EUnetHTA 21

EUnetHTA 21 – Individual Practical Guideline Document

D4.4 – OUTCOMES (ENDPOINTS)

Version 1.0, 25/01/2023
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
V0.4	**24/11/2022**	Fourth draft for CSCQ validation
V0.5	**08/12/2022**	Fifth draft for CSCQ validation
V0.6	**10/01/2023**	Sixth draft for CEB endorsement
V1.0	**25/01/2023**	Date of Publication

Disclaimer

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 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
CEB	

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) reviewed and discussed several drafts of the deliverable before validation. The Consortium Executive Board (CEB) then endorsed the final deliverable before publication.

Associated HTAb & Stakeholders participating in public consultation

The draft deliverable was reviewed by associated HTAb and was open for public consultation between 03.10.2022 and 01.11.2022.

Associated HTA bodies who reviewed	Dachverband der Österreichischen Sozialversicherung, [DVSV], Austria Norwegian Institute of Public Health, [NIPH], Norway Directorate for Pharmaceutical Affairs Ministry for Health [DPA], Malta Swedish Agency for Health Technology Assessment and Assessment of Social Services [SBU], Sweden Health Information and Quality Authority [HIQA], Ireland Finnish Medicines Agency [FIMEA], Finland
Stakeholders who reviewed during public consultation	International Association of Mutual Benefit Societies (AIM), Belgium AstraZeneca (AZ), Europe

	Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung, chronischer Erkrankung und ihren Angehörigen e.V. (BAG SELBSTHILFE), Germany German Medicines Manufacturer´s Association (BAH), Germany European Association of Hospital Pharmacists (EAHP), Belgium Edwards Lifesciences, Europe European Federation of Pharmaceutical Industries and Associations (EFPIA), Belgium European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) HTA SIG, Europe European Organisation for Research and Treatment of Cancer (EORTC), Belgium European Huntington Association (EHA), Belgium European Confederation of Pharmaceutical Entrepreneurs (EUCOPE), Belgium GKV-Spitzenverband – GKV-SV, Germany GSK, Europe IGES Institut GmbH and HealthEcon AG (IGES LifeScience), Germany OAK Access, Netherlands Les Entreprises du Médicament (Leem), France SKC Beratungsgesellschaft mbH (SKC), Germany European Union of General Practitioners/ Family Physicians (UEMO), Belgium Verband Forschender Arzneimittelhersteller (vfa) e.V., Germany HTAi Patient and Citizen Involvement in HTA Interest Group (PCIG), Global European Society for Medical Oncology (ESMO) Lymphoma Coalition, Lymphoma Coalition Europe (LCE), France Medtronic (Mdt), Switzerland Patient Focused Medicines Development (PFMD), Belgium F. Hoffmann La Roche (Roche), Switzerland European Hematology Association (EHA), the Netherlands Bayer AG & Bayer Vital GmbH, Germany Takeda Pharmaceuticals International AG (Takeda), Belgium, Switzerland MedTech Europe (MTE), Belgium European Organisation for Rare Diseases (Eurordis), France European Alliance for Vision Research and Ophthalmology (EU EYE), Belgium **Outside EU/EEA countries** Quality HTA, Canada ISPOR – The Professional Society for Health Economics and Outcomes Research

TABLE OF CONTENTS

1 INTRODUCTION 7
1.1 Problem statement, scope and objectives 7
1.2 Relevant articles in Regulation (EU) 2021/2282 7
2 DEFINITIONS AND GENERAL CONSIDERATIONS 8
2.1 Definitions 8
2.2 General considerations 9
3 CLINICAL RELEVANCE 10
3.1 Definition of patient-centred outcomes 10
3.2 Determinant outcomes for specific therapeutic areas 11
3.3 Surrogate outcomes 12
4 SAFETY 14
4.1 Terminology for JCA 14
4.2 Safety: overall and specific adverse events 14
4.3 Information to be reported for safety outcomes 15
5 VALIDITY, RELIABILITY AND INTERPRETABILITY OF OUTCOMES MEASUREMENT INSTRUMENTS 16
5.1 Definitions and general considerations 16
5.2 Validity and reliability of scales 17
5.3 Interpretability of scales 18
6 REFERENCES 21
APPENDIX A: SPECIFIC DEFINITIONS OF OUTCOMES USUALLY USED IN ONCOLOGY 26

LIST OF ACRONYMS - INITIALISMS

AE	Adverse event
COA	Clinical Outcome Assessment
CDAI	Clinical Disease Activity Index
CEB	Consortium Executive Board
ClinRO	Clinically reported outcome
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
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
ICHOM	International Consortium for Health Outcomes Measurement
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
ObsRO	Observer-reported outcome
ORR	Objective response rate
OS	Overall survival
PASS	Patient-acceptable symptomatic state
PerfO	Performance Outcome
PFS	Progression-free survival
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
SAE	Serious adverse event
sets-STAD	Core Outcome Set Standards for Development

Sets-STAR	Core Outcome Set Standards for Reporting
SUSAR	Suspected unexpected serious adverse reaction
TTP	Time to progression
WHO	World Health Organization
WHO-ART	World Health Organization adverse reaction terminology

1 INTRODUCTION

1.1 Problem statement, scope and objectives

Clinical outcome assessments (COAs), used in clinical studies, are a key component of health technology assessment (HTA). They provide the measure of the clinical benefit of the targeted treatment on how patients feel, function, or survive (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 EUnetHTA 21 practical guideline D4.2 *Scoping process*), in which 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 an important 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) and Article 8(6)), 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 estimated based on COAs will be described as required in the scoping process on the basis of the predefined parameters. However, the ability to conclude on the clinical added value of a treatment can be impacted by factors such as the appraisal of the level of validity and reliability of outcomes measurement instruments or of the relevance of 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 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 COA are provided in EUnetHTA 21 practical guideline D4.6 *Validity of clinical studies* regarding outcomes assessed in original 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, disease control, function, 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**" (corresponding 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)) defines in an accurate way how the outcome is assessed at a patient-level (including use of a specific outcomes measurement instrument; see Section 5). For instance, if the outcome is mortality, the measure of the outcome could be whether the patient is alive or dead 28 days after inclusion. If the outcome is pain, the measure of the outcome could be the change in the level of pain on a patient-reported numeric rating scale (from 0-10) at 24 hours 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, "proportion of deaths 28 days after inclusion" and "mean change in the level of pain 24 hours after the initiation of the treatment" are summary measures corresponding to the two measures of outcome described above. It is sometimes argued in the literature that this difference between an outcome and its measure (or summary measure) is the difference between outcome (as the concept) and "**endpoint**" (as the measure or summary measure) (3,4). However, there is no internationally agreed definition. The two terms are frequently used interchangeably (5). In this guideline, 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 primarily understood as a comparison of the summary measure of outcomes between groups (active and/or placebo). 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 estimand framework complements the PICO framework in precisely defining the treatment effect of interest (2), and while the two concepts are not equivalent, overlap between the two frameworks on other attributes such as population, treatment, intercurrent events are covered in the EUnetHTA 21 practical guideline D4.2 *Scoping process*, D4.6 *Validity of clinical studies* and D4.5 *Applicability of evidence*.

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.

First, the main source of information can come from activity by healthcare professionals. **Clinically reported outcomes** (ClinROs) are assessed by healthcare professionals during clinical examination of a patient and involve clinical judgments of patients' observable signs, behaviours or other physical manifestations. 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. **Performance outcomes** (PerfOs) are clinician-reported outcomes but require active patient involvement to complete a standardized task (e.g., 25-foot walk test with ankle-worn sensor, cognitive tests).

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*" (11). They are measured by **patient-reported outcomes measures** (PROMs), which are mostly self-reported outcomes assessed by administered questionnaires with pre-specified response formats (e.g., Likert scale or

numeric scales). Other formats such as unstructured surveys are possible. 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. Some PROMs measure health status (for instance, the EQ-5D instrument measures health status as a combination of five broad concepts (12)). Other outcomes such as symptoms (including fatigue and pain), anxiety, depression, functioning, impairment, disability and impact on daily living can be assessed using PROMs. PROMs that assess HRQol are usually divided into generic instruments and disease-specific or population-specific instruments (13). Generic instruments can be used in various situations and can allow a comparison of the level of HRQoL of populations affected by different medical conditions. Nonetheless, for specific medical conditions, they can have insufficient content validity (i.e., they cannot capture adequately all the facets of the considered condition). In those cases, the use of disease-specific or population-specific instruments should be considered. In addition, PROMs that assess health status, such as the five dedicated items of the EQ-5D, or HRQoL, such as the SF-6D, allow the measure of utility values, which are generally used for economic evaluation. When they are used for that purpose, they can be referred to as **multi-attribute utility instruments** or **generic utility instruments** (14).

A third source of information can be observers. An **observer-reported outcome** (ObsRO) is a measurement based on an observation by someone else than the patient or a health professional (8). This may be a parent, spouse, or other 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 dependent of 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). A distinction should be made between an ObsRO and a **proxy-reported outcome**. A proxy-reported outcome is a measurement based on a report by someone other than the patient reporting as if he or she is the patient (8). A proxy-reported outcome is not a PRO. A proxy report is also different from a clinician or observer report, as the observer, in addition to reporting his or her observation, may interpret or give an opinion based on the observation.

A particular case is the increasing use of **patient-generated health data** such as outcomes using **connected health technologies** (also called **digital outcomes**). These devices can allow an automated measure of outcomes in settings other than the usual visits for clinical studies, such as in home settings (4,15). They can be roughly divided into 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). They can be collected at a high frequency, but analyses can be challenging due to data handling in the context of the European general data protection regulation requirements, or the large datasets that are collected. Due to the lack of experience, it is important to provide validation of such outcomes and discuss their acceptability with HTA bodies (HTAbs). As with any other outcomes validation the approach should be based on a proper process, including the validation being performed separately and before any studies are initiated with the intent to use such outcomes.

Lastly, categories for classifying outcomes are not mutually exclusive, as some outcomes measurement instruments require the collection of elements from multiple sources. For example, the Clinical Disease Activity Index (CDAI) for rheumatoid arthritis requires clinical and patient-reported elements (16) .

2.2 General considerations

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 relative 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 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 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. 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]"*.

Summary

- Outcomes are concepts for estimating treatment effectiveness.
- The measure of an outcome defines accurately how the outcome is assessed as a variable.
- Effect measure are primarily statistics used to compare the measure of outcomes between groups. Other statistics can be used for other purposes (e.g., 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 a MS wants to specify a measure of an outcome, the wording should follow this template: *"[Outcome of interest] measured preferably as [insert measure]*".
- If a 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 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

- Accurate definition (concept, main source of information, measure, timing, summary and 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 (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 (17,18).

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) (19,20). 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) (21), the Wilson and Cleary biopsychosocial model (22) and the Montreal Accord on PROs (8) can provide further information on outcomes that can be assessed in healthcare.

The EUnetHTA guideline *Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints* recommends that outcomes relevant for HTA should be long-term or final (17). **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 (17). 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 the "event" is well defined and that only validated tools for measurement are used. Time points for COA 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) (23). Initially, these initiatives were in medical fields such as rheumatology (see the OMERACT initiative (24)) 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 (23). 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 (25), as do other sources, for example the International Consortium for Health Outcomes Measurement (ICHOM) (26), 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 and HTDs, it is more likely that patient-centred outcomes will be identified.

- By contributing to less heterogeneity in COA in original clinical studies, COS use 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 COAvia the development of COS may be undermined. It was suggested that to optimise COS usefulness and adoption, 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 studies and systematic reviews needs improvement, as uptake is still low in most research areas (31).

Even though the recommendations from well-established COS should be considered in the selection of outcomes for the scoping process, if such COS are available, it should be noted that COS are not defined from a HTA perspective. Therefore, other health outcomes, deemed relevant by patients, caregivers, clinical experts, or 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, it is important that this document reflects outcomes for assessing the safety and effectiveness 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.
- Other health outcomes, deemed relevant by patients, clinical experts, or HTAbs, can 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 (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.

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 include levels of cholesterol and haemoglobin A1c, 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 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 assessment of the clinical added benefit 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. 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 (17).

Points of attention for the assessment scoping process

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

- If evidence for a patient-centred outcome such as morbidity, overall mortality and HRQoL is likely to be available, then this should be requested during the scoping process;
- 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 (17).

Level of evidence

As detailed in *Endpoints used in relative effectiveness assessment: surrogate endpoints* (17), 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 interventional, 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

If a HTD submits a surrogate outcome to replace an outcome requested by a MS, or if no patient-centred outcome is requested or available, 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. The HTD can also provide scientific literature which demonstrates the link.

For all outcomes requested in the assessment scope, the HTD should provide, in addition to the previously reported follow-up, the latest available data cut, regardless of how immature it is. 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, the latest data cut 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 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 if 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.

There are several frameworks that may be useful when assessing surrogate outcomes. These include reports by Ciani et al. (37), (39), Grigore et al. (40) and Bujkiewicz et al. (41) and guidelines on preparing a submission to the Australian Pharmaceutical Benefits Advisory Committee (42).

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 (43). 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 scoping process, 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. In this situation (only

request for ‘safety’ without any precision), there will be no detail on specific AEs. 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 (44). Other terminology can be used in observational studies, such as the International Classification of Diseases (ICD) (45) and the WHO Adverse Reaction Terminology (WHO-ART), although this is no longer maintained. Patient-reported information related to safety could also be used, such as PRO-CTCAE (Patient-reported outcome Common Terminology Criteria for Adverse Events). 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 non oncology trials (46). A WHO scale has also been developed (47). 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 scoping process. 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 (drug and study) 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.

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.
- A report describing all different AEs must be provided as an appendix to the JCA report.

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 treatments (48). 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 (49). 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 (48). 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** (50). 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 (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 (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) (51). 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 (52). 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. 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.

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 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.

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 of the outcomes measurement instrument and/or instructions.

5.2 Validity and reliability

For appropriate usage, any measurement instrument needs to meet a sufficient level for two main properties: **validity** and **reliability** (48). However, in the context of this document, only COAs (i.e., ClinROs, PROs, PerOS, ObsROS) are considered. As the focus is on outcomes, considerations related to the validation of diagnostic tests or any instrument measuring phenomena with no prognostic value are beyond the scope of this guideline.

Validity refers to the extent to which an outcomes measurement instrument measures what it is supposed to measure (48). For example, if a PROM is designed to measure anxiety levels, the resulting score(s) must correlate with general anxiety symptoms across conditions but not correlate with other mental health outcomes like 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) (53) or **bias** in measurement (i.e., systematic errors). Reliability refers to the extent to which a measure produces similar results under consistent conditions (48). 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. 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 sub properties (e.g., content validity, criterion validity, structural validity, inter-rater reliability, test-retest reliability, internal consistency). Moreover, they are frequently not fully assessed in a single study; investigation of these properties is an ongoing process (54). 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 are close to the concept of **fit for purpose** COA (i.e., the level to which a COA is sufficient to support its proposed use) (1,54). 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 (54). De Vet et al. (48) 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 (55). 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 (56) ,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 (48). 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 (57). 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**) (58).

A sufficient level of validity and reliability for an outcomes measurement 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 original 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 outcomes measurement instruments are validity and reliability.
- The assessment of the measurement properties of instruments is performed by specific studies with 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.

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

5.3 Interpretability

I**nterpretability** 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*" (55). 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 CDAI, 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) (59). 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 enhance interpretability. If in general outcomes that can primarily be conceived as continuous phenomena (e.g., HRQoL) should be assessed

and analysed using first measures and methods that are consistent with this continuous property, analysis on the categorical scale 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 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 classify each patient as having achieved a treatment benefit or not. 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 (11). 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 (60). For example, it can be 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 (61–63). This is also frequently called the minimal clinically important difference (MCID) (61), although it has been used less in recent years. Although the term MID has been used to describe a threshold to interpret between-group differences in scores (e.g., difference in mean change from baseline), the intention within this guideline 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 (64). 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 (62). 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** (62). 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 (62). 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 (65). 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 (66). 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 (67). MIDs are sometimes identified on the basis of expert opinion (62). 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) (68–70).

Another possible responder definition, albeit less common, is the concept of **patient acceptable symptomatic state** (PASS), mostly used in rheumatology (71). 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 (62). 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 measures 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 bibliography 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 measure, 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. 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. US Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009.

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

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

14. 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.

15. 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.

16. 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.

17. European Network for Health Technology Assessment. Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints [Internet]. 2015. Available on: https://www.eunethta.eu/wp-content/uploads/2018/02/WP7-SG3-GL-clin_endpoints_amend2015.pdf?x69613

18. European Medicines Agency. Clinical efficacy and safety guidelines [Internet]. Available on: https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-guidelines/clinical-efficacy-safety-guidelines

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

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

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

22. 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.

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

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

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

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

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. janv 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. 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.

40. 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.

41. 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.

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

43. European Network for Health Technology Assessment. Endpoints used in Relative Effectiveness Assessment. SAFETY. [Internet]. 2015. Available on: https://www.eunethta.eu/wp-content/uploads/2018/03/WP7-SG3-GL-safety_amend2015.pdf?x69613

44. 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/

45. World Health Organization. International Classification of Diseases 11th revision. 2018.

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

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

48. 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).

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

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

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

60. 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.

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

62. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S, 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.

63. 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.

64. 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.

65. 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.

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

67. 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.

68. 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.

69. 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.

70. 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.

71. 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. oct 2005;32(10):2025-9.

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

73. 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.

74. 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.

75. 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.

76. 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

As in other treatment areas the OS has been regarded as the final patient-centred outcome in oncology (72). 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 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 (73). The corelation between PFS and OS is not always confirmed by the final results, especially in studies of targeted therapy or immunologic agents (74).

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 (72).

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 (72).

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 (75). 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 (76).

Use of COAs 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 outcomes.