidemiologic, pathophysiologic, therapeutic and other scientific evidence. Some examples of surrogate endpoints are blood pressure as a surrogate endpoint for cardiovascular disease or HIV1-RNA viral load as an indicator of viral suppression for HIV interventions.

If data for intermediate endpoints are being assessed, caution must be exercised in directly extrapolating from these to final endpoints unless a clear biological or medical rationale, or a strong or validated link, has been demonstrated.

Final endpoints relate to the final therapeutic objective for the use of the health care intervention, not just to clinical outputs, which is why they have greater relevance for the patient and for overall prioritisation. Where relevant, final endpoints are defined as survival indicators which reflect the probability or the frequency of survival over a specifically defined interval (e.g. years of life gained). An assessment of final endpoints makes it possible to compare different interventions conditional on the final consequences being comparable. It should be acknowledged that the relationship between a surrogate and the endpoint of interest can never be considered as definite. Even if well established for a given health technology, this relationship may be challenged with another health technology that provides the same effect on the surrogate, but with an unexpected effect on the final endpoint. If progression-free survival (PFS) is used as an endpoint there should be sufficient independent evidence to demonstrate that this is associated with overall survival.
In oncology, PFS is an intermediate endpoint that is relevant on its own right. The use of progression free survival has not the same impact in adjuvant and in metastatic disease. In the adjuvant setting, PFS use appears acceptable; in the metastatic setting, data on PFS alone is insufficient and should be coupled with quality of life assessment and survival data, the maturity of which will be considered on the case by case basis.

An REA will often take place at a point when limited trial evidence is available. As such, real-world experience of the effectiveness of the treatment is likely to be limited and the results will inevitably rely to some extent on assumptions and modelling. Extrapolation from intermediate to final endpoints will introduce an additional degree of uncertainty in a REA and should be appropriately investigated and discussed. Where possible, comparisons between treatments should ideally be expressed in terms of final instead of intermediary endpoints.(9)

2.1.3. Patient-reported outcomes (PROs)

A PRO is an umbrella term used to describe any outcome evaluated directly by the patient himself/herself, without interpretation by clinicians or others, and based on patients' perception of a disease and its treatment (see EUnetHTA guideline on HRQoL). PRO data may be collected via questionnaires completed by the patient themselves or via an interview. The latter will only qualify as a PRO where the interviewer is recording the patient's views without any interpretation to form a professional assessment or judgement of the patient's condition. PRO may capture simple measures (e.g. pain measured by the Likert scale), more complex measures (e.g. activities of daily living, functional status on the WOMAC scale), or multidimensional measures (e.g. HRQoL) and satisfaction with treatment).(10,11)

The most commonly used PRO questionnaires assess one of the following constructs:

- Symptoms (impairments) and other aspects of well-being
- Functioning (i.e. disability)
- General health perceptions
- Health-related quality of life
- Perception of or satisfaction with health care

HRQoL represents a specific type/subset of PROs, distinguished by its multi-dimensionality. HRQoL measures may be generic (e.g SF-36) or disease specific, collected using questionnaires comprising global scores and a number of subscale scores that measure different domains of patient's health-related quality of life, such as physical, functional, social and emotional status.(10) A number of condition-specific HRQoL measures are also available (e.g. KOOS).(12)

The clinical relevance of some PROs can be difficult to determine, except in cases where a PRO is a main efficacy endpoint for a given disease (e.g. pain used to assess the efficacy of an analgesic). The concept of a patient's minimal perceptible clinical improvement has been proposed to translate a change in HRQoL score into a marker of clinical improvement.(12) However, clinical relevance of the observed change can be better interpreted by defining responders (e.g. a responder may be a patient who reports that pain has decreased by at least 50% compared to baseline score) although this can have implications for the power to detect treatment benefit (13). Some HRQoL measures have been shown to be unresponsive and some other to be over-responsive to modest changes in status, hence it is important to establish what constitutes a clinically meaningful difference in scores. HRQoL measures can highlight situations where clinicians and patients have divergent views on what is considered important to patients.(14) As HRQoL may be influenced by both benefits and harms of a treatment and is not directly linked to treatment effect, it is possible to detect improvements in a single clinical endpoint in the absence of a change in HRQoL, and vice versa. It should be clear that the instrument used is fit for purpose in the context of use, with evidence to support its validity, reliability and responsiveness so that the results can be able to be more easily interpreted.

As with any endpoint measure, it is critical that a HRQoL measure has demonstrated validity and reproducibility. There are circumstances where it is preferable to use a condition-specific HRQoL measure as compared to generic HRQoL measures (see EUnetHTA guideline on HRQoL)

2.1.4. Composite endpoints

If a single primary endpoint is not suitable and cannot be selected from multiple measurements associated with the primary objective, then another strategy is to combine several measures into a composite endpoint based on a specified algorithm (see EUnetHTA guideline on composite endpoints). Composite endpoints combine multiple (2 and more) relevant single events (e.g. mortality, non-fatal myocardial infarction, stroke, hospitalisation and revascularisation procedures) into one endpoint showing the overall and clinically relevant treatment effect. They are often used where statistical power is poor to increase event rates (e.g. slowly progressive and rare diseases) and decrease sample size and to avoid the issue of multiple testing.

Composite endpoints can make it possible to measure the overall benefit of a treatment in a reasonable timeframe. However, the interpretation can cause problems particularly if the combination consists of endpoints with very different clinical importance.(15) Each of the endpoints included in the composite must meet the requirements of validity, reproducibility, appropriateness, accurate measurement, etc. It is important that patients are followed up after the first non-fatal event as they may subsequently experience further events, including a fatal event.(16) If non-fatal endpoints are included in a composite endpoint, it is important to state whether all non-fatal endpoints were evaluated or just the first one to occur.

Besides the main statistical analysis of the composite endpoint, two sets of analyses should be provided in the study reports:

- the analysis of each component as it counts in the composite endpoint (first event of the composite for a given patient)
- the analysis of each component independently of its role in the composite (notwithstanding a previous occurrence of another component)

To be interpretable as showing an effect of the treatment on the composite endpoint globally, results of REAs using a composite main endpoint should ideally:

- be based on a composite of endpoints relevant for REA
- show a homogeneous response across all the components. The extent to which the response should be homogeneous is open to discussion. At least, the point estimates of each of the components should be in the same direction.

In the event of heterogeneity where the effect on the composite is driven by the effect on one of the components, interpretation of the results will be difficult. Although it may be tempting to conclude that the treatment has a significant impact on the component, it is likely that the data are underpowered to draw such a conclusion.

Valid results on comparable composite endpoints could be handled by considering meta-analytical approaches. If the composite endpoint is not given in disaggregated form it may not be viable to combine the results of several studies due to differences in definition (e.g. use of different components). Varying definitions of composite end points can lead to substantially different results and conclusions.(17). In these cases, each component of a composite endpoint should be individually assessed.

2.1.5. Reproducibility and validity

Reproducibility refers to whether repeated measurements return the same value when there is no underlying change in the condition. Differences can arise due to the individual who takes the measurement, the instruments used to make the measurements, or the context in which the measurement is made. Where appropriate, the inter-rater reliability should be investigated. Inter-rater reliability evaluates the degree to which different raters or observers give consistent estimates of the same phenomenon. Depending on the measure being used, substantial variability may occur across raters. Inter-rater reliability does not apply for self-reported endpoints.

Validity refers to how accurately an instrument measures the endpoint it was intended to measure. Direct measures of final endpoints are presumed to have validity. Clearly, any clinical endpoint must have established validity as shown in independent empirical studies. Reproducibility and validity are not independent as a scale cannot be valid if it is not reproducible, but could be reproducible without being valid.(18)

Endpoint adjudication committees (EPACs) are used in some trials to independently assess endpoints whilst being blinded to study treatments. The purpose of EPACs is to assure the validity of assessments for main trial outcomes. It is not clear to what extent the use of EPACs improve the precision or validity of trial results.(35)

2.1.6. Types of data

Endpoints can be in the form of continuous (e.g. blood pressure, HbA1C), binary (e.g. mortality, disease-free after 6 months), ordinal/categorical (e.g. Likert scale, Clinical Global Impression), or count data (e.g. number of hospitalisations).(19) Categorical endpoints, particularly when expressed as dichotomous endpoints, can be open to manipulation when derived from a continuous measure. For example, the distinction between healthy and ill in an endpoint can be set to show a treatment in the best light if there is no commonly agreed cut-off. Dichotomising does not introduce bias if the split is chosen *a priori* and made at the pre-specified cut-

off. However, if the cut-point is chosen based on analysis of the data, in particular by splitting at the value which produced the largest difference in endpoints between categories, then severe bias will be introduced. The definition of endpoints that are binary by nature, such as myocardial infarction, may still vary considerably across studies.(20) The type of endpoint will depend on the technology under consideration and the trial aims. However, any variations, interpretation or assumptions should be discussed.

2.2. Presentational aspects

The manner in which endpoints are presented in published studies impacts on their interpretation and the feasibility of a meta-analysis. The observed relationship between a treatment and an endpoint is typically expressed in a variety of ways (e.g. odds ratio, risk difference, risk ratio, hazard ratio, standardised mean difference, weighted mean difference, number needed to treat).(21) Each method has merits for clinical interpretability in different contexts. In a responder analysis the result will be presented as a rate of responders rather than a mean difference, although ideally both should be presented.

It is important to note that just because an effect in a study is statistically significant this does not mean that it is of a clinically relevant magnitude. Consequently, it is not sufficient to simply state that a statistically significant effect has been found, or state the confidence interval or p-value. The effect size and the associated uncertainty must be stated and its clinical relevance explained.(22)

2.2.1. Absolute or relative

The endpoints of a trial can be expressed in terms of relative, absolute and numbers needed to treat. It is always advisable to present and compare all three forms as the choice of measure is sometimes chosen to maximise the perceived effect. In some instances the absolute difference will be small whereas the relative difference might be large.(23) Ideally, both relative and absolute measures should be presented.(24) If both are not included then justification of the format reported should be included. Absolute measures are generally useful to clinicians as they provide a more realistic quantification of treatment effect than relative measures.(25)

Relative measures have the advantage of usually being stable across populations with different baseline values and are useful when combining the results of different trials in a meta-analysis. However, relative measures have the disadvantage of not reflecting the baseline values of the patients with respect to the endpoint being measured and the question of whether the relative reduction is really independent of baseline values needs to be specifically addressed. Relative measures do not take into account the patient’s chance of achieving the intended endpoint without the treatment, although this is also the case for absolute measures where all study arms include active treatments. Therefore, they do not give a true reflection of how much benefit the patient would derive from the treatment.(26)

Despite the advantages of absolute measures, they are of limited generalisability due to their dependence on the baseline values. It would be inappropriate, for example, to extrapolate published absolute measures from one population to another population with a different baseline value. Pooling absolute measures in a meta-analysis is highly problematic due to fact that the variation in baseline values not accounted for.(27) By extension, where data are presented without a subgroup analysis it is feasible to apply relative effects to different subgroups with the understanding that baseline values will vary by subgroup and that any interaction between subgroup characteristics and treatment effect is ignored. It is not possible to make such a generalisation using absolute measures.

Absolute measures of efficacy and safety endpoints allow the benefit/harm ratio within a trial to be evaluated. Extrapolation of relative measures of efficacy and safety to populations with different risk profiles allows the benefit/harm ratio to be projected to different populations.

In a meta-analysis of studies with a binary endpoint, the choice of effect measure may have a considerable impact on the analysis, and also on the degree of observed heterogeneity. Empirical studies show that binary endpoint measures of relative effect are more likely to be consistent across trials than measures of absolute effect. This issue is of major relevance to indirect comparisons of two or more sets of trials. An assumption of a meta-analysis is that the effect measure is appropriate.

2.2.2. Time to event

Survival analysis measures *if* the endpoint occurred as well as *when* the endpoint occurred. The statistical method of survival analysis (time to event analysis) is not only used to analyse mortality but is also appropriate for other dichotomous endpoints (e.g. myocardial infarction or combined endpoints).

Survival improvement implies a direct clinical benefit to patients. Common survival outcomes include overall survival, disease-free survival, and progression-free survival. Overall survival is the gold standard for demonstrating clinical benefit and as such should be used where possible. Defined as the time from randomisation to death, this endpoint is unambiguous and is not subject to investigator interpretation. In assessing progression-free survival, patients must be evaluated on a regular basis to ensure that the time of progression is measured accurately.

A key issue in survival analysis is censoring – when time to event data is not known or available for all study participants.(28) Different studies may use quite different censoring, rendering their findings incompatible. Data analysis cut-off dates and schedule of assessment have an impact on the probability of observing events related to the time frame. Incomplete reporting has been shown to be a common problem affecting the definition of survival terms and the numbers of patients at risk.

2.3. Study level issues

The use of multiple endpoints can give rise to Type I error whereby the probability of false-positive findings by chance is increased. A single primary endpoint and multiple secondary endpoints should be defined in the study protocol and appropriate adjustment made for multiple testing.(29) In reality there may be multiple primary endpoints. If multiple primary endpoints are included they should be justified (30). There is debate as to whether or not secondary endpoints should even be reported if the effect on the primary endpoint is not significant.(32)

Reported endpoints may not be per protocol – in many instances, studies put forward the endpoint(s) where the most significant effect was observed.(31,36) To reduce reporting bias, where a systematic review is used all relevant endpoints reported in the literature should be included in an assessment.(33) Where a single study is used, all primary and secondary endpoints should be included.

It is important to consider if the baseline value of the endpoint was measured and reported and whether it was used in the analysis. If assessments have been made over time it should be clear whether the endpoints were assessed at fixed points in time or at variable time-points. The intervals between assessments should be similar between treatment arms to avoid information bias. Ideally the patient, investigator and treating physician should all be blinded to treatment for the purposes of endpoint assessment.

Subgroup analysis may be essential where there are potentially large differences in patient characteristics or treatment benefit that may be observed between groups. Subgroups should have been defined *a priori* with plausible reasons for expecting different treatment effects across subgroups. Subgroup analysis can also pose problems for generating false-positive results – if enough subgroups are tested then one will generate a significant treatment effect. Where possible, trials need to be suitably powered for subgroup analysis and this should be taken into account when analysing trial data. Subgroup analyses are often at risk of Type II error, whereby a genuine treatment effect is not detected because the study was underpowered for that analysis.

The number of patients required to achieve a given statistical power for a study is a function both of the risk in the control group and of the hypothesised reduction in the risk due to treatment. Thus, with declining mortality, the required sample size gets larger. One strategy employed is to enrol high-risk patients, allowing the design of trials with a smaller sample size, but the results may not be generalisable. It is also noted that many reports do not detail the nature of the power and sample size calculations or whether they are testing for superiority, inferiority or equivalence which will impact on the ability to detect a statistically significant difference. As such, some trials may misrepresent their results which will impact on the results of a subsequent meta-analysis.

Endpoint measurement is prone to detection bias if adequate blinding has not been used in a study.(34) At least, an individual who is measuring a clinical endpoint should not be aware of which treatment the patient

has received. Problems with blinding may be partly overcome by the use of an endpoint adjudication committee.

Bias can also be introduced by systematic withdrawals or exclusions from the trial for patients receiving the intervention. In any analysis of endpoint data, it should be clear how missing values were handled as different exclusion or imputation techniques may have implications for bias. If no adjustment for multiple hypothesis testing is made to reduce the risk of spurious findings then an explanation should be given for the non-adjustment.

3. Conclusion

Clinical endpoints are regarded as a means to measure treatment benefit in terms of how a patient feels, functions or survives. That impact is usually in the form of improved health status (e.g. survival, cure, remission), but it may also be worsening health status (e.g. adverse reactions, hospitalisations, deaths). The endpoints reported should have to be clearly relevant to the disease, condition, complaint or process of interest as well as the aim of treatment. Clinical endpoint should be reproducible and valid. A reproducible endpoint will facilitate comparisons across studies and jurisdictions. A valid endpoint will measure what was intended to be measured. Reports of clinical endpoints should be interpreted in terms of reproducibility, validity, and statistical and clinical relevance. Endpoint evaluation by the patient, investigator or treating physician should ideally be done in a blinded fashion.

Clinical endpoints should be presented in natural units and their interpretation should be unambiguous. The clinical endpoint should be measurable for all or most patients within a reasonable time frame.

Although HRQoL measures constitute clinical endpoints, given the broad definition of what constitutes a clinical endpoint (how a patients feels, functions and/or survives), changes in these measures may be difficult to link to the intervention and they are susceptible to confounding by other factors. Therefore, HRQoL measures are not adequate to be assessed as the primary or the only relevant endpoint for REA; they should be better assessed simultaneously with other morbidity and/or mortality endpoints.

Endpoints should be long-term or final and all-cause mortality is preferred when relevant for the scope of the assessment. Overall survival is the preferred clinical endpoint in survival analysis. If it is not feasible to measure final endpoints, then surrogate or intermediate endpoints may be acceptable provided there is compelling independent evidence of a strong association or correlation of effects on the surrogate or intermediate endpoint with the effect on the final endpoint of interest.

Multiple clinical endpoints can be presented, including adverse effects. The hierarchy of endpoints, when adequate, will depend on the disease itself and the aim of treatment. If a composite is used, then it should be possible to disaggregate to the constituent endpoints. If non-fatal endpoints are included in a composite endpoint, it is important to state whether all non-fatal endpoints were evaluated or just the first one to occur.

Both relative and absolute measures should be presented. Absolute measures are useful to clinicians as they provide a realistic quantification of treatment effect which is meaningful for treatment evaluation and prognosis. However, due to the dependence of absolute measures on baseline risk, relative measures are more generalisable across studies. Where a continuous endpoint is converted to dichotomous, there should be a clear justification for the choice of cut-point.

Depending on the country in which the REA is being assessed, there may be a preference for quality of life measures, irrespective of the disease area, whereas other countries may prefer more disease specific measures in an assessment. In all cases, the clinical endpoints reported should meet the characteristics described here. Above all, the statistical and clinical significance are distinct and both should be made evident.

Annexe 1. Methods of documentation and selection criteria

(related to original guideline elaboration in JA1)

Keywords used for the bibliographic research:

- Technology assessment
- Relative effectiveness
- Comparative effectiveness
- Clinical outcome
- Treatment outcome
- End point
- Composite end point
- Follow-up

Sources of information

Sources with English literature were selected in preference, such as:

Data-bases

PubMed (www.ncbi.nlm.nih.gov/pubmed/)
EBSCOhost (search.epnet.com/)

Websites

National Guideline Clearinghouse (guideline.gov/)
National Institute for Health and Clinical Excellence (www.nice.org.uk/)
ISPOR (www.ispor.org/)
Pharmaceutical Benefits Advisory Committee (PBAC)
(www.health.gov.au/internet/main/publishing.nsf/Content/Pharmaceutical+Benefits+Advisory+Committee-1)
Centre for Reviews and Dissemination, University of York
(www.york.ac.uk/inst/crd/)

Guidelines, reports, recommendations already available

European Medicines Agency (www.ema.europa.eu/)
U.S. Food and Drug Administration (www.fda.gov/)

Books

Measuring Patient Outcomes, M.T. Nolan & V. Mock.

Other

Google (www.google.ie/) and Google Scholar (scholar.google.com/)
ScienceDirect (www.sciencedirect.com/)
Wiley-Interscience (onlinelibrary.wiley.com/)
CADTH/CEDAC (cadth.ca/)
Hand searching of references cited in relevant documents
Cochrane Database of Systematic Reviews
(www.thecochranelibrary.com/view/0/index.html)
The Cochrane Collaboration (www.cochrane.org/)
The Cochrane Methodology Register (cmr.cochrane.org/)

Bibliographic search strategy

Where time limits could be specified (e.g. PubMed), the databases were searched for the period 01/01/2000 to 30/06/2010. The searches were restricted to human subjects and the English language.

Database searches used the following search strategy:

PubMed
("technology assessment"[Title/Abstract]) OR "relative effectiveness"[Title/Abstract]) OR "comparative effectiveness"[Title/Abstract]) AND ("clinical outcome"[Title/Abstract] OR "treatment outcome"[Title/Abstract] OR "end point"[Title/Abstract] OR "composite end point"[Title/Abstract] OR "follow up"[Title/Abstract])

EBSCO
(TI "technology assessment" or AB "technology assessment" or TI "relative effectiveness" or AB "relative effectiveness" or TI "comparative effectiveness" or AB "comparative effectiveness") and (TI "clinical outcome" or AB "clinical outcome" or TI "treatment outcome" or AB "treatment outcome" or TI "end point" or AB "end point" or TI "composite end point" or AB "composite end point" or TI "follow up" or AB "follow up")

Selection criteria

Publications were selected as relevant to the current review if clinical endpoints were directly discussed. Discussion could be on the merits of different endpoints or on methodological issues pertaining to the comparison of endpoints. The aim was to obtain publications that give a more general overview rather than providing detailed information specific to a single illness or condition.

In excess of 150 reports, papers and presentations were found by searching Google, Google Scholar and the listed “other” sources. Of these, 56 were found to be directly relevant to endpoints. Due to overlaps in content, only 34 of these articles were used in the final document. The results of the PubMed and EBSCOhost searches were collated to produce a list of 301 articles for consideration. However, none of the articles from the database search provided additional information. To determine if this was due to the choice of search string, a content analysis was undertaken to compare keywords in the relevant articles to those from the database searches. There was a similar distribution of keywords in both groups of articles. Our conclusion is that the subtleties of the research question are such that the construction of a more specific search strategy may not be feasible. We are satisfied that the overall literature search has produced sufficient information to complete the task.

Annexe 2. Bibliography

(1) Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. ClinPharmacolTher 2001; 69 (3): 89-95.

(2) Zanolla L, Zardini P. Selection of endpoints for heart failure clinical trials. European Journal of Heart Failure 2003 Dec;5(6):717-23.

(3) Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technology Assessment 2004;8(36).

(4) Mancia G, Grassi G. Efficacy of antihypertensive treatment: which endpoints should be considered? Nephrol Dial Transplant 2005 Nov 1;20(11):2301-3.

(5) Berger ML, Mamdani M, Atkins D, Johnson ML. Good Research Practices for Comparative Effectiveness Research: Defining, Reporting and Interpreting Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report - Part I. Value in Health 2009;12(8):1044-52.

(6) Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science 1998;282(5387):245-6.

(7) Asmar R, Hosseini H. Endpoints in clinical trials: does evidence only originate from 'hard' or mortality endpoints? Journal of Hypertension 2009;27.

(8) Zanolla L, Zardini P. Selection of endpoints for heart failure clinical trials. European Journal of Heart Failure 2003 Dec;5(6):717-23.

(9) Cleemput I, Van Wilder P, Vrijens F, Huybrechts M, Ramaekers D. Guidelines for pharmacoeconomic evaluations in Belgium. Brussels: Health Care Knowledge Centre (KCE); 2008. Report No.: KCE Reports 78C (D/2008/10.273/27).

(10) Eton DT, Shevrin DH, Beaumont J, Victorson D, Cella D. Constructing a Conceptual Framework of Patient-Reported Outcomes for Metastatic Hormone-Refractory Prostate Cancer. Value in Health 2010;13(5):613-23.

(11) Wiedermann BL. Should You Settle for a Surrogate? AAP Grand Rounds 2009 Apr 1;21(4):41.

(12) Roos E, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. Health and Quality of Life Outcomes 2003;1(1):64.

(13) Snapinn SM, Jiang Q. Responder analysis and the assessment of clinically relevant treatment effect. Trials 2007; 8:31.

(14) Shumway M. Preference Weights for Cost-Outcome Analyses of Schizophrenia Treatments: Comparison of Four Stakeholder Groups. Schizophrenia Bulletin 2003 Jan 1;29(2):257-66.

(15) Kleist P. Composite Endpoints for Clinical Trials: Current Perspectives. International Journal of Pharmaceutical Medicine21(3).

(16) Chi GYH. Some issues with composite endpoints in clinical trials. Fundamental & Clinical Pharmacology 2005;19:609-19.

(17) Lim E, Brown A, Helmy A, Mussa S, Altman DG. Composite Outcomes in Cardiovascular Research: A Survey of Randomized Trials. Annals of Internal Medicine 2008 Nov 4;149(9):612-7.

(18) Goetz CG, Poewe W, Rascol O, Sampaio C, Stebbins GT, Fahn S, et al. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations. Movement Disorders 2003;18(7):738-50.

(19) Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.3). Canberra: Pharmaceutical Benefits Advisory Committee (PBAC); 2008.

(20) Shaw LJ, Iskandrian AE, Hachamovitch R, Germano G, Lewin HC, Bateman TM, et al. Evidence-Based Risk Assessment in Noninvasive Imaging. J Nucl Med 2001 Sep 1;42(9):1424-36.

(21) Bewick V, Cheek L, Ball J. Statistics review 11: assessing risk. Crit Care 2004 Aug;8(4):287-91.

(22) O'Connell RL, Gebski VJ, Keech AC. Making sense of trial results: outcomes and estimation. The Medical Journal of Australia 2004;180(3):128-30.

(23) Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. CMAJ 1995 Feb 1;152(3):351-7.

(24) Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010 Mar 24.

(25) Replogle WH, Johnson WD. Interpretation of absolute measures of disease risk in comparative research. Fam Med 2007 Jun;39(6):432-5.

(26) Akobeng AK. Understanding measures of treatment effect in clinical trials. Arch Dis Child 2005 Jan;90(1):54-6.

(27) Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. BMJ 1999 Jun 5;318(7197):1548-51.

(28) Bland M. An introduction to medical statistics. 3rd ed. Oxford: Oxford University Press; 2000.

(29) Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.

(30) CPMP. Points to consider on multiplicity issues in clinical trials. London, European Agency for the Evaluation of Medicinal Products; 2002.

(31) Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. JAMA. 2009;302(9):977-984.

(32) O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. Controlled Clinical Trials 1997;18(6):550-6.

(33) Bekkering GE, Kleijnen J. Procedures and methods of benefit assessments for medicines in Germany. The European Journal of Health Economics 2008;9(Supplement 1):5-29.

(34) Centre for Reviews and Dissemination. Systematic Reviews - CRD's guidance for undertaking reviews in health care. York: Centre for Reviews and Dissemination (CRD), University of York; 2008.

(35) Hata J, Arima H, Zoungas S, Fulcher G, Pollock C, Adams M, et al. Effects of the Endpoint Adjudication Process on the Results of a Randomised Controlled Trial: The ADVANCE Trial. PLOS ONE 2013; 8(2):1-7

(36) Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR. Comparison of protocols and registry entries to published reports for randomised controlled trials. Cochrane Database Syst Rev. 2011 Jan 19;(1):MR000031.

