HTA CG | MEMBER STATE COORDINATION GROUP ON HEALTH TECHNOLOGY ASSESSMENT

Guidance on outcomes for joint clinical assessments

Adopted on 10 June 2024 by the HTA CG pursuant to Article 3(7), point (d), of Regulation (EU) 2021/2282 on Health Technology Assessment

Contents

List of abbreviations

Abbreviation	Definition
AE	Adverse event
ADE	Adverse Device Effects
ClinRO	Clinician-reported outcome
CI	Confidence interval(s)
COA	Clinical outcome assessment
COMET	Core Outcome Measures in Effectiveness Trials
COS	Core outcome set
COSMIN	Consensus-based Standards for the Selection of Health Measurement Instruments
CTCAE	Common Terminology Criteria for Adverse Events
DD	Device Deficiencies
DFS	Disease-free survival
EFS	Event-free survival
EU	European Union
EUnetHTA	European Network for Health Technology Assessment
Haemoglobin A1C	HbA1C
HRQoL	Health-related quality of life
HTA	Health technology assessment
HTAb	HTA body
HTAR	HTA Regulation (EU) 2021/2282
HTD	Health technology developer
ICHOM	International Consortium for Health Outcomes Measurement
JCA	Joint clinical assessment
MCID	Minimal clinically important difference
MedDRA	Medical Dictionary for Regulatory Activities
MID	Minimal important difference
MOS SF-36	Medical Outcome Study Short Form 36
MOS SF-6D	Medical Outcome Study Short Form 6 Dimensions
MS	Member state
ObsRO	Observer-reported outcome
ORR	Objective response rate
OS	Overall survival
PASS	Patient-acceptable symptomatic state
PerfO	Performance Outcome

Abbreviation	Definition
PFS	Progression-free survival
PGIC	Patient global impression of change
PGRC	Patient global rating of change
PICO	Population, Intervention, Comparator, Outcome
PRO	Patient-reported outcome
PROM	Patient-reported outcome measure
PRO-CTCAE	Patient-reported outcome Common Terminology Criteria for Adverse Events
PT	Preferred terms
RECIST	Response Evaluation Criteria in Solid Tumors
SAE	Serious adverse event
sets-STAD	Core Outcome Set Standards for Development
sets-STAR	Core Outcome Set Standards for Reporting
SOC	System organ class
TTP	Time to progression
WHO	World Health Organization

1 Problem statement, scope, and objectives

Clinical outcome assessments (COAs), used in clinical studies (whether they are interventional (also called experimental), or observational), are a key component of health technology assessment (HTA). They provide the estimate of the effectiveness of the targeted treatment on how patients feel, function, or survive, and the estimate of treatment safety (1). In the context of joint clinical assessment (JCA), outcomes are relevant in two different steps. The first step is during the scoping process (as described in the Health Technology Assessment Regulation (HTAR) guidance on the scoping process), in which Member States (MSs) are expected to request their needs in terms of health outcomes (HTA Regulation (EU) 2021/2282, Article 8(6)) when defining the assessment scope regarding the Population(s), Intervention, Comparator(s), and Outcome(s) (PICO). Defining relevant outcomes is an important component of this process. The second step is the production of the JCA report by assessors and co- assessors based on the dossier submitted by the health technology developer (HTD) and the assessment scope previously defined for a given health technology. While MSs are required to give due consideration to the JCA reports published (Article 13(1)), the clinical relevance or interpretation of the estimate of relative effectiveness and safety may differ between MSs when drawing conclusions regarding the clinical added value of a treatment at a national level. As a JCA report must not contain any judgment on the clinical added value of a health technology, appropriate factual reporting of the methodological and statistical elements and results of the analyses of the outcomes requested is essential to allow MSs to perform the appraisal they deem appropriate according to their national decision-making process (Article 9(1)).

According to the HTAR (Article 8(6)), the assessment scope should reflect the needs of the MSs, and the JCA report should not contain any ranking of health outcomes (Recital (28)). Neither the HTAR nor the HTAR guidance on the scoping process propose criteria for the definition of health outcomes by MSs. However, the health outcomes which are requested during the scoping process will be considered in the JCA report. Indeed, the relative effectiveness and safety of the health technology as estimated based on COAs will be described as required in the assessment scope based on the predefined parameters. However, the ability to conclude on the clinical added value of a treatment can be impacted by several factors such as the appraisal of the level of validity and reliability of outcome measurement instruments or of the validity of surrogate outcomes.

The objectives of this guidance are twofold. The first objective is to provide guidance for MSs in defining relevant outcomes during the scoping process. The second is to help assessors and co-assessors in assessing and reporting all the elements that MSs need for the national appraisal of the clinical added value of a health technology. Of note, the analysis and reporting recommendations for assessors are made with the

implicit assumption that appropriate analyses and information is provided by the HTD. As such, this guidance also has practical implications for the submission dossier and assessment report which should be taken into account in the preparation of these documents and associated guidance.

For simplicity, effectiveness is the term used to describe efficacy or effectiveness throughout this document. Furthermore, treatment, intervention and health technology are all terms used for any health technology (medicinal products and medical devices) that can be assessed.

2 Definitions, sources of information, and general considerations

2.1 Definitions

“Outcome” is any concept that can be used for the estimation of relative effectiveness and safety of a health technology, such as mortality, remission, disease control, function, health-related quality of life (HRQoL), and symptoms (1).

Outcomes are distinct from the way in which they are measured, and an outcome can be measured in several ways. The “measure of an outcome” corresponds to the attribute “variable or endpoint” of the estimand framework of the International Council for Harmonisation of technical requirements for pharmaceuticals for human use (2). It is defined by how the outcome is assessed at a patient-level (including use of a specific outcome measurement instrument; see Section 5). For instance, if the outcome is mortality, the measure of the outcome could be how many days the patient is alive during follow-up (if analysed as time-to-event data), or whether the patient is alive or dead at a specific time point (e.g., one year after inclusion). If the outcome is pain, the measure of the outcome could be, for example, the change in the level of pain on a patient-reported numeric rating scale (from 0-10) at 6 months after initiation of the treatment.

An accurate definition of the measure of the outcome allows the estimation of a population-level summary measure (or summary statistics). For example, regarding mortality, overall survival estimated by the Kaplan-Meier method or “proportion of deaths one year after inclusion” are summary measures corresponding to each way of measuring mortality described above. For pain, “mean change in the level of pain 6 months after the initiation of the treatment” is a possible summary measure.

It is sometimes argued in the literature that “outcome” defines the concept whereas “endpoint” defines the measure or summary measure (3,4). However, the two terms are frequently used interchangeably (5). In this guidance, we only use the terms “outcome”, “measure of an outcome”, and “summary measure”.

Lastly, “effect measures” are the statistics that are used to express the effectiveness of a treatment (6). HTA, according to the HTAR (Recital (2)), “focuses specifically on the added value of a health technology in comparison with other new or existing health technologies” (i.e., assessment of relative clinical effectiveness and relative safety). Thus, effect measures are understood as a comparison of the summary measures of outcomes between groups (e.g., between intervention and control). Broadly, effect measures are either difference measures (e.g., mean difference in change, risk difference) or ratio measures (e.g., risk ratio, odds ratio, hazard ratio). However, other statistics can be used to express other aspects of a treatment effect such as within-group change (7). The precise description of the treatment effect of interest reflecting the clinical question posed by the trial objective, in terms of five attributes (treatment,

population, variable, intercurrent events, population-level summary), can be achieved using the estimand framework (2). It can be seen as a complement to the PICO framework. The two frameworks are not equivalent but overlaps between attributes of the two frameworks are covered in the HTAR "*Guidance on reporting requirements for multiplicity issues, subgroup, sensitivity and post-hoc analyses in joint clinical assessments*".

2.2 Types of outcomes according to source of information

It can be useful to classify outcomes according to the main source of information via which they are collected (4,8–10). Identification of adequate source(s) of information can help in defining relevant outcomes during the scoping process.

Categories for classifying outcomes are not mutually exclusive, as some measurement instruments for some outcomes require the collection of elements from multiple sources. For example, the Clinical Disease Activity Index for rheumatoid arthritis requires clinical and patient-reported elements (11).

2.2.1 Clinician-reported outcomes (ClinROs) and Performance outcomes (PerfOs)

Clinician reported outcomes (ClinROs) are assessments made by healthcare professionals based upon clinical examinations of patients and involve clinical judgments of patients' observable signs, behaviours, or other physical manifestations, often in combination with healthcare professionals' own assessment of symptoms reported by patients and / or their caregivers. ClinROs can be assessed using only the results of a clinical examination, or in combination with technologically assessed ClinROs (also referred to as biomarker data, assessed using for example laboratory tests or medical imaging) to report clinical findings or events, such as pulmonary or cardiac function. ClinROs can also be rated on specific scales using outcomes measurement instruments (see Section 5). Performance outcomes (PerfOs) are a particular case of outcomes reported by healthcare professionals, as they require active patient involvement to complete a standardised task (e.g., 25-foot walk test with ankle-worn sensor, cognitive tests).

2.2.2 Patient-reported outcomes (PROs)

- Patient-reported outcomes (PROs) are defined as "any report of the status of the patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" (12). They are measured by patient-reported outcome measures (PROMs), which are often self-administered questionnaires with pre-specified response formats (e.g., Likert scale or numeric scales). Other formats such as unstructured surveys can also be used. The PRO concept is sometimes equated to HRQoL, although this is a limited interpretation, as HRQoL is only a subset of the outcomes that can be

measured using PROMs. While there is no universal definition of HRQoL, there is generally consensus that, a) it is a multi-domain concept, and b) HRQoL instruments assess the subjective perception of the impact of a disease and its treatment(s) on one's daily life, physical, psychological and social functioning and well-being (13,14). Some PROMs measure health status (for instance, the EQ-5D instrument measures health status as a combination of five broad concepts (15)). Of note, some of the PROMs that measures heath status, such as the five dedicated items of the EQ-5D, or HRQoL, such as the Medical Outcome Study SF-6 Dimensions (MOS SF-6D), can be used to measure utility. When they are used for that purpose, they can be referred to as multi- attribute utility instruments or generic utility instruments (16). Other outcomes such as symptoms (including fatigue and pain), anxiety, depression, functioning, impairment, disability and impact on daily living can be assessed using PROMs.

2.2.3 Observer-reported outcomes (ObsROs)

An observer-reported outcome (ObsRO) is a measurement based on an observation by someone other than the patient or a healthcare professional (8). This may be a parent, a significant other relative, or another non-clinical caregiver who is able to regularly observe and report on specific aspects of the patient's health. An ObsRO measure does not include professional medical judgment but is still dependent on the interpretation of the observer. For patients who cannot respond for themselves (e.g., infants or cognitively impaired), observer reports should preferably target the reporting of events or behaviours that can be directly observed. As an example, observers cannot validly report an infant's pain intensity but can report infant behaviour thought to be caused by pain (e.g., crying). In situations where patient cannot respond for themselves, ObsROs are also referred to as outcomes reported by a "proxy" (8).

2.2.4 Digital outcomes

There is increasing use of patient-generated health data provided directly via health technologies (also called digital outcomes). Some medical devices can offer automated measures of outcomes in non-clinical settings, such as the home (4,17). Digital outcomes might include those used for COAs (an example would be actigraphy instead of 6-minute walk test), those that impact the actual use of the product in everyday life (improve adherence) or a combination of both (technologies that can be used for COAs but also impact treatment decisions by for example patient feedback if co-packaged with the product). Digital outcomes can be collected at a high frequency, even continuously, but analyses can be challenging due to data handling in the context of the European general data protection regulation requirements, or because large datasets may be collected that are challenging to analyse. They do not constitute a particular source of information per se. Depending on the purpose and usage of the technology, the outcomes they produce can be considered ClinROs, PerfOs, or PROs. But, due to their novelty and due to some specific methodological challenges

associated to these types of outcomes, it is of particular importance that measurement instruments in this outcome domain are well validated in the respective study populations before their use in a clinical study. For example, “well validated” in the context of novel medical devices measuring an outcome of interest means that the device has at least undergone testing and validation processes to demonstrate its validity and reliability in measuring the intended outcome. During validation process, studies should compare the performance of the device against established standards or reference methods.

2.3 General considerations

Concept and measure

During the scoping process, defining an outcome as a *concept* only (e.g., HRQoL without further specifications on its *measure*) maximises the opportunity for an HTD to provide at least one result relevant to that outcome. However, the HTD could provide a result using a measure of the outcome that could be considered inappropriate (e.g., because the measure is appraised as having an insufficient level of validity). The validity, reliability, and interpretability (see Section 5) of the measure of the outcome provided by the HTD would therefore need to be appraised by MSs based on the elements reported within the JCA, as part of their national decision-making process. Conversely, a more specific request (e.g., HRQoL measured as a change in score for the Medical Outcome Study Short-Form-36 (MOS SF-36 PROM) may help in specifying a measure considered appropriate by an MS, but with a higher risk of not receiving those specific results if the outcome was assessed differently in evidence submitted by the HTD. To alleviate this issue, a general recommendation could be to formulate a request as such: “[Outcome of interest] measured **preferably as** [insert measure]”. This phrasing allows MS to express a preference regarding a specific measure but will lead to the report of results even if the evidence submitted by the HTD regarding the outcome of interest has been measured another way. Requesting an outcome without the specific mention of a measure (or with mention of a preference only) entails the risk of selective reporting, especially when an outcome has been measured in several ways in clinical study results submitted by the HTD (e.g., HRQoL measured by two different outcome measurement instruments). Therefore, in that situation, all results from analyses pre-specified in the original studies assessing a treatment effect on the outcome of interest must be reported in the JCA.

Timing

As sufficient duration of follow-up is an important element when interpreting the results of a specific outcome, a related issue is the *timing* of outcome assessment. A request such as “rate of major adverse cardiovascular events 2 years after inclusion” specifies a timing. However, the timing aspect is at risk of not obtaining results, if, for example, follow-up was not sufficiently long in the clinical study submitted as evidence. Such a request of one specific time point could also hamper the presentation of results

according to statistical modelling such as mixed models for repeated longitudinal data. A general recommendation would be to formulate a request as follows: "[Outcome of interest] measured **preferably** [insert timing of assessment]". As for measures, this phrasing allows MS to express a preference without hampering presentation of results in the JCA report if timing of assessment is different in evidence submitted by the HTD.

Effect measure

Lastly, a more detailed level would be to request a specific *effect measure*. In general, we would advise that specifying an effect measure is not desirable. Indeed, the choice of an effect measure is highly dependent on underlying assumptions regarding statistical analyses. Therefore, it is first the responsibility of the HTD to provide results expressed in terms of effect measures according to good clinical and statistical practice. Nonetheless, if a MS wants to specify an effect measure, this should be done using the previously mentioned template: "[Outcome of interest] with treatment effect expressed **preferably as** [insert effect measure]". As for measures, requesting an outcome without the specific mention of an effect measure (or with mention of a preference only) entails the risk of selective reporting, especially when treatment effect have been expressed using several effect measures in evidence submitted by the HTD (e.g., because of the conduct of sensitivity analyses). Therefore, in that situation, all results from pre-specified analyses assessing treatment effect on the outcome of interest in the original studies must be reported in the JCA.

Summary

- Outcomes are concepts for estimating treatment effectiveness and safety.
- The measure of an outcome defines accurately how the outcome is assessed as a variable.
- Effect measures are statistics that compare the estimates of outcomes between groups.

Points of attention for the assessment scoping process

- Proposing a very specific outcome can impact the reporting of results in a JCA (e.g., as an outcome only, or by specifying a measure, time point for assessment and/or by specifying an effect measure).
- If a specific measure of an outcome is desired by a MS, the wording should follow this template: "[Outcome of interest] measured preferably as [insert measure]".
- If a specific time point for assessment is desired by a MS, the wording should follow this template: "[Outcome of interest] measured preferably [insert timing of assessment]".

- MSs are advised not to specify effect measures. The HTD is responsible for presenting results using appropriate effect measures in accordance with good clinical and statistical practice.
- If a MS still wants to specify an effect measure, the wording should follow this template: “[Outcome of interest] with treatment effect expressed preferably as [insert effect measure]”.

Requirement for JCA reporting

- Provide a detailed and relevant definition of outcomes (concept, main source of information, measure, timing, observation period, summary and effect measure) of any reported outcome. If necessary, references provided by the HTD on how an outcome is defined, measured at a patient-level and population-level, and how effect measure is expressed.
- Baseline values must be reported whenever necessary, as results associated to some outcomes are better understood in the context of baseline values (e.g., when looking at change in scores of an outcome measurement instrument).
- When no specific measures or only preferences are requested in the assessment scope, all measures from analyses pre-specified in the original studies to assess the treatment effect on the outcomes of interest must be reported in the JCA report.
- When no specific effect measures or only preferences are requested in the assessment scope, all effect measures from analyses pre-specified in the original studies to assess the treatment effect on the outcome of interest must be reported in the JCA report.
- Risk of bias associated with the report of outcomes such as omissions of important measures, timepoints, changing of measure scoring, additions of unplanned analyses, differences between study protocol and statistical analysis plan, as well as the rationale for these deviations (as provided by the HTD) must be reported in the JCA report.

3 Clinical Relevance

Several outcomes are considered adequate in clinical studies and in HTA methodology to measure the clinical benefit to the patient. Some outcomes may be fully acceptable as support for the risk/benefit ratio assessment of a certain therapy but are less suitable for the needs of JCA. This may be the case for surrogate outcomes (see the definition in Section 3.3). Surrogate outcomes may or may not reflect a direct patient-centred benefit and their clinical relevance and fit to the JCA need to be considered by MSs. Outcome assessment can be reported at different time points in clinical studies. The timing of outcomes assessments is crucial to understanding the long-term and short-term effects of an intervention. The choice of when to assess outcomes depends on the nature of the intervention and the research question. Chronic diseases or preventive interventions may necessitate long-term follow-up to observe sustained effects. In that context, long-term or final outcomes (i.e., the occurrence of an irreversible event of primary interest such as death) are preferred in HTA. However, in acute conditions, short-term outcomes may be more relevant, e.g., symptoms, HRQoL or adverse events. The acceptability of an outcome is subject to MSs' interpretation of its relevance within their national process for decision-making and thus may differ between MS.

3.1 General considerations for patient-centred outcomes

Patients may not consider that all outcomes are equally important. In contrast to physician-centred care, the term "patient-centred outcomes" refers to outcomes that directly measure mortality, morbidity and outcomes related to patients' feelings, beliefs, preferences, needs and functions (such as the ability to perform activities in daily life) (18–20). Decisions on what is a patient-centred outcome for the PICO question for a particular health technology should ideally be taken in close collaboration with patients who either live with the medical condition and/or are knowledgeable about the condition, and healthcare professionals related to the specific disease area. Some core outcome sets (COS) (see Section 3.2 for more details) are defined with patient involvement, such as the International Consortium for Health Outcomes Measurement (ICHOM) or the Core Outcomes Measures in Effectiveness Trials (COMET) initiatives (21,22). While pondering the potential limits of the use of COS in the context of HTA is necessary (see Section 3.2), these may be helpful for MSs in defining patient-relevant outcomes during the scoping process. However, the final decision for outcome selection is up to each individual MS. It is expected that there will be an overlap in choices of what are considered patient-centred outcomes for JCA with PICO question requests in most cases. In addition, classifications such as the International Classification of Functioning, Disability and Health of the World Health Organization (WHO) (23), the Wilson and Cleary biopsychosocial model (24), and the Montreal Accord on PROs (8) can provide further information to input into MSs' discussions around health outcomes.

As noted above, frequently, long term or final outcomes are preferred in HTA. All-cause mortality is an outcome that is objective, easy to measure and definite since the final time point is death. Mortality might be measured either as overall survival or mortality rates/survival rates for a given period (e.g., 1-year mortality or 5-year mortality). However, for example, for diseases with expected long-term survival, it might be impossible to obtain long-term mortality data from clinical studies at the time at which the JCA report is generated. Short-term data may then have to be reported, when available, in combination with requested or validated surrogate outcomes for long-term data (see Section 3.3 for more details on surrogate outcomes). Duration of follow-up and surrogacy should be clearly stated in JCA reports. A supposed impossibility of obtaining long-term data should not refrain MSs from formulating the desired outcome during the scoping process. COA related to patients' response to the therapy can be reported either as morbidity events or in terms of "time to event" (in the case of the occurrence of irreversible binary events) or as the change in clinical status or symptoms. A range of outcomes measurement instruments may be used to capture relevant information about patients' health status and the disease response to a given therapy. It is crucial that an "event" is well defined and that only validated tools for measurement are used. Time points for COA and the frequency of these assessments, as well as the expertise of outcome assessors involved (e.g. when assessing ClinROs, PerfOs), are of importance for the results and should be reported.

Lastly, Section 5.3 discusses the assignment of a qualitative meaning to an instrument's quantitative scores or change in scores, which can be considered as part of the concept of "clinical relevance".

Points of attention for the assessment scoping process

- Outcomes relevant for HTA should be long-term or final where possible.
- If it is not feasible to measure final outcomes, then intermediate or surrogate outcomes may be acceptable if there is evidence of a strong association or correlation of effects on the surrogate or intermediate outcome with the effect on the long-term or final outcome. A supposed impossibility of obtaining long-term data should not refrain MS from formulating the desired outcome during the scoping process (see also Section 3.3).
- Depending on the research question also short-term outcomes (collected for an appropriate study duration) are relevant, e.g. symptoms, HRQoL or adverse events.

3.2 Outcomes for specific therapeutic areas

Identifying a standardised set of outcomes, defined as a COS, that should be assessed and reported as a minimum in all clinical trials, is an ongoing process carried out by the international healthcare and scientific community for several specific medical or therapeutic areas (21). COS were first defined in rheumatology which is typically a

chronic heterogeneous condition affecting more than one organ (see the OMERACT initiative (25)). COS have subsequently been defined in various medical fields and healthcare settings (21). The relevance of COS is highlighted when facing common conditions such as cancer. Key initiatives include the COMET COS database (26), the ICHOM (22), the Core Outcome Set Standards for Development (sets-STAD) (27), and the Core Outcome Set Standards for Reporting (sets-STAR) (28).

There are several potential benefits from COS:

- by involving a wide range of stakeholders, such as patients, caregivers and health care professionals, it is more likely that patient-centred outcomes will be identified;
- by reducing heterogeneity in COA in original clinical studies, the use of COS may facilitate the conduct of evidence synthesis.

Initiatives for defining COS are also proposed for specific types of outcomes in a given medical field. A recent review investigated the scope, outcomes, and development methods for consensus-based COS for cancer, and the approaches and criteria for selecting outcomes measurement instruments to assess core PROs (29). The conclusion was that there is a lack of recommendations on how to measure core PROs, such that efforts to standardise COA via the development of COS may be undermined. It was suggested to optimise the usefulness and adoption of COS that valid and reliable outcomes measurement instruments for assessment of core PROs should be recommended.

A study proposing a methodological approach for assessing the uptake of a COS for rheumatoid arthritis revealed that the COS was measured and reported in approximately 80% of recent trials of a disease-modifying antirheumatic drug (30). However, a systematic review concluded that COS uptake in new clinical studies and evidence syntheses needs improvement, as uptake is still low in most research areas (31).

COS can be helpful in suggesting and standardizing the definition of relevant outcomes during the scoping process. However, MSs should choose outcomes based on their own needs for decision-making, and consider that other health outcomes deemed relevant by patients, caregivers, clinical experts, or Health Technology Assessment bodies (HTAbs), can complement the use of COS.

Since cancer is one of the leading causes of death worldwide and the stepwise approach to performing JCA in the HTAR establishes oncological medicines as the first group of therapeutics to undergo JCA, this document has incorporated outcomes for assessing safety and effectiveness of new cancer treatments. Specific definitions of outcomes typically used in oncology are provided in Appendix A.

Points of attention for the assessment scoping process

- Consider recommendations from available well-established COS when selecting outcomes.
- Other patient-centred outcomes, deemed relevant by patients, clinical experts, or HTAbs, can complement the use of COS.

3.3 Surrogate outcomes

3.3.1 General considerations

A surrogate outcome is an outcome that is intended to replace an outcome of interest that cannot be observed in a specific clinical study (4). It is a variable that provides an indirect measurement of effect in situations in which direct measurement of a patient-centred effect is not feasible or practical (32). A surrogate outcome may be a biomarker that is intended to substitute for a patient-centred outcome, or it may be an intermediate outcome. A surrogate outcome is expected to only predict the treatment effect of an outcome that is not observed in a clinical study. A biomarker is a surrogate which can be defined as a characteristic that is an objective measure of an indicator of normal biological processes, pathogenic processes or pharmacological responses to an intervention (33). Examples of biomarkers used as surrogacy for mortality and/or morbidity and/or disease remission include levels of cholesterol, haemoglobin A1C (HbA1C), and antibody titre after vaccination. An intermediate outcome is a surrogate outcome such as a measure of a function or of a symptom (disease-free survival, angina frequency, exercise tolerance) but is not the long-term or final outcome of the disease, such as survival or the rate of irreversible morbid events (stroke, myocardial infarction) (34).

The use of surrogate outcomes in the assessment of the relative effectiveness of a health technology can be controversial since the validity of surrogate outcomes has rarely been fully established in a rigorous manner (35–38). Only a few surrogate outcomes have been shown to be true measures of tangible clinical benefit.

Points of attention for the assessment scoping process

- Final patient-centred outcomes such as mortality, morbidity, and HRQoL should be requested during the scoping process.
- Requesting a validated surrogate outcome only to replace a patient-centred outcome of interest should only be done if absolutely necessary.
- Surrogate outcomes can be requested in addition to patient-centred outcomes where relevant. However, only surrogate outcomes for which validity has previously been clearly established should be requested where possible. This may not be possible at the scoping stage in many instances, although in some cases this might have been established by previous JCAs or in other literature on the same indication.

3.3.2 Association between surrogate outcomes and patient-centred outcomes

If the HTD is unable to provide data for an outcome of interest that has been specified in the scope, but another outcome (regardless of whether this has been requested in the scope or not) is believed to provide indirect information regarding the outcome of interest (i.e. is considered a surrogate outcome for the outcome of interest), this should be described in the dossier. The HTD should explain for which outcome of interest a surrogate is applied and demonstrate the strength of the association between the surrogate outcome and the outcome of interest and the association of treatment effects on the surrogate and the outcome of interest (see Section 3.3.3 for more details on level of evidence for surrogacy). For example, if the outcomes "mortality" and "HbA1C" are both specified as outcomes in a scope, the HTD is only required to demonstrate the strength of the association between HbA1C and mortality, if the data for mortality is not available or very limited (e.g. due to very limited events), and the HTD considers HbA1C to provide information about the technology's expected effect on mortality.

3.3.3 Level of evidence for surrogacy

The appraisal by MSs of the association between the surrogate and the long-term or final outcome should take into account all of the following levels of evidence (37):

- Level 1: evidence demonstrating that treatment effects on the surrogate outcome correspond to effects on the patient-centred outcome (from clinical trials); comprises a meta-analysis of several randomised controlled trials (preferably from a systematic literature review); and establishment of correlation between effects on the surrogate outcome and the patient-centred outcome in the respective disease stage and sufficiently restricted to the interventions investigated. While MSs can appraise the validity of a surrogate outcome submitted by the HTD with regard to their national decision-making process, it is highlighted in literature that this level can be considered as required (with evidence showing a sufficiently strong correlation) for demonstrating the validity of a surrogate outcome (39,40);
- Level 2: evidence demonstrating a consistent association between the surrogate outcome and the final patient-centred outcome (from interventional, epidemiological or observational studies);
- Level 3: only evidence of the biological plausibility of an association between the surrogate outcome and the final patient-centred outcome (from pathophysiological studies and/or an understanding of the disease process).

There is no universally accepted threshold for the establishment of sufficient correlations between the surrogate and the patient-centred outcome at trial level (i.e., correlation of the effects) and patient level (i.e., correlation of the outcomes). However, a correlation of at least 0.85 is described as "high" and can be used as a criterion for

validation of surrogate outcomes (39). There are several other useful approaches for the validation of surrogate outcomes. In general, these methods are based on a meta-analytic approach (40). The concept of the surrogate threshold effect is helpful for decision-making because it represents the minimum effect regarding the surrogate outcome that is required to conclude that there is also high certainty of an effect on the patient-centred outcome (41).

3.3.4 Uncertainty surrounding the use of surrogate outcomes

A surrogate outcome may lead to greater uncertainty surrounding the benefit of the technology under assessment. Therefore, elements described below should be considered in a JCA for allowing MS to appraise the validity of surrogacy for their national decision-making process (see requirements for JCA reporting of this section). In addition, there are several frameworks that may be useful when assessing uncertainty surrounding the use of surrogate outcomes. These include reports by Ciani et al. (37,42), Grigore et al. (43), Bujkiewicz et al. (44), and guidelines on preparing a submission to the Australian Pharmaceutical Benefits Advisory Committee (45).

Requirements for JCA reporting

The assessor should consider the following for the JCA report:

- The level of evidence for the association between the surrogate outcome and the final patient-centred outcome.
- Details on whether this association is based on biological plausibility and/or empirical evidence.
- A description of whether this association has been studied in the disease stage, population and intervention of interest.
- In cases for which the association between the surrogate outcome and the final patient-centred outcome has previously been examined but for a different disease stage, population or intervention, the assessment report should consider the implications for the validity of this association in the current population and intervention of interest.
- The strength of the association between the surrogate outcome and the patient-centred outcome.
- The strength of the association between the treatment effect on the surrogate outcome and the patient-centred outcome.
- Any uncertainties associated with the evidence and quantified this if data are available.
- The limitations of the use of a surrogate outcome should be explicitly explained.

- An indication of whether a patient-centred outcome is likely to be available at a later date.

- Clearly outline any remaining areas of uncertainty.

3.4 Composite outcomes

Composite outcomes are treatment effects expressed in terms of an outcome which comprises a combination of two or more 'associated' clinical events (46). Frequently, composite outcomes comprise combinations of mortality and morbidity events (such as the outcome "major adverse cardiac events" which can combine death but also morbidity events such as stroke and myocardial infarction), or combinations of different morbidity events (e.g., skeletal events in prostate cancer trials). The most common rationale for choosing a composite outcome as an outcome is that each individual event / component is rare and by combining them, the percentage of the study population with an event increases, thus increasing the power of the study (47). Other rationales can be to mitigate the issue of multiplicity when performing statistical hypothesis testing (as only one outcome is compared between intervention and control instead of multiple ones), or to aggregate the measure of the treatment effect by pooling events that are considered clinically similar or coherent (e.g., such as a composite outcome including deep vein thrombosis and pulmonary embolism) (46).

A composite outcome can be analysed as a binary outcome (occurrence of any event), but more often, the time to the occurrence of the first event is the measure of interest in the context of a time-to-event analysis.

Interpretation of the treatment effect using composite outcomes is a critical issue and requires, at minimum, that the effects of an intervention on individual components are expected to align (i.e., similar effects of a treatment on each individual component are expected from published research and/or clinical expertise), and that the effect of one component is not mediated by another one. Consideration should also be given to the clinical relevance of the individual components (e.g., acute symptomatic events are commonly not combined with asymptomatic events). Last, it is also expected that each individual component is assessed with a valid and reliable measure (see Section 5 for more details).

In addition to *a priori* considerations on the meaningfulness of composite outcomes and the selection of individual components, the interpretation of results for a composite outcome also depends on the availability of results for each individual component.

As a general rule, results of individual components of a composite endpoint should be presented both as an actual contribution to the composite endpoint as well as a separate variable. In case of the actual contribution of the individual components, there may arise the problem of competing risks because a patient who experienced one

component will not be counted for another component. In this case, the interpretation of effect measures for the contribution of each individual component is not straightforward. Methods that correctly account for competing risks are recommended (such as the cause-specific hazard model (48) and the subdistribution hazards model (49)), for an appropriate interpretation of effect measures of the contribution of each component to the composite outcome.

Points of attention for the assessment scoping process

- When proposing a composite outcome, it is advised to carefully consider its interpretability. Published research and clinical expertise may aid in the selection of individual components that are expected to be affected by a given treatment in a similar fashion.
- If a MS wants to specify a composite outcome, the following wording template should be used: “[composite outcome of interest] preferably comprising the following individual components: [insert components]”.

Requirements for JCA reporting

- For all composite outcomes, in case the individual components have been requested by a MS as individual outcomes, in addition to the results according to the composite outcome, report results for individual outcomes separately.
- In case where only a composite outcome has been requested as part of a PICO, the JCA report should entail results for the contribution of each individual component to the composite outcome.
- If a composite outcome is reported, assess carefully whether the reported results follow the study protocol and statistical analysis plan of the corresponding clinical study with respect to the individual components. In case of deviations (e.g., omission of a component in the final results, changes in the number and/or definition of the individual components during the conduct of the study), report any rationale that the HTD provides.

4 Safety

4.1 Terminology for JCA

It is important that a JCA uses consistent and precise terminology to avoid confusion and misleading conclusions.

This guidance is not intended to duplicate the definitions already provided for safety terminology. In the context of JCA for medicinal products, the term “adverse event” (AE) must be used, and the terms “adverse reaction”, “adverse drug reaction”, “side effect”, “serious incident”, and “adverse effect” should be avoided as they imply a causal relationship of a single event to the intervention. The term “safety” must be used, and “tolerability” and “toxicity” should be avoided.

The fact that the definition of AE and serious adverse event (SAE) in medical device studies (50) differ minimally from medicinal products (51) is left aside in the following recommendations, which means that the product-specific definition should be applied in each case.

Requirements for JCA reporting

- Use the term “safety” instead of “tolerability” or “toxicity”.
- Use the term “adverse event”, instead of “adverse reaction”, “adverse drug reaction”, “side effect”, “serious incident”, or “adverse effect”.

4.2 Information to be reported

Safety outcomes can be defined according to different terminologies. Medical Dictionary for Regulatory Activities (MedDRA) is used for interventional studies (52). Patient-reported information related to safety could also be used, such as Patient-Reported Outcome Common Terminology Criteria for Adverse Events (PRO-CTCAE) (53). In the context of a JCA, the use of MedDRA terminology when reporting safety outcomes is preferred. If a different terminology is used, the HTD should provide a detailed justification.

Safety outcomes can be graded for severity using different scales. Common Terminology Criteria for Adverse Events (CTCAE) is typically used for interventional studies in oncology but can also be used in non-oncology trials (54). A WHO scale has also been developed for grading acute and sub-acute AEs in oncology (55). When the severity of AEs has been graded in the primary study, the JCA must describe the scale used.

Seriousness (serious, nonserious) should also be reported. A serious adverse event is an AE that results in death, is life-threatening, requires hospitalisation or prolongation of existing hospitalisation, results in persistent or significant disability or incapacity, or is a birth defect.

Discontinuation due to an AE (or “adverse event leading to withdrawal”) must be reported. Interruption due to an AE must also be reported.

In general, there is no rationale to only report AEs potentially related to the health technology under study. Specifically, in unblinded studies, there is a high risk of bias in the assessment of causality of AEs. Therefore, a report of AEs irrespective of any assessment of causality must always be available. Nonetheless, for medical devices, attribution of causality to the procedure or device of AEs can be in specific cases straightforward. Some AEs can obviously be related to the device itself, e.g., premature battery depletion, or to the procedure, e.g., surgery incidents (50).

In summary, when safety is required as an outcome in the assessment scope without further specifications, the following descriptive results (i.e., absolute numbers of patients with events and percentages per treatment arm without relative effect measures and nominal p-values) must be reported in the main text of the JCA for each treatment group (i.e., intervention and comparator):

- AEs in total (i.e., all AEs combined irrespective of seriousness),
- Serious AEs,
- Severe AEs with severity graded to pre-defined criteria (e.g., according to CTCAE Grade ≥ 3),
- Death related to AEs (e.g., according to CTCAE Grade 5),
- Treatment discontinuation due to AEs,
- Treatment interruption due to AEs.

For the assessment of medical devices, the following must be reported in addition to the bullet points above (descriptive results in summary as well as for each AE individually):

- Adverse Device Effects (ADE Any adverse event related to the use of an investigational medical device or a comparator)
- Device Deficiencies (DD: Any inadequacy in the identity, quality, durability, reliability, safety or performance of an investigational device, including malfunction, use errors or inadequacy in information supplied by the manufacturer)

If one or several specific AEs are of interest for a MS, they should be requested explicitly in the assessment scope (e.g., symptomatic osteonecrosis of the jaw with bisphosphonates). In that case, in addition to the aforementioned results, these specific AEs will be described irrespective of seriousness.

For medicinal products, results regarding relative safety (i.e., with effect measures, nominal p-values and 95% confidence intervals (CI)) for the AE categories described above and in addition for AEs according to system organ class (SOC) and preferred terms (PT) must be provided in a dedicated appendix of the JCA report (52). AEs according to SOC and PT of any severity must only be included in the appendix if they occur with an incidence of ≥ 5 % in any treatment group. Serious and severe AE (e.g. according to CTCAE Grade ≥ 3) must be included regardless of their incidence.

Requirements for JCA reporting

- MedDra terminology for reporting safety outcomes should be preferred. The use of a different terminology should be adequately justified by the HTD.
- Describe the main source of information (healthcare professionals, medical technology, patients) for reporting safety outcomes.
- Report irrespective of causality for medicinal products.
- Report irrespective of causality ADE and DD for medical devices.
- Report the median (min; max) and mean (SD) observation period (data collection for AEs)
- Provide in the main part of the JCA report a tabular descriptive report, including numbers (X out of XY subjects) and percentage (% of subjects) for AEs in each treatment group (i.e., intervention and comparator) according to the following categories: total AEs irrespective of seriousness, serious AEs, severe AEs (with the instrument which was used to grade the severity), death related to AEs, treatment discontinuation due to AEs, treatment interruption due to AEs.
- For medical devices, in addition provide in the main text of the JCA a descriptive report for AEs in each treatment group according to the categories ADE and DD. Also provide a descriptive report of each individual AE corresponding to these categories.Provide in the main text of the JCA report a descriptive report of specific AEs, that have been requested in the assessment scope.
- For medicinal products, results regarding relative safety (i.e., with n [%] per treatment group and effect measures including 95% CI and nominal p-values) for all the AE categories described above and in addition for AEs according to SOC and PT are included in a dedicated appendix of the JCA report. AE according to SOC and PT of any severity must only be included in the appendix if they occur with an incidence of ≥ 5 % in any treatment group. Serious and severe AE (e.g. according to Common Terminology Criteria for Adverse Events Grade ≥ 3) must be included regardless of their incidence.

5 Validity, reliability, and interpretability of outcomes measurement instruments

5.1 Definitions and general considerations

Outcomes measurement instruments mapping a predefined collection of information onto a scale measuring a specific outcome (e.g., HRQoL, objective response rate) are used in clinical studies assessing the effectiveness of treatment. Such instruments come with instructions for collecting the set of pieces of information necessary (i.e., the items). A conceptual framework outlines the interrelationships among the items and associated domains being measured by the instrument (56). A measurement model allows transformation of the responses to the items onto one scale for a unidomain concept, or a profile of multiple scales for a multidomain concept (57). For example, for PROMs, a frequent measurement model computes the sum of the codes for responses to the items of a given scale, but more complex measurement models can be involved. Outcomes are frequently measured on a continuous scale. The resulting measure can be called a score (58). Categorical scales are also used.

The same outcome (e.g., functioning) can be assessed with different instruments that use different sources of information (see Section 2.1) (8). PROMs and OBsROs (as well as some ClinROs) can generally be regarded as less objective than performance measures or some technologically assessed ClinROs, because they (implicitly or even explicitly) entail subjective appraisal by the patient, a healthcare professional or a proxy. For example, a performance measure of physical functioning can assess an objective manifestation (e.g., the number of metres a patient can walk in 6 min), while a PROM item for the same outcome can involve the patient’s judgment (e.g., asking the patient if it feels difficult to run 100 m) (59). If the patient's judgment is of explicit interest, the corresponding assessment should be conducted by the patient and not by healthcare professionals, as it is known that the latter are not always able to provide fully valid information for the patient's view (60). These differences in perspective need to be considered in formulating requests during the assessment scoping stage and in allowing MSs to assess the relevance of chosen scales submitted as evidence by HTDs. It is important to note that the use of the adjectives “objective” or “subjective” does not prejudge the quality of the measurement properties of an outcome measurement instrument. It only distinguishes instruments which involve the subjective appraisal of a person, from those which do not. Distinguishing these differences in perspective in detail and thus the actual outcome collected can require full access to the verbatim items and sometimes even literature on scale development and validation.

Last, it is frequent to categorize some outcomes measurement instruments as “generic” (i.e., they are not tailored for a specific medical condition, e.g., the MOS SF-36) or “specific” (i.e., they are tailored to be used for a specific medical condition, e.g., the European Organisation for Research and Treatment of Cancer Quality of Life

questionnaire (EORTC QLQ-C30) for assessing HRQoL specifically for patients who suffers from cancer) (14). Generic instruments can be used in various populations and can allow a comparison of the level of the targeted outcome of populations affected by different medical conditions. However, for specific medical conditions, they can have insufficient content validity (i.e., they do not capture adequately all the facets of the considered condition). In those cases, the use of disease-specific or population-specific instruments should be considered.

Summary

- Who and/or what is the main source of information (healthcare professionals, medical technology, patients) for answering items can change the perspective of measurement for the same outcome.
- An accurate understanding of what outcome is measured by an outcomes measurement instrument can be facilitated by access to the full verbatim instrument and/or instructions, as well as literature on scale development and validation.

Requirements for JCA reporting

- The HTD should provide references for the outcomes measurement instruments, including the full text of the instrument and the instructions for using it.
- The main source of information (healthcare professionals, medical technology, patients) for answering items should be described.

5.2 Validity and reliability

Appropriate use of any measurement instrument involves consideration of validity and reliability (57). However, in the context of this document, only COAs (i.e., ClinROs, PROs, PerfOS, ObsROS) are considered. As the focus is on outcomes, considerations related to the validation of diagnostic tests are beyond the scope of this guidance.

Validity refers to the extent to which an outcomes measurement instrument measures what it is supposed to measure (57). For example, if a PROM is designed to measure anxiety levels, the resulting score(s) must correlate with general anxiety symptoms across conditions but is expected to correlate to a lesser degree with other mental health outcomes like depression levels and have low to null correlation with conditions that are known to be independent of anxiety levels. Depending on the type of insufficiency (see same section below about different types of validity), instruments with an insufficient level of validity will either lead to indirectness (i.e., an estimate for an outcome that is different to the outcome of interest) (61) or bias in measurement (i.e., systematic errors). Reliability refers to the extent to which a measure produces similar results under consistent conditions (57). Measures that are reliable are accurate, reproducible, and consistent from one testing setting to another. Thus,

reliability assesses the extent to which a measure is free from measurement errors (i.e., random errors).

Development, modification, or use of an existing outcomes measurement instrument should involve patients; may involve a review of existing literature, and if needed, caregivers (where appropriate) and healthcare professionals to identify the most relevant concepts in a given disease or treatment paradigm, and to ensure the selected scale items and instructions are clear, comprehensive and consistently understood (14). For example, for development of a PROM, qualitative studies are usually conducted to identify valid items and frame corresponding questions. Then, responses to these items are collected from a sample of patients and specific statistical analyses are performed to estimate quantitative indices of validity and reliability, select the final set of items, and establish the measurement model.

Validity and reliability are multi-faceted attributes and they cannot be assessed using just one index for each; they can be categorised into several subproperties (e.g., content validity, criterion validity, structural validity, inter-rater reliability, test-retest reliability, internal consistency) (62). Moreover, they are frequently not fully assessed in a single study; investigation of these properties is an ongoing and thorough process (63). Indeed, validity and reliability are not an off/on or yes/no designation. Instead, they come in degrees. Moreover, they are not properties of the instrument itself. Rather, they speak to how scores are interpreted and used. Along with interpretability (see Section 5.3) the combination of these concepts is close to the concept of fit for purpose COA (i.e., the level to which a COA is sufficient to support its proposed use) (1,63). Depending on the quality of the measurement properties of an instrument, it implies that scores from an assessment can be appropriate for one kind of inference or use but not for another (63). De Vet et al. (57) provide a more detailed methodological background. A consensus taxonomy of the measurement properties of outcomes measurement instruments has been developed by the international Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) group (62). The same group has proposed a risk of bias tool to assess the quality of studies assessing the measurement properties of PROMs for use in systematic reviews (64), as well as a tool on the quality of studies assessing the reliability of outcomes measurement instruments (10).

A measurement on a scale is valid and reliable only if it was computed according to an evidence-based measurement model (57). In particular, if a PROM leads to a measure of a profile of scales, a unique overall score can only be computed if the measurement model allows it. In addition, analyses should never pool outcome measures collected from different sources that are considered distinct according to the use of the instrument (e.g., such as analysing the scores coming from the child and adult version of an HRQoL instrument by computing one pooled mean and standard deviation), unless a measurement model has been validated for allowing such pooling.

Instruments are usually constructed in one language first (e.g., English) and can be translated thereafter. Translation is at risk of altering the measurement properties of an instrument because of cultural differences, especially for PROMs (65). While there is no consensus on a unique method to achieve such translation, it is widely accepted that the process of translating an instrument must follow specific steps (i.e., cross-cultural adaptation) (66).

A sufficient level of validity and reliability for an outcomes measurement instrument does not ensure that an estimation of treatment effectiveness has high certainty of results, as the design, conduct and analyses of the study can lead to biases and/or random errors. Therefore, assessment of the certainty of results in a JCA report must follow the principles detailed in the other relevant HTAR guidances.

While it is currently beyond the scope of the guidance to provide an in-depth description of psychometric evaluation, the JCA report should contain enough information for each MS to be able to appraise the validity and reliability of each measurement instrument (e.g., describe the purpose and structure of an instrument, especially PROMs, and listing references, as provided by the HTD, allowing the access to the specific studies assessing the measurement properties).

Summary

- The two main properties of any outcome measurement instruments are validity and reliability.
- The assessment of the measurement properties of instruments is performed by specific studies using an appropriate design and statistical analyses.
- Validity and reliability are multi-faceted attributes that cannot be appraised in a binary manner. Depending on the quality of its measurement properties, scores from an assessment can be fit for purpose for one kind of inference and not for another.
- A taxonomy of measurement properties is proposed by the international COSMIN group.
- Translation of outcomes measurement instrument (especially PROMs) requires cross- cultural adaptation and follows specific steps.
- Studies that assess measurement properties of an instrument should be independent studies of the ones that are submitted as evidence by an HTD for answering the assessment scope of a JCA.

Points of attention for the assessment scoping process

- Carefully consider the quality (measurement properties, purpose) of any specific instrument requested to measure an outcome.

Requirements for JCA reporting

- Provide a short and appropriate description of the purpose and structure of an instrument, especially PROMs (number of scales, definition of the outcome measured by each scale, number of items per scale).
- References, as provided by the HTD, allowing the access to the specific studies assessing the measurement properties (and measurement model) of the instruments that are used.

5.3 Interpretability

Interpretability can be defined as “the degree to which one can assign qualitative meaning – that is, clinical or commonly understood connotation – to an instrument’s quantitative scores or change in scores” (62). Quantitative measures are usually expressed on a continuous or discrete scale with arbitrary boundaries (e.g., a score from 0 to 100) with, for a given value, no particular meaning attached to it. Thus, to enhance the interpretability of the results, at least one value on the scale has to be linked to a specific meaning regarding treatment effectiveness.

Enhancing the interpretability can be done by classifying patients into categories defined by relevant thresholds. For example, using the Clinical Disease Activity Index, patients can be categorized into three groups: active disease (when the score is >10), low disease activity (when the score lies between >2.8 and ≤10), and remission (when the score is ≤ 2.8) (67). Here, relative treatment effectiveness can be expressed by a difference in the proportion of patients who have switched from categories (and/or by using an effect measure such as a risk ratio). This expression of treatment effectiveness can be used to enhance interpretability.

In general, a threshold, called a responder definition, can be used to classify whether or not a patient has experienced an improvement or a deterioration of his or her condition. This can be done either by assessing whether or not a patient reached a pre-specified level of success, or by assessing whether the change in scores is as least equal to a pre-specified threshold (12). For example, for HRQoL related to mental health, a responder definition could be defined as “an improvement in MOS SF-36 mental health domain of at least 10 points”. According to this definition, patients would be classified into two categories: those who have experienced a sufficient improvement (and therefore considered as having experienced a sufficient treatment benefit on the given outcome), and those who have not. There exist many methods for estimating such responder definition for a given scale, which are partly subject of scientific debate and are accompanied by different terminology. Most of the methods are based on linking an observed estimation of the change in scores of the scale of interest in a sample of patients of the population of interest, to another phenomenon that can come from various perspectives (60). For example, using data from a cohort of patients with at least two time points, observed change in scores can be linked to observed results

in other medical outcomes such as disease severity, symptoms, prognosis, functional impact (e.g., a minimum change in score associated with a specific gain in functioning) or a global impression of change from a healthcare professional.

The patient's perspective is frequently used by linking a change in score to the subjective meaning of what is a relevant change according to patients. This approach is called the minimal important difference (MID) and can be defined as the minimal change in score perceived as an improvement or deterioration by the patient (68–70). This is also frequently called the minimal clinically important difference (MCID) (68), or meaningful change. Although the term MID can be used to describe a threshold to interpret between-group differences in scores (e.g., difference in mean change from baseline), we use it within this guidance is to refer to a threshold for interpreting within-patient change over time. Hundreds of clinical studies have been performed to propose plausible MID values for hundreds of PROMs (71). Although this approach was initially developed for PROMs, it can be useful for other outcomes measurement instruments.

The methods that are considered the most appropriate for estimating MIDs are anchor-based methods, as they explicitly link a change in score to the patient's perception (69,72). A change in score is linked to the response for a unique item: a patient global rating of change (PGRC) or patient global impression of change (PGIC). A PGRC is an overall assessment of a change compared to baseline performed by the patient. For instance, a PGRC can be phrased as follows: "Since the beginning of your treatment, overall, do you think your quality of life is now…" Proposed responses could be "a lot better", "a little better", "about the same", "a little worse" and "a lot worse".

MIDs are also frequently estimated using distribution-based methods (69). In contrast to anchor- based methods, only the overall variability in scores is used in distribution-based methods. Thus, these methods have been criticised because they do not explicitly refer to the meaning of the change for patients (69). They are still used as secondary approaches as some authors argue they can complement anchor-based methods in order to "triangulate" a plausible range where the true MID value lies (73). Two approaches are most common. The first is based on estimation of Cohen's d, which is computed by dividing the mean change in score by the standard deviation for the score at baseline. On the basis of results from experimental psychology, Cohen proposed a rule of thumb whereby d values of 0.2, 0.5 and 0.8 approximate effect sizes considered as small, moderate and large, respectively (74). Although not initially developed for responder definitions, d values of 0.2 and 0.5 are still proposed as plausible MID values (71,75). A second approach relies on disentangling changes in score from measurement errors. For example, on the basis of empirical observations, 1 standard error of measurement has been suggested as a plausible MID (76). MIDs are sometimes identified on the basis of expert opinion (69). Such MIDs are only a representation of what experts think about a change that patients consider significant. Numerous factors have been identified explaining variability in MID values, such as

dependency to the baseline level of the construct of interest, the direction of change (i.e., improvement or deterioration), the length of the period to which the PGRC refers to, and the patient population (a MID value can be different for a same outcomes measurement instrument depending of the disease and patient population assessed) (77–79).

Another possible responder definition, albeit less common, is the concept of patient acceptable symptomatic state (PASS), mostly used in rheumatology (80). Instead of focusing on the change in score that is perceived as beneficial by patients, the idea is to find the minimum score above which patients consider their health state as acceptable.

A graphical display for each treatment group of the change in score using a cumulative distribution function (estimated as the cumulative proportion of patients above a threshold for the change in score) is frequently recommended to enhance the interpretability (69). This allows estimation of the difference in proportion of patients who experienced a change in score at least as large as any threshold that can be defined for the change in score continuum (e.g., for multiple plausible MID values).

Lastly, uncategorised data must always be presented, i.e., outcomes conceived as continuous phenomena (e.g., HRQoL) should always be assessed by using measures and methods that are consistent with this continuous property (e.g., change in scores over time). Analysis on the categorical scale (i.e., using a responder definition) could complement the analysis on the continuous scale, and vice-versa. Nonetheless, to avoid the risk of data dredging and inflated type-1-error-rate, one measure of treatment effect should be pre-specified as a primary analysis in the protocol and the statistical analysis plan of the corresponding study. Elements regarding pre-specification and control of type-1-error must be reported in the JCA report according to the rules defined in the HTAR "*Guidance on reporting requirements for multiplicity issues, subgroup, sensitivity and post-hoc analyses in joint clinical assessments"*. MSs will consider presentation of results (i.e., according to the original scale and/or according to a responder definition) the way they deem appropriate for their national decision-making process.

Summary

- To enhance interpretability, a responder definition that classifies whether or not a not a patient has experienced an improvement or a deterioration is useful.
- A responder definition can be derived from numerous perspectives.
- Outcomes can be analysed with corresponding summary measures and effect measures to complement the analysis on the continuous scale. As a responder definition leads to discretisation of variables initially measured on a continuous

scale, categorical response-scores, however, should complement continuous scores but not replace them.

Requirements for JCA reporting

- The characteristics of the scale on which outcomes are measured (continuous, discrete or qualitative; boundaries; unit of measurement, if any; labels for the categories; direction of interpretation).
- The responder definition, if proposed, and as provided by the HTD (rationale and methods for estimation, perspective, rule for classifying patients, information on pre-specification of responder definition).
- References, as provided by the HTD, to allow full access to the bibliography justifying the responder definitions used.
- The measure of an outcome that was pre-specified as part of the primary analysis for each outcome measure (e.g., on a continuous or categorical scale).
- Along with results expressed according to a responder definition (summary measure, effect measure), also report results expressed using the original quantitative scale.
- Results expressed via a graphical representation such as a cumulative distribution function are highly encouraged.

6 References

1. US Department of Health and Human Services, US Food and Drug Administration, Center for Drugs Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments. 2022;57.

2. International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use. Addendum on estimands and sensitivity analysis in clinicalt trials to the guideline on statistical principles for clinical trials. E9(R1). 2019.

3. McLeod C, Norman R, Litton E, Saville BR, Webb S, Snelling TL. Choosing primary endpoints for clinical trials of health care interventions. Contemp Clin Trials Commun. 2019;16:100486.

4. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. US Food and Drug Administration - US National Institutes of Health; 2021.

5. US National Cancer Institute. Definition of an endpoint [Internet]. 2022. Available on: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/endpoint

6. Higgins JPT, Li T, Deeks JJ. Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022). Cochrane. 2022.

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

8. Mayo NE, Figueiredo S, Ahmed S, Bartlett SJ. Montreal Accord on Patient-Reported Outcomes (PROs) use series – Paper 2: terminology proposed to measure what matters in health. J Clin Epidemiol. 2017;89:119-24.

9. Walton MK, Powers JH, Hobart J, Patrick D, Marquis P, Vamvakas S, et al. Clinical Outcome Assessments: Conceptual Foundation—Report of the ISPOR Clinical Outcomes Assessment – Emerging Good Practices for Outcomes Research Task Force. Value Health. 2015;18(6):741-52.

10. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol. 2020;20(1):293.

11. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther. 2005;7(4):R796.

12. US Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009.

13. European Medicines Agency. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. 2006.

14. Fayers PM, Machin D. Quality of life: the assessment, analysis, and interpretation of patient-reported outcomes. 2nd ed. Chichester ; Hoboken, NJ: J. Wiley; 2007.

15. EuroQol Research Foundation. EQ-5D-3L User Guide [Internet]. 2018. Available on: https://euroqol.org/publications/user-guides

16. Richardson J, McKie J, Barriola E. Multi attribute utility instruments and their use. In: Encyclopedia of health economics. Elsevier Science. San Diego: A.J Culyer; 2014. p. 341-57.

17. Huhn S, Axt M, Gunga HC, Maggioni MA, Munga S, Obor D, et al. The Impact of Wearable Technologies in Health Research: Scoping Review. JMIR MHealth UHealth. 2022;10(1):e34384.

18. Epstein AM. The Outcomes Movement — Will It Get Us Where We Want to Go? N Engl J Med. 1990;323(4):266-70.

19. Barr JT. The outcomes movement and health status measures. J Allied Health. 1995;24(1):13-28.

20. Patient-Centered Outcomes Research Institute. Patient-Centered Outcomes Research [Internet]. 2013 [Accessed 3 apr 2024]. Available on: https://www.pcori.org/research/about-our-research/research-we-support/establishing-definition-patient-centered-outcomes-research/patient-centered-outcomes-research

21. Comet initiative. COMET initiative. Core Outcome Measures in Effectiveness Trials [Internet]. 2022. Available on: https://www.comet-initiative.org/

22. International Consortium for Health Outcomes Measurement. ICHOM - Patient-Centered Outcome Measures [Internet]. 2022. Available on: https://www.ichom.org/patient-centered-outcome-measures/

23. World Health Organization, éditeur. International classification of functioning, disability and health: ICF. Geneva: World Health Organization; 2001. 299 p.

24. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA J Am Med Assoc. 1995;273(1):59-65.

25. Omeract. OMERACT. Outcome Measures in Rheumatology. [Internet]. 2022. Available on: https://omeract.org/

26. Comet initiative. COMET initiative database. [Internet]. 2022. Available on: https://www.comet-initiative.org/studies

27. Kirkham JJ, Davis K, Altman DG, Blazeby JM, Clarke M, Tunis S, et al. Core Outcome Set-STAndards for Development: The COS-STAD recommendations. PLOS Med. 2017;14(11):e1002447.

28. Kirkham JJ, Gorst S, Altman DG, Blazeby JM, Clarke M, Devane D, et al. Core Outcome Set–STAndards for Reporting: The COS-STAR Statement. PLOS Med. 2016;13(10):e1002148.

29. Ramsey I, Eckert M, Hutchinson AD, Marker J, Corsini N. Core outcome sets in cancer and their approaches to identifying and selecting patient-reported outcome measures: a systematic review. J Patient-Rep Outcomes. 2020;4(1):77.

30. Kirkham JJ, Clarke M, Williamson PR. A methodological approach for assessing the uptake of core outcome sets using ClinicalTrials.gov: findings from a review of randomised controlled trials of rheumatoid arthritis. BMJ. 2017;j2262.

31. Williamson PR, Barrington H, Blazeby JM, Clarke M, Gargon E, Gorst S, et al. Review finds core outcome set uptake in new studies and systematic reviews needs improvement. J Clin Epidemiol. 2022;150:154-64.

32. International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH Harmonised Tripartite Guideline. Statistical Principles for clinical trials E9. 1998.

33. Atkinson A, Colburn W, DeGruttola V, DeMets D, Downing G, Hoth D, et al. Biomarkers Definitions Working Group. Clin Pharmacol Ther. 2001;69:89-95.

34. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? Jama. 1999;282(8):790-5.

35. Haslam A, Hey SP, Gill J, Prasad V. A systematic review of trial-level meta-analyses measuring the strength of association between surrogate end-points and overall survival in oncology. Eur J Cancer. 2019;106:196-211.

36. Schuster Bruce C, Brhlikova P, Heath J, McGettigan P. The use of validated and nonvalidated surrogate endpoints in two European Medicines Agency expedited approval pathways: A cross-sectional study of products authorised 2011–2018. Kesselheim AS, éditeur. PLOS Med. 2019;16(9):e1002873.

37. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Use of surrogate end points in healthcare policy: a proposal for adoption of a validation framework. Nat Rev Drug Discov. 2016;15(7):516-516.

38. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125(7):605-13.

39. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015;175(8):1389.

40. Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Stat Methods Med Res. 2010;19(3):205-36.

41. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat. 2006;5(3):173-86.

42. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward. Value Health. 2017;20(3):487-95.

43. Grigore B, Ciani O, Dams F, Federici C, de Groot S, Möllenkamp M, et al. Surrogate endpoints in health technology assessment: an international review of methodological guidelines. Pharmacoeconomics. 2020;38(10):1055-70.

44. Bujkiewicz S, Achana F, Papanikos T, Riley RD, Abrams KR. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints [Internet]. 2019. Available on: http://www.nicedsu.org.uk.

45. Pharmaceutical Benefits Advisory Committee. Australian Government, Department of Health and Ageing. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. 2016.

46. Cordoba G, Schwartz L, Woloshin S, Bae H, Gotzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ. 2010;341(aug18 3):c3920-c3920.

47. Euopean Medicines Agency. Gudeline on multiplicity issues in clinical trials. Draft. 2017.

48. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389-430.

49. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc. 1999;94(446):496-509.

50. European Medical Device Coordination Group. Safety reporting in clinical investigations of medical devices under the Regulation (EU) 2017/745 [Internet]. 2022 [Accessed 5 apr 2024]. Available on: https://health.ec.europa.eu/system/files/2022-11/md_mdcg_2020-10-1_guidance_safety_reporting_en.pdf

51. European Medicines Agency. Serious adverse reaction [Internet]. [Accessed 5 apr 2024]. Available on: https://www.ema.europa.eu/en/glossary/serious-adverse-reaction

52. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. MedDRA - the Medical Dictionary for Regulatory Activities [Internet]. 2022. Available on: http://www.meddra.org/

53. US National Cancer Institute. Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE®) [Internet]. 2023. Available on: https://healthcaredelivery.cancer.gov/pro-ctcae/

54. US National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE) [Internet]. 2022. Available on: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm

55. World Health Organization. Cancer treatment: WHO recommendations for grading of acute and sub acute toxicity. Cancer. 1981;47:207-14.

56. Rothman ML, Beltran P, Cappelleri JC, Lipscomb J, Teschendorf B. Patient-Reported Outcomes: Conceptual Issues. Value Health. 2007;10:S66-75.

57. Vet HCW de, Terwee CB, Mokkink LB, Knol DL, éditeurs. Measurement in medicine: a practical guide. Cambridge: Cambridge Univ. Press; 2011. 338 p. (Practical guides to biostatistics and epidemiology).

58. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994. 752 p. (McGraw-Hill series in psychology).

59. Schwartz CE, Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. Health Qual Life Outcomes. 2004;2(1):16.

60. Sneeuw KCA, Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease. J Clin Epidemiol. 2002;55(11):1130-43.

61. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol. 2011;64(12):1303-10.

62. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.

63. Edwards MC, Slagle A, Rubright JD, Wirth RJ. Fit for purpose and modern validity theory in clinical outcomes assessment. Qual Life Res. 2018;27(7):1711-20.

64. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;27(5):1171-9.

65. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures: Spine. 2000;25(24):3186-91.

66. Epstein J, Santo RM, Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. J Clin Epidemiol. 2015;68(4):435-41.

67. Olivieri M, Gerardi MC, Spinelli FR, Di Franco M. A Focus on the Diagnosis of Early Rheumatoid Arthritis. Int J Clin Med. 2012;03(07):650-4.

68. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol. 2002;14(2):109-14.

69. the Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL), Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res. 2012;22(3):475-83.

70. Vanier A, Sébille V, Blanchin M, Hardouin JB. The minimal perceived change: a formal model of the responder definition according to the patient's meaning of change for patient-reported outcome data analysis and interpretation. BMC Med Res Methodol. 2021;21(1):128.

71. Vanier A, Woaye-Hune P, Toscano A, Sébille V, Hardouin JB. What are all the proposed methods to estimate the Minimal Clinically Important Difference of a Patient-Reported Outcome Measure? A systematic review. In: Philadelphia, 18-21 Oct, 24th annual conference of International SOciety of Quality Of Life. 2017.

72. Terluin B, Eekhout I, Terwee CB. Improved adjusted minimal important change took reliability of transition ratings into account. J Clin Epidemiol. 2022;148:48-53.

73. Leidy NK, Wyrwich KW. Bridging the Gap: Using Triangulation Methodology to Estimate Minimal Clinically Important Differences (MCIDs). COPD J Chronic Obstr Pulm Dis. 2005;2(1):157-65.

74. Cohen J. Statistical power analysis for the behavioral sciences. 2. ed., reprint. New York, NY: Psychology Press; 2009. 567 p.

75. Norman GR, Sloan JA, Wyrwich KW. The truly remarkable universality of half a standard deviation: confirmation through another look. Expert Rev Pharmacoecon Outcomes Res. 2004;4(5):581-5.

76. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999;52(9):861-73.

77. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 2010;63(5):524-34.

78. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? PharmacoEconomics. 2000;18(5):419-23.

79. Woaye-Hune P, Hardouin JB, Lehur PA, Meurette G, Vanier A. Practical issues encountered while determining Minimal Clinically Important Difference in Patient-Reported Outcomes. Health Qual Life Outcomes. 2020;18(1):156.

80. Tubach F, Wells GA, Ravaud P, Dougados M. Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues. J Rheumatol. 2005;32(10):2025-9.

81. Delgado A, Guddati AK. Clinical endpoints in oncology - a primer. Am J Cancer Res. 2021;11(4):1121-31.

82. Hernandez-Villafuerte K, Fischer A, Latimer N. Challenges and methodologies in using progression free survival as a surrogate for overall survival in oncology. Int J Technol Assess Health Care. 2018;34(3):300-16.

83. Hess LM, Brnabic A, Mason O, Lee P, Barker S. Relationship between Progression-free Survival and Overall Survival in Randomized Clinical Trials of Targeted and Biologic Agents in Oncology. J Cancer. 2019;10(16):3717-27.

84. European Medicines Agency. Guideline of the clinical evaluation of anticancer medicinal products. [Internet]. 2021. Available on: https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-evaluation-anticancer-medicinal-products-man-revision-6_en.pdf

85. Gyawali B, Hey SP, Kesselheim AS. Evaluating the evidence behind the surrogate measures included in the FDA's table of surrogate endpoints as supporting approval of cancer drugs. EClinicalMedicine. avr 2020;21:100332.

86. RECIST working group. RECIST. The official site of the RECIST Working Group. [Internet]. 2022. Available on: https://recist.eortc.org/

Appendix A Specific definitions of outcomes usually used in oncology

Overall Survival (**OS**) is regarded as the final patient-centred outcome in oncology (81). Improvement in OS clearly demonstrates a clinical benefit which is meaningful to the patients. However, measuring OS often requires a large number of patients and long follow-ups. Long-term OS-data for the technology under assessment may be influenced by treatment given in further steps, sequential use of other agents, or even cross-over treatments, making it difficult to attribute the OS result to a specific medical intervention.

In oncology, the most frequently reported disease related outcomes are **progression free survival (PFS)** as a surrogate for OS, **event free survival (EFS)**, or **disease-free survival (DFS)**.

Since the therapy of cancer disease is often sequential and choice of therapy varies with the type of tumour and stage, there are some outcomes that are typically used in particular settings to capture the effect at a given time-point. Some of those outcomes are presented below.

Progression free survival (PFS) is defined as the time from randomization until first evidence of disease progression or death. PFS is measured by censoring patients who are still alive without progression at the time of evaluation or those who were lost to follow up and thus the data are available earlier, within the timeframe of the trial. PFS is a frequently used surrogate outcome in oncology since it can be reported within a shorter time of follow-up and the results may be obtained with a lower number of patients. However, the correlation between PFS and OS seems to differ across cancer types and therapy lines (82). The correlation between PFS and OS is not always confirmed by the final results, especially in studies of targeted therapy or immunologic agents (83,84).

Time to progression (TTP) is defined as the time from randomization until first evidence of disease progression. Since PFS and TTP are similar, it is important for studies to clarify what is meant by evidence of disease progression. Clear definition of TTP is important to avoid confusion when comparing results from different studies (81).

Disease free survival (DFS) is defined as the time from randomization until evidence of disease recurrence. DFS is often used as a surrogate outcome for therapies in adjuvant setting. DFS has been used as a surrogate outcome for OS in clinical trials for stage III colon cancer, in an adjuvant setting in lung cancer, and in breast cancer. The definition of ‘disease-free interval’ is not always clear and the validity of an incidental finding of cancer regardless of symptoms has been questioned. It is strongly recommended that the recurrence is defined when using DFS as an outcome (81).

Event-free survival (EFS) is defined as the time from randomization to an event which may include disease progression, discontinuation of the treatment for any reason, or death. According to Gyawali at al., while EFS and DFS used to be interchangeable, the patient is not technically “disease-free” at the time of randomization in a neoadjuvant setting; EFS is an outcome mainly used for neoadjuvant settings while DFS is usually applied in adjuvant settings (85). If EFS is used as a surrogate outcome for OS it needs to be validated for each unique tumour type, treatment, and stage of disease.

Objective response rate (ORR) is a measure of antitumor activity and defines a proportion of patients that respond either partially or fully to the therapy according to a predefined set of response criteria. Response Evaluation Criteria in Solid Tumors (RECIST) is the most commonly used set of evaluation criteria. RECIST provides a simple and pragmatic methodology to evaluate the activity and efficacy of new cancer therapeutics in solid tumours, using validated and consistent criteria to assess changes in tumour burden (86).