eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

EUnetHTA 21

EUnetHTA 21 – Individual Practical Guideline Document

D4.4 – OUTCOMES (ENDPOINTS)

Version 0.3, 29/09/2022

Template version 1.0, 03/03/2022

DOCUMENT HISTORY AND CONTRIBUTORS

Version	Date	Description
V0.1	22/06/2022	First draft for CSCQ and NC-HTAb review
V0.2	24/08/2022	Second draft for CSCQ and NC-HTAb review
V0.3	29/09/2022	Third draft for public consultation

Disclaimer

This Practical Guideline was produced under the Third EU Health Programme through a service contract with the European Health and Digital Executive Agency (HaDEA) acting under mandate from the European Commission. The information and views set out in this Practical Guideline are those of the author(s) and do not necessarily reflect the official opinion of the Commission/Executive Agency. The Commission/Executive Agency do not guarantee the accuracy of the data included in this study. Neither the Commission/Executive Agency nor any person acting on the Commission's/Executive Agency's behalf may be held responsible for the use which may be made of the information contained herein.

Participants

Hands-on Group	Gemeinsamer Bundesausschuss [G-BA], Germany Haute Autorité de Santé [HAS], France National Authority of Medicines and Health Products [INFARMED], Portugal National Centre for Pharmacoeconomics [NCPE], Ireland Norwegian Medicines Agency [NOMA], Norway
Project Management	Zorginstituut Nederland, [ZIN], The Netherlands
CSCQ	Agencia Española de Medicamentos y Productos Sanitarios [AEMPS], Spain
CEB	Austrian Institute for Health Technology Assessment [AIHTA], Austria Belgian Health Care Knowledge Centre [KCE], Belgium Gemeinsamer Bundesausschuss [G-BA], Germany Haute Autorité de Santé [HAS], France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen [IQWIG], Germany Italian Medicines Agency [AIFA], Italy National Authority of Medicines and Health Products [INFARMED], Portugal National Centre for Pharmacoeconomics [NCPE], Ireland National Institute of Pharmacy and Nutrition [NIPN], Hungary Norwegian Medicines Agency [NOMA], Norway The Dental and Pharmaceutical Benefits Agency [TLV], Sweden Zorginstituut Nederland [ZIN], The Netherlands

The work in EUnetHTA 21 is a collaborative effort. While the agencies in the Hands-on Group will be actively writing the deliverable, the entire EUnetHTA 21 consortium is involved in its production throughout various stages. This means that the Committee for Scientific Consistency and Quality (CSCQ) will review and discuss several drafts of the deliverable before validation. The Consortium Executive Board (CEB) will then endorse the final deliverable before publication.

Copyright

TABLE OF CONTENTS

DOCUMENT HISTORY AND CONTRIBUTORS **2**
TABLE OF CONTENTS **3**
1 INTRODUCTION **5**
1.1 Problem statement, scope, and objectives *5*
1.2 Relevant articles in Regulation (EU) 2021/2282 ..... *5*
2 DEFINITIONS AND GENERAL CONSIDERATIONS **6**
2.1 Definitions *6*
2.2 General considerations *7*
3 CLINICAL RELEVANCE **8**
3.1 Definition of patient-centred outcomes *8*
3.2 Determinant outcomes for specific therapeutic areas *9*
3.3 Surrogate outcomes *10*
4 SAFETY **12**
4.1 Terminology for JCA *12*
4.2 Safety: overall and specific adverse events *12*
4.3 Information to be reported for safety outcomes *12*
5 VALIDITY, RELIABILITY, AND INTERPRETABILITY OF SCALES **13**
5.1 Definitions and general considerations *13*
5.2 Validity and reliability of scales *14*
5.3 Interpretability of scales *15*
6 REFERENCES **17**
APPENDIX A: SPECIFIC DEFINITIONS OF OUTCOMES USUALLY USED IN ONCOLOGY **21**

73 # LIST OF ACRONYMS - INITIALISMS

AE	Adverse event
CEB	Consortium Executive Board
COMET	Core Outcome Measures in Effectiveness Trials
COS	Core outcome set
COSMIN	Consensus-based Standards for the Selection of Health Measurement Instruments
CSCQ	Committee for Scientific Consistency and Quality
CTCAE	Common Terminology Criteria for Adverse Events
DAS 28	Disease Activity Score 28
DFS	Disease-free survival
EFS	Event-free survival
EMA	European Medicines Agency
EU	European Union
EUnetHTA	European Network for Health Technology Assessment
HaDEA	European Health and Digital Executive Agency
HRQoL	Health-related quality of life
HTA	Health technology assessment
HTAb	HTA body
HTAR	HTA Regulation (EU) 2021/2282
HTD	Health technology developer
ICD	International Classification of Diseases
JCA	Joint clinical assessment
MedDRA	Medical Dictionary for Regulatory Activities
MCID	Minimal clinically important difference
MID	Minimal important difference
MOS SF-36	Medical Outcome Study Short Form 36
MS	Member state
ORR	Objective response rate
OS	Overall survival
PASS	Patient-acceptable symptomatic state
PFS	Progression-free survival
PGRC	Patient global rating of change
PICO	Population, Intervention, Comparator, Outcome
PRO	Patient-reported outcome
PROM	Patient-reported outcome measure
SAE	Serious adverse event
SUSAR	Suspected unexpected serious adverse reaction
TTP	Time to progression
WHO	World Health Organization
WHO-ART	World Health Organization adverse reaction terminology

74

1 INTRODUCTION

1.1 Problem statement, scope and objectives

Clinical outcome assessment is a key component of health technology assessment (HTA). It is the measure of the clinical benefit of the targeted treatment on patient-centred outcomes (see the definition in Section 3.1). In the context of joint clinical assessment (JCA), outcomes are relevant in two different steps. The first step is during the scoping process, when member states (MS) are expected to request their needs in terms of health outcomes (HTA Regulation (EU) 2021/2282 (HTAR), Article 8(6)) when defining PICO (Population, Intervention, Comparator, Outcome) questions. Defining relevant outcomes is a key component of this process. The second step is when assessors and co-assessors produce the JCA report based on the dossier submitted by the health technology developer (HTD) and the PICO question(s) previously defined for the health technology under assessment. While MS are required to give due consideration to the JCA reports published (Article 13 (1)), the clinical relevance or interpretation of the measure of relative effectiveness may differ between MS when drawing conclusions regarding the clinical added value of a treatment at a national level. Therefore, appropriate reporting of the methodological and statistical elements and results of the analyses of the outcomes requested is essential (Article 9(1)).

According to the HTAR (Recital (28)), health outcomes should not be ranked and the assessment scope should reflect MS needs. Neither the HTAR nor EUnetHTA 21 practical guideline D4.2 (Scoping process) proposes criteria to be used by MS when defining health outcomes. However, health outcomes requested during the assessment scoping stage have an important impact on the result of a JCA. Indeed, the relative effectiveness of the health technology as assessed in terms of health outcomes will be described as required in the scoping process on the basis of the predefined parameters. However, the conclusions that MS can draw regarding the clinical added value of a treatment can be impacted by factors such as appraisal of the validity and reliability of the measurement scales of instruments or of the relevance of intermediate or surrogate outcomes.

The objectives of this guideline are twofold. The first objective is to provide guidance for MS in defining relevant outcomes during the scoping process. The second is to help assessors and co-assessors in assessing and reporting all the necessary elements that MS need to carry out for national appraisal of the clinical added value of a health technology. Thus, all the requirements for reporting and assessment mentioned in this guideline suggest that HTDs are supposed to present the necessary elements in their submission dossiers (Article 9(3)).

In the context of JCA, outcomes cannot be dissociated from the way in which they are statistically analysed. Complementary elements related to the assessment of the certainty of results associated with outcomes of interest are provided in EUnetHTA 21 practical guideline D4.6 (Validity of clinical studies) regarding outcomes assessed in individual clinical studies, and EUnetHTA 21 methodological and practical guidelines D4.3.1 and D4.3.2 (Direct and indirect comparisons) regarding outcomes assessed in evidence synthesis studies. EUnetHTA 21 practical guideline D4.5 (Applicability of evidence: practical guideline on multiplicity, subgroup, sensitivity, and post-hoc analyses) provides complementary details on specific issues such as multiple hypothesis testing, subgroup, sensitivity and post hoc analyses.

For simplicity, effectiveness is the term used to describe efficacy or effectiveness throughout the rest of this document. Furthermore, treatment, intervention and health technology are all terms used for any health technology that can be assessed.

1.2 Relevant articles in Regulation (EU) 2021/2282

Articles from Regulation (EU) 2021/2282 directly relevant to the content of this practical guideline are:

- Recital 2,
- Recital 28,
- Article 8: Initiation of joint clinical assessments,

- Article 9: Joint clinical assessment reports and the dossier of the health technology developer,
- Article 13: Member States' rights and obligations.

2 DEFINITIONS AND GENERAL CONSIDERATIONS

2.1 Definitions

"**Outcome**" is any concept that can be used for estimating treatment effectiveness, such as mortality, remission, health-related quality of life (HRQoL), symptoms and safety. Outcomes are distinct from the way in which they are measured. The "**measure of an outcome**" defines in an accurate way how the outcome is assessed (including use of a specific instrument; see Section 5). For instance, if the outcome is mortality, the measure of the outcome could be "proportion of deaths 28 days after inclusion". If the outcome is pain, the measure of the outcome could be "change in the level of pain on a patient-reported visual analogue scale of 100 mm at 24 hours after initiation of the treatment". It is sometimes argued in the literature that this difference between an outcome and its measure is the difference between outcome (as the concept) and "**endpoint**" (as the measure) [1,2]. However, there is no internationally agreed definition. The two terms are frequently used interchangeably [3]. In this guideline, we only use the terms outcome and measure of an outcome. Lastly, **effect measures** are the statistics that are used to express the effectiveness of a treatment [4]. 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". Thus, effect measures are primarily understood as a comparison of the measure of outcomes between two interventions groups. 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 the absolute effect or a within-group change [5].

It can also be useful to classify outcomes according to the main source of information via which they are collected [6]. Identification of adequate source(s) of information can help in defining relevant outcomes during the assessment scoping stage.

First, the main source of information can come from activity by healthcare professionals. In general, the resulting outcomes can be called **clinician-reported outcomes** [7]. These can be divided into two subcategories. **Clinically reported outcomes** are assessed by healthcare professionals during clinical examination of a patient and involve clinical judgments of patients' observable signs, behaviours or other physical manifestations. **Technologically assessed outcomes** require the use of technology such as laboratory tests or medical imaging.

Second, the main source of information can be the patients. **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*" [8]. They are measured by **patient-reported outcomes measures** (PROMs), mostly in the form of self-administered questionnaires. The PRO concept is sometimes equated to HRQoL. However, HRQoL is only a subset of the outcomes that can be measured using PROMs. Some PROMS measure health status (for instance, the EQ-5D instrument measures health status as a combination of five broad concepts [9]). Other outcomes such as symptoms, fatigue, pain, anxiety, depression, functioning, impairment, disability and impact on daily living can be assessed using PROMs. Sometimes, instruments that would normally be answered by patients are instead reported by an observer with shared experience. An example would be a caregiver if the patient is unable to answer the items. These cases are referred to as PROs answered by "**proxies**". This distinction is important because the person who is assessing the outcome can impact the accuracy of the information.

Third, there are other specific cases. **Performance outcomes** are close to clinically reported outcomes but require active patient involvement, (e.g., tests of walking, cognitive tests). There is also increasing use of **patient-generated health data** such as outcomes using **connected digital health technologies** (e.g., monitoring devices for medical adherence). These devices can allow an automated measure of outcomes in settings other than the usual visits for clinical studies, such as in home settings [2,10]. Use of such technologies could lead to benefits such as better compliance or expansion of participation in

clinical studies for populations with limited access to clinical facilities [11,12]. However, use of such technologies risks limiting the eligibility for clinical studies to participants with sufficient digital literacy or sufficient access to technologies such as an efficient internet connection [12].

Lastly, categories for classifying outcomes are not mutually exclusive, as some instruments require the collection of elements from multiple sources. For example, the Disease Activity Score 28 (DAS 28) for rheumatoid arthritis requires clinical, technological and patient-reported elements [13].

2.2 General considerations

During the assessment scoping stage for JCA, the definition of outcomes requested by MS should be as appropriate as possible, as this can impact assessment of the results submitted by a HTD in a JCA report. Therefore, general guidance can be useful for formulating outcomes that are the most relevant during the assessment scoping stage.

Defining an outcome at the broadest level (e.g., HRQoL without further specifications) maximises the opportunity to obtain a result. 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 adequacy of the measure of the outcome provided by the HTD therefore needs to be appraised by the MS on the basis of the elements reported within the JCA. 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 a MS, but with a higher risk of not obtaining 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]*". 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, but also at the 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 could also be to formulate a request as such: "*[Outcome of interest] measured preferably at [insert timing of assessment]*".

Lastly, a more detailed level would be to request a specific effect measure. While this practical guideline does not endorse any criteria to be filled by MS when requesting health outcomes, 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. For example, hazard ratios estimated using a Cox model require that the proportional hazards assumption approximately holds. If not, hazard ratios are not valid estimates and another effect measure should be used, such as the restricted mean survival time. 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 an 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]*".

Summary

- Outcomes are concepts for estimating treatment effectiveness.
- The measure of an outcome defines accurately how the outcome is assessed.
- Effect measure are primarily statistics used to compare the measure of outcomes between two intervention groups. Other statistics can be used for other purposes (absolute effect, within-group change).

Points of attention for the assessment scoping process

- Proposing an outcome with a more or less specific definition (e.g., as an outcome only, or by specifying a measure, time point for assessment and/or by specifying an effect measure) can impact the reporting of results in a JCA.
- If an MS wants to specify a measure of an outcome, the wording should follow this template: *"[Outcome of interest] measured preferably as [insert measure]*".
- If an MS wants to specify a time point for assessment, the wording should follow this template: "*[Outcome of interest] measured preferably at [insert timing of assessment]*".
- Effect measures should not be specified by MS. The HTD is responsible for presenting results using appropriate effect measures in accordance with good clinical and statistical practice.
- If an 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

- Accurate definition (concept, main source of information, measure, timing, effect measure) of any reported outcome.

3 CLINICAL RELEVANCE

3.1 Definition of patient-centred outcomes

Several outcomes are considered adequate in confirmatory clinical trials 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 and biomarkers (see the definitions in Section 3.2). In general, long-term or final outcomes (i.e., the occurrence of an irreversible event of primary interest such as death) are preferred in HTA. In terms of the relevance of different outcomes for PICO questions or JCA, the research question and the disease and treatment investigated will be most important. The acceptability of an outcome is subject to MS interpretation of their relevance within their national process for decision-making and thus may differ between MS. Both the EUnetHTA collaboration and the European Medicines Agency (EMA) have published detailed guidelines on the choice of outcomes in trials and for assessment of the relative effectiveness of therapies [14,15].

Not all outcomes are considered equally important to patients. 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) [16,17]. Deciding what is a patient-centred outcome for the PICO question for a particular therapy should ideally be done in close collaboration with patients and healthcare professionals who either live with the medical condition and/or are knowledgeable about the condition. However, the final decision is up to the 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.

Classifications such as the International Classification of Functioning, Disability and Health of the World Health Organization (WHO) [18], the Wilson and Cleary biopsychosocial model [19] and the Montreal Accord on Patient-Reported Outcomes [6] can provide further information on outcomes that can be assessed in healthcare.

The EUnetHTA guideline recommends that outcomes relevant for HTA should be long-term or final [14]. **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** (OS) or mortality rates/survival rates for a given period (e.g., 1-year mortality or 5-year mortality). For diseases with expected long-term survival, it might be impossible to obtain mature mortality data from clinical trials at the time at which the JCA report is generated. If it is not feasible to measure a final outcome , 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 final outcome [14]. Outcome measurements 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 clinical evaluation measurements and scales may be used to capture relevant information about patients' health status and the disease response to a given therapy. It is crucial that the "event" is well defined and that only validated tools for measurement are used. Time points for assessment of different outcomes and the frequency of these assessments may be of importance for the number of results reported.

Points of attention for the assessment scoping process

- The EUnetHTA guidelines recommend that 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 final outcome.

3.2 Determinant outcomes for specific therapeutic areas

Efforts are being conducted to identify a standardised set of outcomes that should be measured and reported, as a minimum, in all clinical trials in specific areas of health or healthcare, defined as a **core outcome set** (COS) [20]. Initially, these initiatives were in medical fields such as rheumatology (see the OMERACT initiative [21]) in which disease manifestation is mostly chronic and heterogeneous and affects more than one organ. In these medical settings, defining a set of the most relevant outcomes is highly challenging, which is why there is a need to define COS at an international level. These initiatives have subsequently been applied in various medical fields and healthcare settings [20]. The relevance of COS is highlighted when facing prevalent conditions such as cancer and multimorbidity. The **COMET** (Core Outcome Measures in Effectiveness Trials) initiative maintains a COS database [22].

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 contributing to less heterogeneity in outcome reporting in individual clinical studies, COS use may facilitate the conduct of meta-analyses.

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 instruments to assess core PROs [23]. The conclusion was that there is a lack of recommendations on how to measure core PROs, such that efforts to standardise outcome assessment via the development of COS may be undermined. It was suggested that to optimise COS usefulness and adoption, valid and reliable 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 [24]. However, a systematic review concluded that COS uptake in new studies and systematic reviews needs improvement, as uptake is still low in most research areas [25].

Even though the recommendations from well-established COS should be considered in the selection of outcomes for the assessment scoping process, if such COS are available, it should be noted that COS are not written from a HTA perspective. Therefore, generic multiattribute utility instruments should complement the use of COS.

Since cancer is the leading cause 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, it is important that this document reflects outcomes for assessing the safety and efficacy of new cancer drug therapies. Specific definitions of outcomes typically used in oncology are provided in Appendix A.

Points of attention for the assessment scoping process

- In the selection of outcomes, recommendations from well-established COS should be considered, if such COS are available.
- Generic multiattribute utility instruments should complement the use of COS.

3.3 Surrogate outcomes

General considerations

A **surrogate outcome** is an outcome that is intended to replace an outcome of interest that cannot be observed in a trial. 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 [26]. 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 **biomarker** can be defined as a characteristic that is objectively (reliably and accurately) measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to an intervention [27]. Examples include levels of cholesterol and haemoglobin A1c.

An **intermediate outcome** is an outcome such as a measure of a function or of a symptom (disease-free survival, angina frequency, exercise tolerance) but is not the ultimate outcome of the disease, such as survival or the rate of irreversible morbid events (stroke, myocardial infarction) [28].

The use of surrogate outcomes in assessment of the clinical added benefit of a health technology can be controversial since the validity of surrogate outcomes has rarely been rigorously fully established [29–32]. Only a few surrogate outcomes have been shown to be true measures of tangible clinical benefit. The guideline "Endpoints used in relative effectiveness assessment: surrogate endpoints" previously developed during EUnetHTA Joint Action 1/2 outlines the methodological issue with the use of surrogate outcomes [14].

Safety is a particularly important consideration when using surrogate outcomes. It is important to accurately capture the risk–benefit profile of an intervention. Even if surrogacy has been demonstrated for a specific efficacy outcome, unexpected side effects of that intervention may lead to an increase in mortality or other unfavourable outcomes. Therefore, safety outcomes of interest should be included at the scoping stage. Other considerations regarding safety are addressed in Section 4.

Points of attention for the assessment scoping process

A validated surrogate outcome should only be used to replace a patient-centred outcome of interest if absolutely necessary:

- If evidence for a patient-centred outcome is likely to be available, then this should be requested during the scoping process instead of surrogate outcomes such as morbidity, overall mortality and HRQoL;
- 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 might have been established by previous JCAs or in other literature on the same indication [14].

Level of evidence

As detailed in "Endpoints used in relative effectiveness assessment: surrogate endpoints" [14], appraisal of the association between the surrogate and the final outcome should take into account the level of evidence:

- **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; and establishment of correlation between effects on the surrogate outcome and the patient-centred outcome;
- **Level 2**: evidence demonstrating a consistent association between the surrogate outcome and the final patient-centred outcome (from epidemiological or observational studies);
- **Level 3**: only evidence of 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).

Association between the surrogate outcome and the patient-centred outcome

The HTD should demonstrate the strength of the association between the surrogate outcome and the patient-centred outcome and the treatment effect. This is often done via regression analysis for single studies, or meta-regression in the case of multiple studies. Ideally the association will be demonstrated at both the individual level and the trial level.

For all outcomes requested in the assessment scope, the HTD should provide data, regardless of how immature they are. The presence of surrogate outcome data, regardless of their validity, does not change this requirement. For example, if an intervention is expected to impact OS, data on OS should always be presented, even if the length of follow-up or the number of events is insufficient.

Uncertainty

A surrogate outcome may lead to greater uncertainty surrounding the benefit of the technology under assessment.

Requirements for JCA reporting

The assessor should 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 if available.
- The limitations of the use of a surrogate outcome should be explicitly explained.
- Details of any additional information required that could decrease the uncertainty surrounding this outcome.
- An indication of whether or not a patient-centred outcome is likely to be available at a later date.
- Clearly outline any remaining areas of uncertainty.

There are a number of frameworks that may be useful when assessing surrogate outcomes. These include reports by Ciani et al. [31, 33], Grigore et al. [34] and Bujkiewicz et al. [35] and guidelines on preparing a submission to the Australian Pharmaceutical Benefits Advisory Committee [36].

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 guideline is not intended to duplicate the definitions already provided for safety terminology [37]. In the context of JCA, the term "**adverse event"** (AE) must be used, and the terms "adverse reaction", "adverse drug reaction", "side effect", "serious incident", "device deficiency", "adverse device effect" and "adverse effect" should be avoided. The term "**safety**" must be used, and "tolerability" and "toxicity" should be avoided.

Requirements for JCA reporting

- Use the term "safety", and not "tolerability" or "toxicity".
- Use the term "adverse event", and not "adverse reaction", "adverse drug reaction", "side effect", "serious incident", "device deficiency", "adverse device effect" or "adverse effect".

4.2 Safety: overall and specific adverse events

During the assessment scoping stage, MS define their required safety outcomes. If **specific adverse events** are of interest for MS, they should require these explicitly (e.g., symptomatic osteonecrosis of the jaw with bisphosphonates).

When "safety" is required as an outcome in the assessment scope without further specifications, only overall safety results (i.e., all AEs combined) will be reported in the JCA report. If some specific AEs were required in the assessment scope, they will be reported in the JCA report. In cases requiring both "safety" and a specific AE, both results will be reported in the JCA report, but limited to the AEs required for the specific part.

Points of attention for the assessment scoping process

- Any need for a specific AE must be explicitly requested.
- A broad request ("safety") will not be associated with any description of a specific AE.

Requirements for JCA reporting

- Specific AEs that are requested must be reported.

4.3 Information to be reported for safety outcomes

Safety outcomes can be defined according to different terminologies. MedDRA (Medical Dictionary for Regulatory Activities) is used for interventional studies [38]. Other terminology can be used in observational studies, such as the International Classification of Diseases (ICD) [39] and the WHO Adverse Reaction Terminology (WHO-ART), although this is no longer maintained. Therefore, a JCA must describe the terminology used when reporting safety outcomes.

Safety outcomes can be graded for **severity** using different scales. CTCAE (Common Terminology Criteria for Adverse Events) is typically used for interventional studies in oncology but can also be used in nononcology trials [40]. A WHO scale has also been developed [41]. Therefore, 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** (SAE) 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.

Any **suspected unexpected serious adverse reaction** (SUSAR) should be reported, even if these are (by definition) not requested during the assessment scoping stage. These are defined as AEs assessed as being unexpected by the sponsor and/or study investigator and meeting the criteria for being

classified as serious. The term "adverse reaction" can be used as an exception in this situation for consistency with the regulatory process.

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

Causality (attributability) between a heath technology and an AE could be described by many terms and scales. There is no rationale, and a high risk of bias in unblinded studies, to only report AEs potentially related to the health technology under study. A safety outcome must always be reported irrespective of causality.

Reporting for overall safety (see above) requires grouping all AEs, without any description of specific AEs.

Requirements for JCA reporting

- Specify the terminology used for coding of AEs.
- Reporting for all AEs combined (overall safety) and specific AEs (if applicable), irrespective of seriousness, as well as SAEs.
- Reporting of severity, with the scale used.
- Reporting of discontinuation and interruption due to AEs.
- Primary reporting irrespective of causality.
- Reporting of SUSARs.

5 VALIDITY, RELIABILITY AND INTERPRETABILITY OF SCALES

5.1 Definitions and general considerations

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 treatments [42]. Such instruments come with instructions for collecting the set of pieces of information necessary (i.e., the **items**). A **measurement model** allows transformation of the responses to the items onto one scale for a unidimensional concept, or a profile of multiple scales for a multidimensional concept [42]. 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** [43]. 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) [6]. PROMs (as well as clinically reported measures) can generally be regarded as less objective than performance measures or some technological measures, because they (implicitly or even explicitly) entail subjective appraisal by the patient (or the healthcare professional). 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) [44]. If the patient's view 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 [45]. These differences in perspective need to be considered in formulating requests during the assessment scoping stage and in allowing MS to assess the relevance of chosen scales submitted as evidence by HTDs. 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.

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.
- Understanding accurately what outcome is measured by an instrument can be facilitated by access to the full verbatim instrument and/or instructions, as well as literature on scale development and validation.

Points of attention for the assessment scoping process

- Specifying the main source of information can have relevance for a given outcome.

Requirements for JCA reporting

- References, as provided by the HTD, allowing retrieval of the full verbatim measurement instrument and/or instructions.

5.2 Validity and reliability of scales

For appropriate usage, any measurement device needs to meet a sufficient level for two main properties: **validity** and **reliability** [42]. However, in the context of this document, only the instruments that are defined in the previous section are considered. As the focus here is on outcomes, considerations related to the validation of diagnostic tests or any device measuring phenomena with no prognostic value are beyond the scope of this guideline. This guideline will also not cover considerations about the measurement properties of routine clinical examination procedures, routine biological and laboratory tests (e.g., measurement of serum creatinine levels), or routine use of medical imaging (e.g., measurement of the size of a particular anatomical structure).

Validity refers to the extent to which an instrument measures what it is supposed to measure [42]. For example, if a PROM is designed to measure anxiety levels, it must not measure depression levels. Depending on the type of insufficiency, 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) [46] or **bias** in measurement (i.e., systematic errors). Reliability refers to the extent to which a measure produces similar results under consistent conditions [42]. 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).

The process of studying the measurement properties (i.e., validity and reliability) of instruments involves conducting specific surveys and (clinical) studies, which has already occurred in part during scale development. For example, for development of a PROM, patient surveys or interviews (qualitative studies) are usually conducted to identify valid items and frame corresponding questions. Responses to these items are collected from a sample of patients and specific statistical analyses are performed to select the necessary items and to establish the measurement model.

Validity and reliability are not one-dimensional properties and they cannot be assessed using just one index for each; they can be categorised into several subproperties. Moreover, they are frequently not fully assessed in a single study; investigation of these properties is an ongoing process. De Vet et al. [42] provide a more detailed methodological background. A consensus taxonomy of the psychometric properties of PROMs has been developed by the international Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) group [47].

Some facets of validity and reliability can have more or less relevance depending on the purpose of the outcome being assessed. For example, if the purpose of an instrument is to assess a multidimensional outcome (e.g., the MOS SF-36 measures HRQoL as a profile of eight dimensions), then an essential element of the validity of the instrument is its structural validity (i.e., the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the outcome to be measured) [42]. The reliability of instruments assessing clinically reported outcomes can be operator-dependent. Therefore, high inter-rater reliability is paramount (i.e., when the assessment on the same patient is performed by different well-trained professionals, it leads to the same result) [42]. PROMs are completed by patients, so high test–retest reliability has more value for these (i.e., if the assessment is performed by the same patient at two time points with identical conditions, the result is the same) [47].

A measurement on a scale is valid and reliable only if it was computed using the measurement model as validated by the authors of the instrument [42]. 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. 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 [48]. Therefore, PROM translation follows specific rules (**transcultural adaptation** [48]), notably including a specific validation phase after translation.

A sufficient level of validity and reliability for an instrument does not ensure that a measure 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 relevant EUnetHTA 21 practical guidelines: D4.6 Validity of clinical studies (for individual clinical studies), D4.3.2. Direct and indirect comparisons (for evidence synthesis studies) and D4.5 Applicability of evidence: practical guideline on multiplicity, subgroup, sensitivity and post-hoc analyses.

Summary

- The two main properties of any measurement instruments are validity and reliability.
- The validation of instruments is performed by specific clinical studies with appropriate design and statistical analyses.
- Depending on the purpose of an instrument, different aspects of validity and reliability can have more or less relevance.
- A taxonomy of psychometric properties for PROM is proposed by the international COSMIN group.
- Translation of PROMs requires transcultural adaptation.

Points of attention for the assessment scoping process

- If a specific instrument is requested for measuring an outcome, the quality of the instrument (measurement properties, purpose) is critical.

Requirements for JCA reporting

- 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 (clinical) studies assessing the measurement properties (and measurement model) of the instruments that are used.

5.3 Interpretability of scales

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*" (47). 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 DAS-28 score, patients can be categorized into three groups: active disease (when the score is greater than 5.1), low disease activity (when the score lies between 2.6 and 3.2), and remission (when the score is less than 2.6) (13). 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). While this expression of treatment effectiveness can enhance interpretability, this analysis on the categorical scale should complement the analysis on the continuous scale. In addition, to avoid the risk of data dredging and inflated type-1-error-rate, one measure of treatment effect should be pre-specified in the protocol and statistical analysis plan as a

primary analysis (see the EUnetHTA 21 practical guideline "*Applicability of evidence: practical guideline on multiplicity, subgroup, sensitivity and post-hoc analyses*").

In general, **responder definition** can be used to decide whether each patient has achieved a treatment benefit. This can be done either by assessing whether or not a patient reached a prespecified level of success, or by assessing whether the change in scores is as least equal to a pre-specified threshold (8). This threshold can be obtained by different methods, which are partly subject of scientific debate and are accompanied by different terminology. Most of the methods are based on linking the change in scores to a phenomenon that can come from various perspectives [49]. For example, it can be medical outcomes such as disease severity, symptoms, prognosis, or functional impact (e.g., a minimum change in score associated with a specific gain in functioning).

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 [50–52]. This is also frequently called the minimal clinically important difference (MCID) [50], although it has been used less in recent years. Hundreds of clinical studies have been performed to propose plausible MID values for hundreds of PROMs [53]. Although this approach was initially developed for PROMs, it can be useful for other measurement instruments.

The methods that are usually considered the most appropriate for estimating MIDs are **anchor-based methods,** as they explicitly link a change in score to the patient's perception [51]. A change in score is linked to the response for a unique item: a **patient global rating of change** (PGRC). 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** [51]. In contrast to anchor-based methods, only the overall variability in scores is used in distribution-based methods. Thus, they are criticized as they do not explicitly refer to the meaning of the change for patients (51). 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 [54]. Although not initially developed for responder definitions, *d* values of 0.2 and 0.5 are still proposed as plausible MID values. 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 [55].

MIDs are sometimes identified on the basis of expert opinion [51]. Such MIDs are only a representation of what experts think about a change that patients consider significant.

Another possible responder definition, albeit less common, is the concept of **patient acceptable symptomatic state** (PASS), mostly used in rheumatology [56]. 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.

Lastly, 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 [51]. 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).

Summary

- To enhance interpretability, a responder definition that classifies which patients are supposed to have experienced a treatment benefit or not is useful.
- A responder definition can be derived from numerous perspectives.
- As a responder definition leads to discretisation of variables initially measured on a continuous scale, outcomes can be analysed with corresponding summary statistics and effect measures to complement the analysis on the continuous scale.

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 (methods for estimation, perspective, rule for classifying patients).
- References, as provided by the HTD, to allow full access to the literature justifying the responder definitions used.
- The measure of an outcome that was prespecified as part of the primary analysis (e.g., on a continuous or categorical scale).
- Along with results expressed according to the responder definition (summary statistics, effect measure), 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. McLeod C, Norman R, Litton E, et al. Choosing primary endpoints for clinical trials of health care interventions. *Contemp Clin Trials Commun* 2019;16:100486.

2. FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. Silver Spring, MD: US Food and Drug Administration–US National Institutes of Health; 2021.

3. US National Cancer Institute. *Definition of an Endpoint*. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/endpoint (accessed 23 Sept 2022).

4. Higgins JPT, Li T, Deeks JJ. Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, et al, editors. *Cochrane Handbook for Systematic Reviews of Interventions version 63* (updated February 2022). London, UK: Cochrane Collaboration; 2022.

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

6. Mayo NE, Figueiredo S, Ahmed S, et al. 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.

7. Mokkink LB, Boers M, van der Vleuten CPM, 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.

8. US Food and Drug Administration. *Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Silver Spring, MD: US Food and Drug Administration; 2009.

9. EuroQol Research Foundation. *EQ-5D-3L User Guide*. Rotterdam, The Netherlands: EuroQol; 2018. Available at https://euroqol.org/publications/user-guides (accessed 23 Sept 2022)

10. Huhn S, Axt M, Gunga HC, et al. The impact of wearable technologies in health research: scoping review. *JMIR mHealth uHealth* 2022;10(1):e34384.

11. Byrom B, Doll H, Muehlhausen W, et al. Measurement equivalence of patient-reported outcome measure response scale types collected using bring your own device compared to paper and a provisioned device: results of a randomized equivalence trial. *Value Health* 2018;21(5):581–9.

12. Cho PJ, Yi J, Ho E, et al. Demographic imbalances resulting from the bring-your-own-device study design. *JMIR mHealth uHealth* 2022;10(4):e29510.

13. Prevoo ML, van 't Hof MA, Kuper HH, et al. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum* 1995;38(1):44–8.

14. European Network for Health Technology Assessment. *Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints*. Diemen, The Netherlands: EUnetHTA; 2015. Available at https://www.eunethta.eu/wp-content/uploads/2018/02/WP7-SG3-GL-clin_endpoints_amend2015.pdf?x69613 (accessed 23 Sept 2022).

15. European Medicines Agency. *Clinical Efficacy and Safety Guidelines*. https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-guidelines/clinical-efficacy-safety-guidelines (accessed 23 Sept 2022).

16. Epstein AM. The outcomes movement — will it get us where we want to go? *N Engl J Med* 1990;323(4):266–70.

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

18. World Health Organization. *International Classification of Functioning, Disability and Health: ICF*. Geneva, Switzerland: World Health Organization; 2001.

19. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995;273(1):59–65.

20. COMET Initiative. *Core Outcome Measures in Effectiveness Trials*. https://www.comet-initiative.org/ (accessed 23 Sept 2022).

21. OMERACT. *Outcome Measures in Rheumatology*. https://omeract.org/ (accessed 23 Sept 2022).

22. COMET. *COMET initiative database*. https://www.comet-initiative.org/studies (accessed 23 Sept 2022).

23. Ramsey I, Eckert M, Hutchinson AD, et al. 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.

24. 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.

25. Williamson PR, Barrington H, Blazeby JM, et al. Review finds core outcome set uptake in new studies and systematic reviews needs improvement. *J Clin Epidemiol* 2022;150:154–64.

26. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Harmonised Tripartite Guideline E9. Statistical Principles for Clinical Trials*. Geneva, Switzerland: ICH; 1998.

27. Atkinson A, Colburn W, DeGruttola V, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69:89–95.

28. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? *JAMA* 1999;282(8):790–5.

29. Haslam A, Hey SP, Gill J, et al. 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.

30. Schuster Bruce C, Brhlikova P, Heath J, et al. 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. *PLOS Med* 2019;16(9):e1002873.

31. Ciani O, Buyse M, Drummond M, et al. Use of surrogate end points in healthcare policy: a proposal for adoption of a validation framework. *Nat Rev Drug Discov* 2016;15(7):516.

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

33. Ciani O, Buyse M, Drummond M, et al. 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.

34. Grigore B, Ciani O, Dams F, et al. Surrogate endpoints in health technology assessment: an international review of methodological guidelines. *Pharmacoeconomics* 2020;38(10):1055–70.

35. Bujkiewicz S, Achana F, Papanikos T, et al. *Multivariate Meta-analysis of Summary Data for Combining Treatment Effects on Correlated Outcomes and Evaluating Surrogate Endpoints*. NICE DSU Technical Support Document 20. London, UK: National Institute for Health and Care Excellence; 2019.

36. Pharmaceutical Benefits Advisory Committee. Australian Government, Department of Health and Ageing. *Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (Version 5.0)*. Canberra, Australia: Commonwealth of Australia; 2016.

37. European Network for Health Technology Assessment. *Endpoints used in Relative Effectiveness Assessment. Safety*. Diemen, The Netherlands: EUnetHTA; 2016. Available at https://www.eunethta.eu/wp-content/uploads/2018/03/WP7-SG3-GL-safety_amend2015.pdf?x69613 (accessed 23 Sept 2022).

38. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *MedDRA – the Medical Dictionary for Regulatory Activities*. http://www.meddra.org/ (accessed 23 Sept 2022).

39. World Health Organization. *International Classification of Diseases*. 11th revision. Geneva, Switzerland: World Health Organization; 2018.

40. US National Cancer Institute. *Common Terminology Criteria for Adverse Events (CTCAE)*. https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm (accessed 23 Sept 2022).

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

42. de Vet HCW, Terwee CB, Mokkink LB, et al., editors. *Measurement in Medicine: A Practical Guide*. Cambridge, UK: Cambridge University Press; 2011.

43. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd edition. New York, NY: McGraw-Hill; 1994.

44. 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.

45. 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.

46. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol* 2011;64(12):1303–10.

47. Mokkink LB, Terwee CB, Patrick 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.

48. Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 2000;25(24):3186–91.

49. 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.

50. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407–15.

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

52. Vanier A, Sébille V, Blanchin M, et al. 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.

53. Vanier A, Woaye-Hune P, Toscano A, et al. What are all the proposed methods to estimate the minimal clinically important difference of a patient-reported outcome measure? A systematic review. Presented at the 24th annual conference of the International Society of Quality of Life, Philadelphia, 18–21 October 2017.

54. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. New York, NY: Psychology Press; 2009.

55. 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.

56. Tubach F, Wells GA, Ravaud P, et al. Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues. *J Rheumatol* 2005;32(10):2025–9.

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

58. 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.

59. Hess LM, Brnabic A, Mason O, et al. 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.

60. 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* 2020;21:100332.

61. RECIST Working Group. *RECIST. The official site of the RECIST Working Group*. https://recist.eortc.org/ (accessed 23 Sept 2022).

APPENDIX A: SPECIFIC DEFINITIONS OF OUTCOMES USUALLY USED IN ONCOLOGY

As in other treatment areas the OS has been regarded as the final patient-centred outcome in oncology (57). Improvement in OS clearly demonstrate clinical benefit which is meaningful to the patients. However, measuring OS often requires a large number of patients and long follow-ups. Long-term survival 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 most often reported disease related outcomes are **progression free survival** (PFS) as 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 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 seems to be frequently used surrogate endpoint 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 (58). The corelation between PFS and OS not always is confirmed by the final results, especially in studies of targeted therapy or immunologic agents (59).

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 (57).

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 be defined when utilizing DFS as an outcome (57).

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 now the outcome reserved for neoadjuvant settings while DFS is applied in adjuvant settings (60). 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. RECIST (Response Evaluation Criteria in Solid Tumours) is the most common 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 (61).

Use of clinical endpoints in cancer treatment continues to expand and evolve as new cancer therapies, like immunotherapy, are developed. There is a need to differentiate outcomes for various treatment lines in oncology. Immune therapy in cancer treatment introduced extended use of biomarkers intended to serve as new surrogate clinical endpoints.