Clinical Outcomes Assessment Selection -

Practical Guidance in Neuroscience Drug Development

Guidance Version 1.2

ECNP Experimental Medicines Working Group

COA Subgroup

Noveber 5th, 2021

https://qrco.de/bcN9Bt

Author List

Coordinators:

Silvia Zaragoza Domingo PhD, Neuropsynchro, EU

Jordi Alonso Caballero PhD, IMIM, CIBER en Epidemiología y Salud Pública (CIBERESP), EU

Montserrat Ferrer Forés PhD, IMIM, CIBER en Epidemiología y Salud Pública (CIBERESP), EU

Coauthors:

Larry Alphs PhD, Denovo Biopharma, US

Peter Annas PhD, H. Lundbeck A/S, EU

Pavel Balabanov PhD, EMA, EU*

Anna-Karin Berger PhD, H. Lundbeck A/S, EU

Kim Bishop PhD, Global Pharma Consultancy, US

Florence Butlen PhD, EU*

Chris Edgar PhD, Cogstate, Ltd., EU

Bill Horan PhD, VeraSci, US

Jan Kottner PhD, Charité-Universitätsmedizin, Berlin, EU

Mark Opler PhD, Medavante-Prophase, US

Monika Vance PhD, Santium, CA

Chris Yavorsky PhD, Valis Bioscience, US

Contributors To The Project

Luke Allen, Cambridge Cognition, UK

David Daniel PhD, Signant Health, US

Michael Davidson PhD, Nicosia University, Chipre

Manuel de Gracia Blanco PhD, Universitat de Girona, EU

Kiri Granger PhD, Cambridge Cognition, UK

Deborah Hall PhD, Heriot-Watt University, Malaysia, Asia

Josep Maria Haro Abad PhD, Hospital Sant Joan de Déu, EU

Rohini Sen PhD, Takeda, US

Daniel Umbricht PhD, Independent consultant, EU

Gerry Dawson PhD, p1vital, EU

Brian Harel PhD, Takeda

John Harrison PhD, Metis Cognition, EU

Judith Jaeger PhD, Cognitionmetrics, US

Jean-Pierre Lindenmayer PhD, New York University School of Medicine, US

Stephen R. Marder PhD, UCLA, US

Anzalee Khan PhD, Nathan S. Kline Institute, US

Vanitha Krishna MA, IQVIA Biotech, US

On Behalf of the ECNP COA Subgroup

(*See* List of Members in Appendix 4)

Funding Sources

Activities related to the editing of this document and licensing copyrighted materials for reprinting have been funded by Neuropsychological Research Organization s.l. (Neuropsynchro), Europe.

Activities related to meeting facilities and networking have been funded by the European College of Neuropsychopharmacology (ECNP), Europe.

Activities related to hosting a website for this protocol will be funded by Santium Media Corporation (SANTIUM), Canada.

*Disclaimer

The views expressed in this document are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the EMA or one of its committees or working parties.

How to Reference This Document

The ECNP COA Group encourages the use of the information included in this document in the pursuit of improving the clinical trial efficiency. Our recommendations and resources, meeting summaries, and more are available to the public for free.

We do ask that you acknowledge the ECNP COA Group as the source whenever using or referencing our materials in any format, written or otherwise, including, but not limited to, print and digital publications, presentations, grant applications, and websites. This applies to both original and derivative works. Contact group leads to discuss educational activities for your research group.

To properly cite the ECNP COA Group Guide:

Please acknowledge the *"ECNP COA Group. Clinical Outcomes Assessment Selection - Practical Guidance in Neuroscience Drug Development. Version 1.0, September 24, 2021*."

If the material was obtained on our website, include a link to the source page: www.CNSCoreOutcomesSet.com.

Contacting Group Leads

Silvia Zaragoza Domingo szaragoza@psyncro.net

Jordi Alonso Caballero Jalonso@IMIM.ES

Creative Commons Statement

Clinical Outcomes Assessment Selection - Practical Guidance in Neuroscience Drug Development

Guidance: Version 1.0

1. Introduction

The last two decades have been fraught with negative or inconclusive results from studies in the area of neuroscience therapy. A high percentage of results from clinical trials in the neurosciences fail to support the safety and efficacy claims of new treatments. This failure rate constitutes a serious challenge to neuroscience discovery, resulting in a low innovation index for neurological and psychiatric diseases compared to that in other fields (Arnerić et al., 2018; van der Doef et al., 2018).

An unfortunate consequence of the "no return" on the massive investments in the development programs is that a number of pharmaceutical companies have pulled out of the neuroscience arena.

There are, of course, a number of reasons the development of new drugs in the neuroscience field has been relatively unsuccessful.

To break the logjam in neuroscience drug research, the development of more innovative study designs and a continued search for genotype-specific biomarkers have been suggested (Eichler & Sweeney, 2018; Sanders et al., 2019).

Since the primary outcome measures in neuroscience drug development are almost always based on clinical outcome assessments, an immediate and intuitive approach would be to identify new clinically relevant and sensitive endpoints and develop and validate better measurement tools.

Problems in neuroscience drug development relate, directly or indirectly, to methodological aspects of interventional trials. The new interdisciplinary network created within the European College of Neuropsychopharmacology (ECNP) focuses on ***Clinical Outcomes Assessment (COA)*** with the aim of addressing the crucial issue of **better selection of outcomes and endpoints for neuroscience clinical trials**.

Specific objectives included in this guidance document are as follows:

- **Promote good practices** on outcome and endpoint selection in early stages of drug development,
- **Maximize the efficiency** of outcome measurement instruments (OMI) selection based on available information on their performance in completed trials,
- **Guide the development and validation of sensitive** new clinically relevant outcomes and endpoints for neuroscience indications, including innovative drug development tools such as digital health monitoring technology,
- **Identify alternative pathways for early drug development** using tools that address novel issues arising from our advancing knowledge of disease mechanisms and heterogeneity with a goal to optimize trial sensitivity (e.g., to support shorter, smaller trials) or to increase our understanding of the relevance of treatment response (e.g., how patient feels, functions, survives). A proof-of-concept (PoC) endpoint does not need to be relevant if (like a biomarker) it is 'reasonably likely to predict' response or outcome.

- **Address the "Patient-centric approach"** as highlighted by recent FDA guidance, to ensure that COAs are relevant and meaningful to the intended target population, i.e., COAs address a meaningful change from patients' perspectives (FDA, 2020a).

Overall, this guidance document aims to promote methods to increase the ***efficiency*** of efficacy detection in early-stage drug development trials (PoC/Phase II), preventing the discarding of new compounds from the companies' pipeline too early.

The standard method for the selection of COAs that is covered in this guidance focuses on the generic area of **core outcome set (COS)** research methods, which are applicable to all therapeutic fields and various health care contexts (acute, chronic, outpatient, inpatient, etc.). To develop this guide, we started by identifying existing successful case experiences on **completed COS projects** for interventional clinical trials for various conditions**, including** projects such as the **NIH-MATRICS** (Green, Nuechterlein, et al., 2004), the **ImproveLTO** (Needham et al., 2017)**,** the **CS-COUSIN** (Kottner et al., 2018) and the **CROSSSD** (Katiri R, 2020), for which the authors presented their work to our work group in lessons-learned sessions followed by a discussion. This activity provided the background for this guide, which will be developed to be applied specifically within the clinical neurosciences field.

Analyzing the legacy of existing COS initiatives in the neurosciences, we can find two different situations: projects aiming to reach consensus on the evaluation of domains for targeted conditions, e.g., cognition, mood, movement disorders, etc., and other projects with a more global scope starting with the identification of the relevant domains and working on COA selection from there. The COS standard approach provides a framework that is useful for drug discovery, even in early stages applicable to various situations from the initial step or from intermediate steps.

As treatment strategies in the neurosciences are becoming increasingly complex, there is a need for a more translational and integrative view to identify and evaluate outcomes while maintaining a patient-centered approach in the early stages of drug research. This guidance document aims to guide translational research of drug developers or independent researchers for the study of the safety and efficacy of new drug therapies. We have mainly adopted a guiding prescription and description approach rather than pointing out specific instruments for conditions of interest. When available, for specific conditions, legacy instruments require consideration, and their pros and cons should be considered from existing experiences with their application; however, a complete COA selection process needs to address how the specific dimensions can be evaluated, as well as the appraisal of the existing evaluation instruments, in addition to the consideration of the regulatory requirements.

As commented, this guidance document has been developed by considering extensive and comprehensive existing and broadly known **COS** guides, i.e., COMET Handbook 2.0 (Williamson et al., 2017) and OMERACT Handbook (Beaton D et al., 2021). The overarching principle for following this strategy is to provide a straightforward and concise method for drug development plans that commonly, at early stages of drug development, have time and resource limitations. Minimal standards for COS projects can also be found in existing initiatives, such as the COS-STAD project (Kirkham et al., 2017), which is generic and applicable to any medical context. Within this guide, we propose a 7-step method, with a special focus in neuroscience therapeutic areas and

highlighting translational issues at the start of drug development. Our aim is to contribute to increasing the likelihood of finding efficacy in clinical trials, provided that the intervention has an effect on the targeted condition.

Additionally, other materials derived from the FDA-scheduled series of public workshops in 2018 to define best practices to *Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments* to measure the patient experience in clinical trials (*see* PFDD Guidance 3 Discussion Document: Select, Develop or Modify Fit-for-Purpose Clinical Outcome Assessments) (FDA, 2020a) were used. This FDA guidance document set provides a clear stepwise process, clarifying the concepts involved in the process of tool validation in connection with labeling claims.

When this process is put into practice, quite often due to time and budget constraints, industrial users might need to shorten the process to decide on the COAs. For that reason, for each of the 7 steps, this guideline also includes mentioning the risks involved in not fully completing each one. It is desirable that the authors of the protocols of specific COS projects provide a discussion section about how shortcomings might have an impact on the final clinical trial results in terms of "signal detection" and the reliability of "PoC studies".

A plan for COA selection needs to be started early in the drug development process, ideally before the initial meetings with regulators (e.g., the FDA pretype C meetings) (*Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products Guidance for Industry*, 2017). In Section 5, we provide advise on how to use this guide and comments on the convenience of starting the process in advance to have a clear roadmap to follow during preclinical and clinical phases.

Some examples in the document are considered to be illustrative of COS projects within the neurosciences field, such as NIH-MATRICS (Marder & Fenton, 2004) psychiatry for the development of treatments for cognitive impairments that are related to schizophrenia, MSOAC (LaRocca et al., 2018) and BICAMS (Langdon et al., 2012) in neurology, for the development of treatments for multiple sclerosis (MS), the CHOICE project for the treatment in children for epilepsy in neurology (Crudgington et al., 2019) and the neuropediatric REiNS project (Plotkin et al., 2013) in neurofibromatosis type 1. All these projects follow the COS standard method, although with some variations; the method starts with a predefined COS selection protocol and specific methodologies to reach a consensus on COS.

The agreement and further use of a standard COS for specific conditions have clear advantages in the research of new treatments, contribute to generating a common language, and facilitate the application of meta-analysis methods of interventional trials when different studies use the same/similar endpoints/measurement methods/instruments.

This COA guidance has been elaborated through the contribution of a number of experts in the neuroscience clinical trials field, applying their knowledge to translational research of drug development. By using this guide, researchers will be in the position to build their own COS/COA selection strategy for specific applications in a condition of interest and for a specific context of use (COU). Materials related to this document and working templates will be found ready for use at the group platform at www.CNSCoreOutcomesSet.com.

We encourage sponsors to use this guide when elaborating their specific COA selection protocols and to standardize the process as much as possible based on the rationale for each interventional context and development phase. Additionally, the evaluation of the risks related to skipping specific steps when conducting the complete process is not possible.

2. COA Decision and Regulatory Environment

Regulators in various geographical areas are continuously issuing guidance documents to clarify requirements for drugs for target diseases or conditions. For some conditions, gold standard instruments are accepted by regulators and used by all drug developers. This is the case for the ADAS Cog and CDR for clinical trials in Alzheimer's Disease, MCCB in Cognitive Impairment in Schizophrenia, and the MDS-UPDRS in Parkinson´s Disease. There are however many situations where regulatory guidance do not specify a single instrument or research guides are out of date, in draft form or nonexistent.

Quite often, no regulatory guidance documents exist or that there are no predefined *gold standard* endpoints or instruments for clinical trials in a particular disease area. For those cases, the closest guidance available is the research guidance produced by scientific societies or those that offer generic COA validation paths (workshops, recommendations, etc.). The most commonly used strategy consists of taking advice from clinical experts and conducting meetings with regulators to agree on the outcomes and endpoint strategies to follow.

Conducting exhaustive literature reviews focusing on the use of COAs in the field is always an option, as is conducting a survey of preferred COAs among a representative sample of end-users, i.e., clinicians, patients or caregivers.

As noted above, although it is not ideal, in novel therapeutic areas, including rare disease indications, clinical experts of the targeted disorder can be the first to propose outcomes and tentatively recommend corresponding clinical instruments (Busner J, 2021; in press). Based on this initial suggestion, the decision process will include other experts, such as clinical scientists, statistics, psychometricians, COA specialists, medical advisors, experts in regulatory affairs, experts on market access, pharmacoeconomists, RWE specialists and experts with other functions and perspectives. The final participating team will vary according to the company and the internal decision-making process. Such roles tend to be complementary; for instance, while clinical experts take into account their clinical experience with a disorder, COA experts also consider the clinical context, the dimensions or concepts of interest, and proprieties inherent to the existing instruments, such as content validity, validation status and known psychometric properties.

For the final proposal of COA instruments, it is important to note that some evaluation instruments used widely in clinical practice might not be fit-for-purpose for regulatory trials, as they may not be designed in a way that makes it likely sensitively detect treatment effects or discriminate between treatment and placebo arms' scores. Some of the properties required for an instrument to be valid for use as a clinical endpoint include measuring well-defined concepts, having a clear recall period, having distinct and nonoverlapping response options representing

clinically meaningful gradations, and having available standardized user annual/training materials (FDA, 2020a).

The processes validating a COA strategy in clinical trials usually starts by meeting with regulatory agencies, which is known as a consultation meeting at the US FDA or EMA, often in the context of an Investigational New Drug application (NDA) or Clinical Trial Application (CTA). These meetings have a predefined agenda(*Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products Guidance for Industry*, 2017). Specifically, during FDA applications, at the initial meetings, the group of experts may or may not be involved.

Later, during scheduled programmatic meetings with regulators, the industry sponsor can agree on the "*validation path*" for a specific targeted endpoint to be used as the primary or key-secondary measure for drug development in the target disease or condition. In this case, COA briefing books and full statistical validation plans should be submitted for approval before using new instruments in clinical research, and this is even more important if these tools will support label claims in the future for the compound.

The FDA is especially active in generating support documents and in organizing public workshops to guide decision processes for COA selection and innovation.

Considering COA Selection in the European Regulatory Environment

In Europe, one of the European Medicines Agency's (EMA) strategic goals is to foster research and the uptake of innovative methods in the development of medicines. This aids in making safe and effective innovative medicines available to patients in a timely manner. The EU-Innovation Network is launched a pilot for simultaneous national scientific advice (SNSA) from national competent authorities (NCAs) on 1 February 2020 to further strengthen early regulatory support for innovation (*see* Innovation in Medicines by the EMA (EMA, 2020a).

Furthermore, the EMA offers scientific advice to support the qualification of innovative development methods for a specific intended use in the context of research and development into pharmaceuticals. Advice is given by the EMA's Committee for Medicinal Products for Human Use (CHMP) on the basis of recommendations by the Scientific Advice Working Party (SAWP). This qualification process leads to a CHMP qualification opinion or CHMP qualification advice. The application can be completed by using the 'IRIS' platform.

Early interactions through the available platforms for regulatory validation of COAs in the intended context of use are to be encouraged (*see* EMA Qualification of Novel Methodologies for Medicine Development through IRIS Platform, EMA, 2020d). The IRIS platform provides a single space for applicants and the EMA to submit requests, communicate, share information and deliver documents concerning a qualification procedure.

It is important that all stakeholders, including researchers and academia, feel encouraged to submit requests to the EMA and have the opportunity to discuss early plans for innovative methodologies with a clear application in pharmacological research.

On 15 December 2005, the EMA launched an "SME Office" to provide financial and administrative assistance to micro-, small- and medium-sized enterprises (SMEs), with the aim of promoting innovation and the development of new human and veterinary medicinal products by SMEs (Carr, 2010; EMA, 2005).

In addition to regular applications, it may also be possible to consider an "expedited qualification" process for endpoints in some conditions, usually when drugs target unmet medical needs or represent a clearly valid added value to existing therapies. Fast Track Designation (United States), Breakthrough Therapy Designation (BTD; United States), and PRIority MEdicines (PRIME) Designation (European Union) are three such programs.

An evolving regulatory environment

In recent years, the FDA has particularly emphasized the relevance of using fit-for-purpose (FFP) scales rather than accepting existing "*gold standard scales*", with the idea that the new FFP instruments are specific to the studied indication and to the clinical trial context (FDA, 2020a). However, it is not always feasible to reach acceptance on innovative COA plans mainly because the cost, resources needed or the time required to complete the full required FDA validation path.

In addition, with the introduction of the more *patient-centric approach* introduced by the same regulatory agency, the field will find issue with previously accepted and commonly used scales in clinical trials. We are facing new challenges, at least with the FDA, whereas EMA seems more prone to stick to "*gold standard*" endpoints and related existing scales.

The *patient-centric* FDA approach will impact future developments because the unique use of ClinROs as primary endpoints, as is the case of Parkinson's disease (MDS-UPDRS), Alzheimer's disease (CRD-SB) or schizophrenia (PANSS), might not be enough. In the case of Parkinson's disease, the MDS-UPDRS Part III (motor examination) is no longer as important as the PRO parts of the same scale; therefore, for the FDA, the focus has shifted toward, for instance, the Non-Motor Part Scales as a coprimary or primary endpoint. Another example is in the AD field, where there is currently a substantial focus on *functions* that are considered meaningful to patients.

The views on this topic vary in other geographical areas, with the Pharmaceuticals and Medical Devices Agency in Japan (PMDA) often following FDA recommendations and, for example, China typically following its own local view of clinical trial endpoints (meaning scales need to be fully validated for a Chinese context).

3. General Comments on COA Selection Strategies

A common strategy when designing clinical trials has been to stick to the COAs used in previous drug development programs, even if the results were eventually negative. As a consequence, when study results are negative, the question of whether it is the compound or the outcome measure that failed often remains unanswered.

Examples of successful development case stories are now available in public databases of labeling claims and endpoints used thorough public FDA databases, the FDA Compendium (FDA, n.d.-c), summarizing COA information for many diseases and conditions in a single resource. The FDA suggests to using the COA compendium as a starting point when considering a COA for use in clinical trials.

Similar trials in public registries of clinical trials are always a valid source of information (clinicaltrial.gov & EMA registry (EMA, n.d.-a)). Additionally, the proposal of existing COA tools that are used in closely related therapeutic areas is also an option to consider at the start of the process, as is the examination of COAs that have been used to measure the targeted domain but have not yet been validated for the specific condition (considered a new COU). It is worth mentioning that when adapting an existing instrument is a valid option, the FDA has recommendations regarding the decision process on how to determine whether to use an existing instrument, to modify an instrument, or to develop a new instrument (*see* FDA, 2020a, p. 2).

When the intention is to use existing validated instruments in a different COU, a full validation plan for the new COU is required before entering into the clinical phases of drug development. Whenever the full COA validation program is not possible to conduct, the sponsor can agree to a compromise in order to collect enough data thorough the phase II program to enable its clinical validation for further use in phase III and, eventually, to support a labeling claim.

Clinical Benefit in the context of FDA regulations

Clinical benefit is defined by the FDA as a positive, clinically meaningful effect of an intervention, e.g., a positive effect on how an individual feels, functions, or survives (FDA-NIH Biomarker Working Group, 2021). This approach is basically *patient centered* and substantially clarifies the role of COAs in the study design as a primary and coprimary or secondary endpoint. *See* Section 7.5.2.2.1.2 for a discussion of performance-based cognitive endpoints. All steps taken during the decision on COAs/OMIs should keep this idea in mind.

The application of this principle in *early drug development* might vary because some COAs, such as PerfOs, although considered objective measures, may not have well-established relationships with patient functioning in daily life. For this situation, for the selection of various types of OMIs–PRO, ObsRO, ClinRO, PerfO or digital health monitoring measures–for phase II or PoC trials, *biomarker type evidence,* e.g., 'reasonably likely to predict' rather than "clinically meaningful", can be adopted in view of future pivotal trials.

PROs in the European Regulatory Context

Regarding the use of COAs in the European context, specifically the use of PROs in the CNS from the regulatory perspective in Europe, in clinical trials, PRO data can be used to help contextualize the observed benefit:

- as an addition to the effects on recognized endpoints to conclude efficacy,
- in chronic, incurable conditions, even stabilization of QoL scores can be sufficient as a claim of benefit.

Often, in the CNS, the use of PROs is limited to serving as additional exploratory endpoint(s) or supportive evidence but are not sufficient on their own to justify positive benefit/risk (B/R) conclusions or specific indication statements.

There are a number of challenges in the development and validation of PROs:

- Testing reliability and validity against those of other PRO instruments and recognized functional scores,
- Validating in an informative population and avoiding cases where there is little room for improvement,
- Testing the sensitivity and reliability before applying the new PRO in pivotal studies,
- Determining the minimal clinically relevant difference, as the mere presence of a statistically significant difference may not be enough,

To move forward, collaboration with stakeholders and regulators is key to developing PROs that target functional domains and that are not captured by existing scales. Early interactions through the available platforms for regulatory validation in the intended context of use are to be encouraged (*see* EMA, 2020b).

4. Early Stages of Drug Development: Experimental vs. Clinically Relevant Endpoints

In the context of phase II or PoC trials, as mentioned above, *fit-for-purpose* instruments are commonly used as tools, and these are seen to be initially sufficient for use in the early stages of drug development. The FDA issued this concept in 2016 in the context of the Fit-for-Purpose Initiative (FPP FDA, 2021a), representing a pathway for regulatory acceptance of dynamic tools for use in drug development programs. Therefore, the designation of 'fit-for-purpose' (FFP) was established due to the evolving nature of these types of drug development tools (DDTs) and the inability to provide a more formal qualification.

More recently, a guidance document was issued by the same regulatory agency with the purpose of clarifying the role of "*what is important to patients*" within the FFP model (FFP FDA, 2021a). Using public workshops, the FDA sought obtain feedback on this topic from a broad range of stakeholders, including patients, patient advocates, academic and medical researchers, expert

practitioners, drug developers, and other interested persons, on topics including electing, developing or modifying fit-for-purpose COAs to measure the patient experience in clinical trials.

According to this document, the FDA determines that a COA is fit-for-purpose based on some general principles (*see* FDA, 2020a, p. 8):

- The COA is appropriate for its intended use
- The COA validity and reliability measure concepts that are clinically relevant and important to patients
- Data can be communicated in a way that is accurate, interpretable, and not misleading (i.e., well defined in the sense of appropriately applied in medical product development).

Convenience to set different validation pathways for endpoints in early stages

One of the specific objectives regarding outcome and endpoint definition is to articulate "different pathways" for the development of tools addressing early phases of drug development, dependent on how we wish to use them and the questions we seek to answer that follow either track of being (a) *biomarker-like* or a (b) *clinically relevant* COA instrument.

There is a substantial difference between both tracks. The first seeks to optimize the sensitivity in the "*experimental early phase*" (e.g., to support shorter, smaller trials with compounds with new/unknown mechanisms of action), and the second track seeks to optimize drugs with well-known mechanisms of action, where the need is more related to *ensuring clinical relevance* for pivotal trials (i.e., benefits on how a patient feels, functions and survives).

In the first model, which is more experimental, an endpoint for a PoC study need not be clinically relevant to patients, as long as, in the manner of a biomarker, it is '*reasonably likely to predict'*. Although it is a primary or secondary endpoint, this gives early-stage endpoints a more "*exploratory*" status.

The application of these two models, *experimental* versus *clinically relevant*, in early stages has the advantage of boosting science toward new compounds with specific mechanisms of action that are not completely well understood/known (direct or indirect), as is the case for many compounds actually used in psychiatry and neurology. The risk of not applying the view of an experimental primary/secondary endpoint as a biomarker in PoC runs the risk of limiting innovative developments to the same compounds with established mechanisms of action, delaying the advancement of new ones.

In the proposed *experimental* COA model, the drive for greater sensitivity can be viewed by regulators as running counter to clinical meaningfulness (i.e., that we are able to capture changes of a magnitude that confer no benefit to patients). Examples of such experimental COAs in neurosciences could be cognition (PerfO) or other experimental COAs directly related to the intended new drug's mechanism of action, including the use of the RdOC (NIMH, n.d.) taxonomy to explore transnosological dimensions in neurosciences beyond specific conditions.

ⓘ A first good example of experimental developments can be found in the field of depression. Innovative experimental models are currently under development for their application in the development of antidepressant agents; there is good evidence that a common feature of antidepressant drugs is that they modulate the processing of emotional information within days of beginning treatment. Specifically, they reduce the negative bias typically seen in depressed patients weeks before they report a reduction in symptoms. This response is also resistant to placebo effects, as patients are largely unaware of the change in their negative bias. There is also strong evidence that this reduction in negative bias may be a necessary prelude to the reduction in symptoms. Thus, this early biomarker of the antidepressant response may be exploited in drug development as a qualified biomarker of efficacy to provide confidence in investing in expensive large-scale clinical trials (Harmer et al., 2009).

ⓘ A second example of the validation of innovative experimental models to support drug development tools in psychiatry is the Reward Task Optimization Consortium (RTOC) (Bilderbeck et al., 2020). This constitutes an initial work toward optimizing a valid, neuroscience-informed and online-administered test battery for the measurement of anhedonia and dysfunctional reward processing for future use in large clinical trials on schizophrenia and mood disorders.

For some conditions, the proposal of innovative experimental strategies for endpoints or outcomes in clinical trials may require very early consultation with regulators to assess the feasibility of the full development program *(see Section 7.6.1)* and reaching formal agreement on the validation path before its use.

5. How to use this guide

It is worth emphasizing that a plan for COA selection needs to be implemented soon in the drug development process, even at late preclinical stages and ideally before the initial meetings with regulators take place (FDA type C meetings or equivalent in EU) to obtain feedback about the measurement strategy from the relevant FDA review division in the US or the EMA division in the EU.

COA selection is not limited to identifying outcomes and instruments used in the past but to designing a strategy tailored for the specific mechanism of action of new drug development in the specific indication and COU.

The use of this guide early enough in the drug development path will produce two main documents, initially a *COA selection protocol* and, if applicable, a *COA development plan*.

COA selection protocols will result in specific concepts/domains to measure and related existing instruments; however, on several occasions, existing instruments will be ready to use in the specific drug development context. This is especially relevant in new areas of drug development

and on orphan indications. All required activities will be summarized within the *COA development plan*.

- The *COA selection protocol* shows the steps proposed for specific drug development in a COU and ideally achieves agreement among all stakeholders. The protocol can be submitted for approval to independent review experts, for instance, members of the ECNP COA Group, to gain an independent view and open it to discussion. Such a COA selection protocol can also be included in public databases (*see* Section 13.1) and eventually published in specialized journals when convenient to preserve industrial properties related to processes.

- The *COA development plan* will result from the execution of the *COA selection protocol* when the conclusion is that additional work on the validation of selected measurement tools is needed (new tools are needed or existing tools need to be used in a new COU). Confirmation of the validity of the instruments in the new COU is needed when there is not sufficient psychometric information about the instrument to ensure its efficiency in the new context of application or when recommended by the COA analysis; the new instruments are to be developed based on existing instruments or item databases or from scratch.

When available, it is an option to select concept-driven instruments that are FFP, providing interpretable outcomes for the purpose intended to enhance the scientific framework of clinical benefit, but as a measure of primary endpoints, such instruments might be seen as a weakness that increases the possibility of a negative study result. This could be mitigated with a more comprehensive COA development plan.

Later, parts if not all of *the COA development plan* document will be useful as a basis for initial discussion with regulators at specific meetings. At this strategic pacification stage, some key documents, such as *COA dossiers* and *briefing books* for the instruments to be used in phase II/III clinical trials, can be extracted from the *Selection Protocol* before its execution, in line with the expectations of the regulatory departments (FDA DDT, EMA) to gain approval for target label claims.

This guide can also be used in the contrary manner: for areas in which a gold standard exists, this guidance allows us to identify its weaknesses and potentially include secondary outcomes to mitigate the risk of a negative result.

6. Building the Working Team for a Specific COS Project

COS developers usually constitute a group with stable members participating in the main decisions and settings of a project. A first task is then to create a COS Project Working Group or Research Steering Group appointed to oversee and manage the project, ideally gathering a multidisciplinary network *(a description of roles and activities can be found in* (Fackrell et al., 2017). The inclusion of multiple perspectives from the inception phases of research, as well as throughout the process, is critical to maximize the impact of research and facilitate research dissemination.

As suggested by the OMERACT Handbook, the group should have international representation (representatives from at least 3 continents), at least one patient partner, and at least 5 participants with content-specific expertise to brainstorm and generate a wide list of possibilities for outcome domains, intended application context and population of intended use.

Reasons for multiple stakeholder involvement include but are not limited to increasing the number of ideas, perspectives, and depth of questions considered; including all sectors affected; establishing credibility and ensuring relevance and meaning to various groups; enhancing quality; increasing the face validity of final proposals; identifying concerns, barriers, and controversies that would not have otherwise been considered; increasing transparency; increasing the uptake and dissemination of the outcome measure or research product; and fostering relationships for future research efforts.

We can add other reasons to this list: increasing technological understanding for future IT developments of any type, increasing the awareness of R&D processes involved in COA developments if needed, etc.

7. Proposed Standard Protocol for Developing a COS Project

The standard COS can include various types of COAs, such as patient-reported outcomes (PROs), clinician-reported outcomes (ClinROs), observer-reported outcomes (ObsROs), proxy-reported outcomes (ProxyROs) and performance-reported outcomes (PerfOs), as performance-based skills assessments and digital health monitoring measures.

These minimal sets should assess a minimum list of impacts that matter most to patients, are likely to demonstrate change (including differences in trial arms related to disease burden, treatment burden, and, if applicable, physical function), and ideally should be assessed during a clinical trial. A standard core set might be relevant across several disease populations or subgroups or be focused on attributes of a specific disease (FDA, 2020a).

We propose a way to standardize the identification of a COS at least within the neurosciences field, keeping in mind that what is included in this guide can also be applied to other fields in health.

The following sections are built with a similar predefined structure to facilitate the work of the reader. Whenever possible, we describe the *goal of each step*, a *basic description* of what is involved, a list of *key considerations* to take into account, what are considered a set of *minimal activities* to be included for its accomplishment, *additional actions* to include to reach higher precision, and a *risk forecast* in the event that the step is skipped in the process. Illustrative examples from completed projects are also provided when relevant.

The content of each section is not to be used as a substitute for other more elaborated or technical manuals but can be used as a minimal set of directions and recommendations without prevent higher accuracy in the process or more elaborated methods.

7.1. Step 1 – Disease Model and Background

7.1.1. Review of existing literature

The goal is to identify potential outcomes from existing work as information provided to inform the consensus process for next steps.

Before working on a COS project, initial deep research on existing similar projects is the starting point. We have grouped all information that will be needed to conduct comprehensive work into a single step. However, the scope of the review can cover any of the following and needs to be considered at the start of a review.

- Identifying existing COS projects, either completed and/or ongoing
- Research on Disease Impact Models/Disease Concept, named qualitative research
- Specific literature search using defined outcomes/instruments (CT study protocols, clinical guidelines, research guidelines, meta-analyses or any other applied methodology for selecting outcomes).

There are several data sources that should be considered:

a) systematic review of published studies
b) reviews of published qualitative work
c) investigation into items collected in national audit sets or focus groups with key stakeholders to understand their views of outcomes of importance

Other sources, if available, can also be of great interest, as are existing protocols within clinical trial registries such as clinicaltrial.gov.

Data extraction is considered in terms of the following: study characteristics, outcomes, outcome measurement instruments and definitions (*see* Annex 3 for a proposed complete summary of data extraction).

Minimal activities to be completed include defining a search strategy in advance (keywords, time frame if relevant for the disease/intervention under study).

In addition, all excluded research works should be listed and the reason for exclusion should be described.

Key aspects to consider are identifying existing completed research works of any type and considering the overlap with other existing groups undertaking similar activities.

It is interesting to consider in detail the discussion sections of the literature selected as a source of relevant information to identify gaps and guide future COS research.

The *risk* in not conducting this initial review work is duplicating efforts unnecessarily instead of continuing the research started by other groups (usually experts in the field) and fall victim to the same issues instead of learning from past experiences. Work done by others in the past can be a great starting point in a step that is quite time consuming and requires a systematic approach.

Another important aspect to consider when reviewing literature for clinical trials is the fact that even if similar constructs are measured in trials, there might be substantial variation regarding the use of measurement instruments, the time of measurement, and the methods of aggregation (Lange, 2020). These variations do not allow firm meta-analysis or qualitative comparison of evidence from clinical trials. However, all the information obtained from the initial review will contribute to the awareness of variations in OMI uses, which needs to be considered in terms of how it is expected to impact the targeted domains and how to mitigate the risk within the COS specifications.

Review of the existence of previous COS projects either completed or ongoing

Public registries of completed or ongoing COS projects are especially useful, as the www.comet-initiative.com is continuously updated. This resource is known as the COMET database and is the only such repository, regardless of the type of sponsor promoting the work. Completed and published COS projects can be found in commonly used databases, such as Medline, CINAHL, Embase, Cochrane Library, and PsychInfo. In addition, specific projects funded by the European Commission in the EU can be found via CORDIS website resources (EU European Commission, n.d. CORDIS).

Evaluation of the overlap in case there is an existing COS project

If no overlap is identified or there is strong justification for developing a new core outcome set, regardless of one existing, then it is advisable to proceed. Overlap may exist, but there might be specific reasons to start a new COS project with the same therapeutic indication. Reasons such as the following can justify the work (*See* OMERACT protocol, Beaton D et al., 2021).

- The existing COS lack relevant stakeholders. Usually, this is the case of old COS work that does not include patients and but only clinicians in the consensus panel.
- The existing COS was performed before specific therapeutic advances were made in the field, with a completely new scenario regarding disease management and impact.
- The existing COS corresponds to a specific subgroup of patients or disorder subtype that does not represent the targeted indication or patient subpopulation.

It is advisable to minimize unnecessary duplication of efforts. Although there might be no exact match for the scope of interest, there may be related COSs, e.g., a COS in a specific subtype of the disorder or conducted in a different country or with a lack of patients as stakeholders. For such cases, all this work can be used to start an outcome matrix to report the outcomes reported in the eligible studies. This information can be used as a starting point and as a tool to match with the new results that show bias or non-consistency and will eventually be useful for analyzing the discussion section for a better understanding of the disorder and its related outcomes.

Currently, there is much interest in and activities related to COS development. There are so many COS initiatives worldwide that must first be reviewed to determine whether there is a related ongoing COS project.

If a COS already exits, the goal should be to improve its quality in terms of additional work related to a particular stakeholder or countries, to include more information than is included in the existing COA, or to develop an alternative consensus method.

7.1.2. Review of existing Qualitative Studies or Disease Impact Models

The goal is similar to that of the previous section, emphasizing the identification of qualitative research work either published or unpublished with patients suffering from the disorder or relatives or caregivers either formal or informal.

The interaction with patients, or other stakeholders, will allow us to identify concepts (symptoms and impacts) and dimensions that describe patients' experiences of the disorder from the patient´s perspectives as well as whether any group has elaborated from this information a disease impact model. For disorders with an impact on caregivers/relatives/partners, it is important to collect their experiences and include them within the model. Examples within the literature of disease impact models, including the model of MS produced by (Martin et al., 2017) and Angelman syndrome, include views of patient families (Willgoss et al., 2021).

Minimal activities include that once all available research works have been identified, the selection criteria should be applied, and those that do not meet the criteria should be excluded. From the selected research, relevant information about concepts, dimensions and overall model, if any, can be extracted.

Identified, qualitative studies will support the definition of concepts included in the model of outcomes to be measured in terms of dimensions and instruments (*see* BEST FDA-NIH Biomarker Working Group, 2021).

Key considerations are that disease types and existing disease subtypes clearly matter in this initial search. Although studies might exist, it could be the case that the disorder type does not match the patient population targeted in the sponsor's new intervention.

Appraisal of how the existing information justifies concept definition in cognition

For cognitive measures, the relationship between concepts and measurement could be slightly more complex due to the low awareness of patients about their own cognitive level of performance, which has been observed in some neuropsychiatric disorders (*see* Section 7.5.2.2.1).

In this case, efforts should be made to identify qualitative studies with reports from patients and proxies or relatives regarding the relationship between cognitive health and functionality in daily life.

All this information will be useful in justifying the content validity of more objective instruments to be used to measure outcomes.

Search strategy for past research

Minimal activities involve summarizing and defining a *search strategy* (keywords, resources to be used, inclusion/exclusion criteria for research works extracted, time frame if relevant for the disease/intervention under study) in advance.

There are several examples to follow to guarantee the understanding of initial work and any potential bias (Prinsen et al., 2016-Additional file 1).

As *additional activities,* this process of evaluating qualitative research can include checking for specific processes such as patient interviews and cognitive debriefing. An example of a qualitative interview guide to collect information about meaningful changes according to patients and/or relatives (defining clinical meaningful within patient change) can be found in neurofibromatosis research (Rose T, 2020). Within this example, among other activities, a patient listening session with the FDA was organized; this informal one-hour discussion was designed to help those involved in the drug approval process understand the obstacles and unmet needs of patients who might benefit from emerging new drug treatments (Children's Tumor Foundation (CTF), 2019).

As *risks*, it can be considered a potential source of bias to overclassify quantitative versus qualitative research projects during the systematic review stage, therefore excluding relevant information from the initial analysis. Some projects use a mixed method approach that combines qualitative and quantitative data in the same project, resulting in valuable information for a new COS activity.

7.1.3. Defining the unmet need (or gaps to fill)

The goal is to describe the rationale for the development of a COS in the context of what is already known with the focus on the contribution to actual clinical application.

Considering all the information collected from various sources, this step will propose areas of improvement within a COS under development for the intervention of interest. Therefore, this section will focus on filling the gaps in the existing knowledge, which are otherwise important to the COS.

An accurate review of the factors considered in past research versus factors that have not been considered in the literature will highlight gaps to address in future steps.

The key consideration is to cover the list of potential gaps in past research mentioned in review papers and handbooks as follows:

- The stakeholder group's views not encompassed in the information collected,
- Insufficient description of patient participants in past research
- Insufficient information about disease features/subtypes in past research
- Specific disorder is not at all or not sufficiently covered in past research
- Insufficient number of participants to ensure representativeness (e.g., relevant in orphan indications)
- There is no mentioning of the type of focus group interviews used in past research.
- Relevant innovations in the field have substantially changed the patient experience
- Relevant nuances across countries limiting the generalization of results from past research
- COS developed for other purposes rather than interventional research, e.g., clinical or nondrug interventions.
- Outcomes requiring technologies that do not exist or are more invasive than the standard ones (momentary evaluations, remote surveillance, etc.) are not covered in past research.

Minimal activities the next steps of this analysis include the need to conduct additional qualitative research before proceeding to the next steps of the project or to acknowledge the results acquired thus far as the reference for future consensus work.

Additional activities can be used to gather complementary information about nuances in targeted geographical areas that will have an impact on the acceptability of the COS for different applications across regulatory agencies.

To better *represent patients* and to understand the heterogeneity of the disorder, the use of IA/machine learning (ML) technology has also been suggested. For some disorders with a high variability expression, this methodology can clearly increase the possibilities of building a more realistic impact disease model.

For some indications, such as orphan disorders, this can be an option due to the difficulty of reaching patients suffering from the disorders.

Rating scales are currently understudied, and other *innovative measurement strategies* are more interesting to pharma, as they are used to measure patient data in real life. To explore boundaries, more innovative methods are desirable for this group.

For *risks,* the appraisal of reviewed literature is critical to plan for future actions that will be mentioned in the *COA development plan* when convenient (see *Section 5*). The risk of not taking this into account includes working with information that is obtained from a non-representative sample and that is biased toward the interest of other researchers/groups rather than the actual interest of early-stage drug development.

In some fields immature in terms of clinical research, it might be difficult to find publications reporting a significant number of outcomes. To identify publications that evaluate an outcome for a specific disease, a deep analysis of the literature can provide information not only about the specific outcomes but also about the state of the art regarding the specific indication, such as preferences, paths to follow, variability in the outcomes measured or how they are measured, and it can also help identify bias for selective outcome reporting (toward favorable ones). The analysis can provide information about the needs in the field regarding standardization, barriers to meta-analyses and other relevant issues, such as outcomes that are missed.

In contrast, when a *gold standard* exists, it is worth mentioning that quite often, consensus on a COS in a specific indication hinders further innovative approaches in clinical trials with new compounds. A consensus on a COS might not be helpful for advancing the field because, once accepted by regulators/communities, it can limit innovation. Once a gold standard is accepted, innovation usually stops. Standardizing a COS does not mean that we cannot continue to innovate. A COS defines the minimum; any other (innovative) outcome can be measured as well. Examples of such situations can be found in the field of AD, where innovation in ADAS Cog has stopped owing to a lack of consensus.

During the group meetings, we listed the following areas where there are still needs for COA improvement.

TABLE 1 SUMMARY OF THE NEED (UNMEET NEED), INCLUDING SPECIFICATIONS FOR EARLY-STAGE DRUG DEVELOPMENT

Schizophrenia with Predominant Negative Syndrome (NSS)

- Drugs are needed to demonstrate efficacy on the negative symptoms of schizophrenia.
- Clearly, NS play a role in driving public health needs for schizophrenia.
- (more details will be specified) a name was mentioned of someone with experience in this area.

Schizophrenia with Cognitive Impairment Associated with Schizophrenia (CIAS)

- Cross-cultural performance-based measures and functionality related to cognition are needed.
- Although the gold standard has been accepted by the FDA with MCBB, improvement in efficiency is needed.
- The topic is struggling to ascertain the relationship between cognitive impairment and functional status, i.e., how much functional status is explained by cognitive impairment.
- Methods: Question of whether clinical trials targeting cognition should include "cognitive remediation" in pharmacological CTs. This strategy can provide a sort of baseline for this type of trial. Recently, a white paper with definitions and endpoints have been issued.
- Type of intervention; this definition of "cognitive remediation" is also useful as type of nondrug intervention and mixed with drug CT.

Developmental Disorders (Autism, Down Syndrome, etc.)

- Could be interesting to define specifics on the commonly targeted domains (behavior, cognition, mood).
- OMIs are usually taken from clinical practice using normative data to interpret scoring. Not always applicable in international clinical trials.
- Lack of measurement tools other than those used in clinical settings.

Clinical Insight in Schizophrenia

- Lack of Insight in schizophrenia is the leading cause of relapse, noncompliance with medications, noncompliance with treatment, etc.
- At the core of major public health significance because it leads to noncompliance and nonattendance to treatment and follow-up intervention programs.
- At present, it can be addressed in many effective ways.
- There are no effective treatments for the disorder, and it can be interesting as a potential treatment target indication.
- Lack of consensus around neurobiological perspectives and treatment.
- Importance as few people pay attention to how to measure and address it from a therapeutic standpoint, potentially credit for an indication.
- Not a unitary concept; there are many contributing factors, such as a lack of awareness of ones thinking, lack of meta-cognitive functioning, including a lack of insight, lack of awareness of owns feelings, and failure to recognize other people's feelings. Could be operationalized much more clearly and could then be an indication for an intervention.
- Not specifically for schizophrenia but also in mania, depression, etc., which suggests its trans-nosological use in other indications, such as HTA and Diabetes.
- Like many other conditions, it is a question of penetrance and degree.
- One issue is the measurement of insight and how it can be measured more ecologically. A review study showed that "insight" can be measured thorough PROs and ecological devices. Digital technology can be useful for tracking the insight in a more ecological way and paralleling behaviors in relation to insight. It can be tracked in the community and used as an indicator for relapse.

Oncology – Esp. Leukemia, Bowen Anal Cancer (Post-APR Surgery), Pancreatic Cancer

- No consensus on PRO COSs in most cancers. Combinations of PROs vary widely from protocol to protocol, even within the same types of cancers. PROs are highly sought-after measures in Oncology.
- Work is already being done on COSs for the most deadly cancers and related treatments (i.e., chemo-radiotherapy, gynecological cancers, colorectal, prostate, lung and breast cancers). However, most of these COSs are biomarker-based, with validated PROs virtually nonexistent or in great shortage, with the PROs that are used originating from the hospital/clinic level with questionable/difficult-to-confirm psychometrics (unpublished, other than mentioned in literature as measures used; references pointing to papers also mentioning only use thereof).
- PRO areas being measured pre- and posttreatment are physical function, psychosomatic symptoms, depression, anxiety, pain, behavioral/personality changes, occupational function, drug-induced symptoms and disease-specific symptoms (i.e., pruritus, dysgeusia, ageusia, chemo brain, fatigue, memory loss, nausea, muscle strength, speech, vision, lumps, swelling, weight loss, tingling sensation, etc.). There are validated PROs used in other therapeutic areas that can be adapted for oncology with minimal pilot samples to adjust cutoff scores for cancer patient groups.
- PRO areas The EORTC and PROMIS scales are the most popular PROs used due to their flexibility via built-in IRT-powered platforms. These two sources, particularly PHO's PROMIS/Neuro-QoL, could be used as a starting point for developing OMIs to maintain credibility with oncology researchers and supplemented with other PROs used in psychiatry, neurology, otolaryngology and/or orthopedics.

Borderline Personality Disorder

- There is a lack of consensus on a COS for clinical trials, as highlighted within published Cochrane reviews.

Schizophrenia Treatment Resistant (TRS)

- There is a lack of homogeneity on the definition and measures of TRS across studies. A better definition of targeted domains would be useful.

Rare Diseases (Specify)

- Endpoint and clinical outcomes assessment for R&D studies of rare diseases is a complex and challenging area requiring new methodological approaches.

Parkinson's Disease – Focus on Cognitive and Behavioral Dimensions

- There is consensus on a COS, but it is still pending validation with ecological measures (Digital Health Monitoring).
- Some dimensions need to be disclosed as an effect of the disorder or an effect of the treatment (e.g., impulsivity).
- Additionally, a COS for genetic variations like GB-PD is needed.

Epilepsy

- Drugs with better efficacy-safety profiles, drug-resistant seizures, disease modifying treatments, comorbidities, among others.

Other Conditions with Clear Unmet Needs:

- **Alzheimer's disease**
  - Including behavioral disturbances
- **Lewy-Bodies dementia, Stroke, Huntington's Disease**

7.1.4. Describing "drug targets" and "drug actions": actual/desired

The goal is to clarify what is known, the mechanism of action of the new drug under research and the pathophysiological pathway by which the therapeutic effect is thought to be achieved.

Usually, this information is included in the investigational product brochure and product clinical development plan. Relevant information should be summarized to provide a brief description of the effect of the study drug and its strength of the effect. Along all the COS projects, the therapeutic target should be kept in mind.

In some situations, a COS will target disorders with a lack of consensus on the best molecular target or targets for drug development. However, the mechanism of action of the study drug is at least well known from preclinical and phase I research.

As key considerations, initial assumptions about *disease dimensions*, *diagnostic subtypes* and *patient segmentations* should be considered at this initial project phase. Additionally, assumptions made by the sponsors in terms of the *expected benefits over disease dimensions* and *potential* risks, i.e., considering vulnerable populations are worth mentioning.

For *additional activities,* the initial clinical outcomes and measurement instruments used with similar drug candidates or mentioned by regulatory research guidance documents can be included in this section (as Concepts of Interest, COIs). All ideas and *know-how* from the research team and *expert advisors* can be included to increase the quality of the initial review and take advantage of the know-how of the research team on the sponsor side.

For molecules under development, studies underway to *improve* their *pharmacological profile* and the aspects of the drug relevant for clinical practice that may improve in the future (dosing, formulation, reduction on side effects, etc.) should be mentioned in this section.

In this section, any relevant findings from basic/clinical research about *biomarkers* can be included, which can be useful for different steps. Evidence from EEG, ERPs, fMRI (bold connectivity to function related to drug mechanism), CSF biomarkers, neuroinflammation biomarkers, metabolic or hormonal factors should be mentioned in this section. Biomarkers are often used in clinical practice to diagnose/stage a disease or to predict/monitor the response to therapy (Sullivan EJ, 2012).

Candidates for *surrogate endpoints* should also be mentioned in this section, as should any potential laboratory measure or physical sign that is to be used as a substitute for a clinically meaningful endpoint. Changes induced by therapy on a surrogate endpoint are expected to reflect changes in a clinically meaningful endpoint and any future validation plans with COA instruments should support the expected relationship (mode details can be included in the corresponding section for future roadmaps; *see* Section *10*).

Although it is advisable to focus on one disorder, it is also worth mentioning other potential therapeutic indications and potential overlap effects in terms of dimensions, i.e., trans-nosological

approach. This can be especially useful if one dimension is prioritized over others at the consensus stage.

Not including this section or including too many details might *risks* negatively impacting the entire COS development process. Although substantial information about a new drug under development was historically provided, a summary of the main features is now suggested to keep the COS project in line with the drug mechanism of action. For novel mechanisms of action, such as epigenomics, it is interesting to make an effort to summarize the state of art relevant to COS development.

Additionally, it is worth mentioning that this step is where *translational efforts* need to be described to progress adequately toward the clinical phases of drug development.

To illustrate this, we can comment on the REiNS COS project (COMET-Initiative, Project 722, n.d. REiNS; Plotkin et al., 2013) and the evaluation of the impact of the disorder on cognition. Interesting lessons learned were communicated following negative results in very promising trials with lovastatin and simvastatin in the treatment of neurofibromatosis type 1 and despite very promising preclinical results. A COS used to test the compounds was developed by the REiNS research team. It was highlighted by (Acosta, 2013), a member of the steering committee working group, showing the relevance of good communication between experts in molecular science, animal models, and human cognition.

As it can be applied in many areas of neurosciences, the result of this interdisciplinary approach allows for the development of a well-planned and systematic model to test the possibilities that science is offering.

7.2. Step 2 – Defining the Scope of use for the Outcome Set

The goal is to clarify the context of use of the COS where the study drug will be tested in later stages (phase III).

This guidance document represents a guide to selecting a COS for PoC/Phase II clinical trials in a sample of patients representative of the COU for the study drug, which results in the initial primary focus for developing a COS for clinical research. Considerations related to the application of the COS within clinical practice can be taken into account based on how familiar clinicians are with the specific outcome and its evaluation.

This step involves the specifics about the COU of the final COS if the study drug is demonstrated to be efficient and safe. This decision can be structured at various levels, such as health problems, intervention type, and setting of prescription/administration of the intervention. A complementary and more detailed description of minimum standards for scope can be found in the COS-STAD recommendations (Standards 1–4), as defined using a different classification (for the COS-STAP Group et al., 2019; Kirkham et al., 2017).

Eventually, for drug development, a COU will be used as a primary, coprimary or secondary endpoint in the design of a clinical trial.

7.2.1. Health condition, condition stage, target patient population

The goal is to define the purpose of the COS, including therapeutic indication and intended use, i.e., to validate a pharmacological intervention. Ultimately, this consists of defining the scope of the COS and its applicability.

The final purpose of the COS/OMI set is the driver of the project, representing the final application of the core set. The sponsor needs to identify and describe interest in the applicability of the new development as much as possible.

Key considerations include being specific regarding *health problems* and their place within the health system or along the natural course of the condition.

At this stage, it is also relevant to describe the level of variability in the expression of the disorder expected across subjects, since this will have an impact on future decisions, such as the number of participants in the consensus process.

Even for final decisions on domains, it is relevant to take into account health priorities, if any. In this line, for some interventions, it might be advisable not to choose too many domains and instead prioritize the most relevant domains.

Regarding *risks,* for disorders with a high variability in the expression of symptoms, the final COS may not completely represent all the patients suffering from the condition.

7.2.2. Defining the intervention

The goal is to specify how it the new drug under research is expected to produce its effect.

For this definition, the therapeutic goal must be specified, for instance, symptomatic, etiologic, preventive, and natural courses, as well as, for instance, acute phases, chronic maintenance periods, early stages of the disorder, prodromal phases, first episodes, etc.

7.2.3. Defining the clinical setting and context of use

The goal is to specify the therapeutic settings in which the new intervention will be conducted as much as possible. An international scope should also be taken into account, considering different health systems.

The context in which decision-making is expected to happen as part of the setting, for instance, in emergency wards, mental health settings for outpatients, and inpatient settings for chronic patients or postacute wards, should be described. The following examples will help sponsors understand the project COU definition; not all the listed will apply, but they can help in the specific context of drug development (FDA, 2020a; Powers et al., 2017).

- Disease definition, including pathogenesis or disease subtype if appropriate.
- The targeted study population, including a definition of the disease and selection criteria for clinical trials (e.g., baseline symptom severity, patient clinical and demographic characteristics, history of previous treatments, language(s)/culture subgroups).

- Identify the targeted study design and objectives.
- Identify the endpoint model and definition and positioning (i.e., planned set of primary and secondary endpoints with testing hierarchy) if known.
- Relationships (known and hypothesized) among all clinical trial endpoints, both COA and non-COA.
- Analysis plan - Hierarchy of all COA and non-COA endpoints intended to support claims corresponding to the planned data analyses.
- Targeted labeling.
- Study setting as impatient vs. outpatient, geographical localization and clinical practice variation.

7.2.4. Describing the final COS application in a "nutshell"

The goal is to summarize the application of the COS within the context of intervention in early-stage RCTs.

The final application of the minimal COS should be expressed in a short sentence or short paragraph as a "nutshell". Examples of such descriptions regardless of the therapeutic field can be found in public databases. For educational purposes, we classify a set of COS projects according to the mentioned rule:

ⓘ Example of projects with a *specific* goal:

- Improve_LTO Core Outcomes Measures for Patient-Centered Clinical Research in Acute Respiratory Failure Survivors.
- METRICS initiative To develop an instrument to measure cognition to support the development of new pharmacological approaches to improving neurocognition in schizophrenia.
- CHOICE The aim of this study was to select a COS to be used in evaluative research of interventions for children with Rolandic epilepsy.
- MS Core outcome set for relapsing type in adults (COMET Project Lucchetta R C, n.d.).

ⓘ Example Projects with a rather *generic* goal:

- PD proposes a global consensus standard set of outcome measures for idiopathic PD (Project COMET-Initiative, 2017; de Roos et al., 2017).

- Bipolar Disorders develops a core outcome set for use in community-based bipolar trials (COMET-Initiative, 2017; Retzer et al., 2020).

There are some *risks* related to being too specific at this stage, as it might increase the sensitivity of detecting changes while decreasing the external validity of the COS in various clinical treatment scenarios. However, the sponsor needs to balance both risks, decreased validity and decreased efficacy detection in the expected targeted domains.

For new areas of research, this step is critical, especially in areas involving innovation in health and promoting changes in disease management in clinical practice. The definition should forecast different geographical areas that might have different clinical practices for either the medical specialties in charge of the disorder or disease management in private/public practice.

- ⓘ To illustrate this, it is worth mentioning attempts to innovate the treatment of cognitive impairment related to schizophrenia (CIAS), which is a clear unmet need that is quite neglected in clinical practice. A new treatment for this indication would include a change in the way the disorder is managed within the health system and private practice.

- ⓘ Another example can be found in the treatment of patients with predominant negative symptomatology, who quite often stay home without attending medical appointments.

Adopting an inadequate definition of the unmet need and the COU of the new treatment might prevent finding the right patients either during clinical research or when the treatment is on the market.

7.3. Step 3 - Stakeholder involvement

The goal is to be explicit on what categories of stakeholders should be considered (patients, public, practitioner, press, policy maker, program manager, professor, payer) and on the minimum requirements for consensus on the specific COS project.

The list of potential stakeholders to consider will usually depend on the final use of the COS. The decision will impact either the validity of the final domain set or the final effective use of the instruments in clinical research (Beaton D et al., 2021; Prinsen et al., 2016).

A potential full list of stakeholders:

a. Patients and health care consumers
b. Researchers (and within this, methodological or content expertise in multiple areas such as outcome measure development, biostatistics, psychometrics, qualitative studies, comparative effectiveness, and clinical trial design)
c. Clinicians (e.g., physicians, surgeons, nurses, psychologists, physical and occupational therapists) and health care providers (meta-level institutions and health care systems)
d. Research funders (government, foundations, benefactors, scientific societies);
e. Government regulatory authorities
f. Health care policy groups, including representatives of various levels of health systems;
g. Pharmaceutical companies or device manufacturers
h. Clinical trial management companies (e.g., CROs) or scale management agencies for licensing and rater training service providers
i. Family members and caregivers—either formal or informal—tutors, monitors,
j. Foundations and patient/health care advocacy groups, patient representative groups;
k. Payers (government systems, insurance/managed care benefits managers).

7.3.1. Key stakeholders

The goal is to identify stakeholders that must be included in a consensus COS definition.

According to COS-STAD recommendations, a consensus should be reached among those who will use the COS in research, as well as among health care professionals with experiences with patients with the condition and patients with the condition or their representatives (including family members and carers), named U, Exp and Pats, respectively, by COS-STAD (Kirkham et al., 2017).

The OMERACT Handbook suggests that a consensus concerning outcomes to be used in a clinical trial must include those *who perform clinical trials, methodologists, pharmaceutical companies, clinicians, regulatory groups, and patients* (Beaton D et al., 2021).

As seen in the previous sections, there are many other parties to consider, and the minimal parties that should be included may vary from project to project. Asking a question such as "Who has the necessary expertise and content knowledge for our project?" can help construct an initial list of the parties that should be involved. Asking "*In what settings and populations are these outcomes intended to be used?*" may provide additional stakeholder groups to consider.

As *key considerations,* the patient groups included in the panel can be patients in an active role, e.g., as participants in a trial or via an interview or survey, or as patient research partners (PRPs) as collaborative partners.

Including *regulators* from the beginning of a project can be key to ensuring the final acceptance of the COAs/OMIs for drug development, which is the ultimate goal. Their participation might be limited by institutional regulations, so they are not as active as other participants. Therefore, representatives from various agencies in the US and the EU or APAC are desirable to capture all views to ensure external validity beyond the current project. For a deeper description of the relevance of regulatory participants in the project, see (Tunis et al., 2017). In addition to providing guidance from the regulator perspective, they can also contribute to boosting scientific advances in the field, accelerating the use of innovative solutions/proposals.

- ⓘ For the MSOAC project in MS, CDER was partnering with an existing consortia, the Critical Path Institute Multiple Sclerosis Outcomes Assessments Consortium (MSOAC), a collaboration in a precompetitive setting for developing COAs suitable for drug development (LaRocca et al., 2018).

- ⓘ Outside neuroscience, the COS developed for clinical trials on hemophilia is an interesting case example that illustrates the engagement of regulators in the process, followed by the clinical use of the COS produced (Iorio et al., 2018).

For *additional activities,* based on advances in digital health and technology, it would also be advisable to include *IT experts* as part of the panel to gather opinions on the technological aspects of future OMI developments and its boundaries. Questions related to digitability or digitalization (remote assessment formats, momentary assessments, digital health monitoring, etc.).

The *minimal activities* involve setting the rationale for choosing stakeholder groups, eligibility criteria, and identification strategies in advance in collaboration with the research team. For each stakeholder group, specific recruitment strategies, e.g., direct, personalized contact, indirect contact via websites, or mailing lists, will be used. This might involve other partners and service providers in the group to reach the potential participants.

The eligibility criteria for the Delphi panels for each stakeholder group can also be further specified based on the availability of potential participants. Examples of inclusion/exclusion criteria for members can be found at Fackrell et al., 2017.

OMERACT manuals recommend starting with at least 100 participants per stakeholder group, as Delphi results must be stratified by patients versus other stakeholder participants to identify any differences in the domains considered important. We suggest that working groups strive to have a minimum of 30 to 50 participants in each of the 'patient' and 'other stakeholders' groups. Because there are diseases/conditions for which it is very difficult/impossible for patients to participate in Delphi studies (e.g., mild cognitive impairment, early-stage cognitive deterioration), these figures must be tailored to the case of interest.

The following section presents the use of IA/ML at several COS project steps; specifically, this step allows us to have a wide representation of the patients and to better understand the heterogeneity of the disorder.

7.3.2. Artificial Intelligence and Machine Learning

A rapidly evolving approach to the development of COAs/OMIs is to apply artificial intelligence (AI) and ML to data from clinical trials. Most neuropsychiatric diseases represent clustering of related, but heterogeneous, constructs whose pathophysiology, distinguishing phenotypic characteristics and relationships to outcome measures are incompletely understood. Due to this construct heterogeneity, responses to treatment are likely differentially observed for the component subpopulations, and the relevancy of PROs/COAs may vary according to the subgroup.

Furthermore, regulatory authorities use the experience of patients to make decisions on whether a treatment entity's benefits outweigh its risks. PROs and COAs are the most common sources of this patient experience data. Regulatory decision-making requires science-based data. However, the cost of performing clinical trials that address all of these considerations is prohibitive. Even if funding were available, patient availability, time and other resource requirements would make the necessary studies impossible to complete using traditional methodologies. However, methodologies based on the incorporation of AI and ML are suited to addressing the multifactorial problem of identifying robust outcome measures that personalize the clinical value of treatments for these heterogeneous disease constructs.

Recent advances in AI, specifically ML-based approaches, have made these techniques more accessible and acceptable. They have been adopted by the biopharmaceutical industry and used to target validation, computational chemistry, and drug repurposing. Methods such as ensemble trees (e.g., boosted trees and random forest trees), support vector machines, and deep neural networks are among the most popular methods being used. Although many specific ML methods

are available, the essence of ML algorithms is that, in some sense, they program themselves. This means that the algorithms 'learn' about features of a data set that have not been manually programmed for the task. Instead, models emerge from their interaction with data, identifying patterns and relationships that would otherwise be extremely difficult to identify.

Success in the application of AI/ML for understanding subpopulation outcomes is now well established for oncologic diseases (Nagy et al., 2020). These same approaches are being applied to datasets on neuropsychiatric diseases. Although it is often assumed that large datasets are necessary to reap the benefits of AI/ML, methods have evolved such that small datasets can be used for initial hypothesis generation and validated in larger trials (Geraci et al., 2018).

The use of AI/ML raises challenges for regulators and clinicians. The AI/ML risk-benefit relationships from patients with specific neuropsychiatric disorders that were initially proposed for recognition must be validated. Thereafter, as larger data sets are obtained from more diverse populations with the disorder, additional relationships may be identified, and earlier relationships may be revised. Regulators must establish the validity of the particular AI/ML method that has been applied as well as the finding. Once verified, regulators must find ways to efficiently and effectively communicate these findings to treating clinicians. As AI/ML becomes more firmly rooted in the healthcare system, such updates could be frequent.

7.3.3. Balancing stakeholders

The goal is to clarify the composition of the panel groups regarding the variety of stakeholders and the number of participants for each stakeholder and whether a specific balance is needed.

A single homogeneous panel approach will result in core outcomes deemed essential by only one stakeholder. If multiple stakeholders participate in the panel, the process can take into account the proportion of stakeholders participating by weighting the participants used for different groups. The weights, i.e., percentage of each stakeholder group can clearly favor one opinion over others. Then, recruitment of participants, whenever possible, needs to be balanced across stakeholder groups.

As *key considerations,* in areas where different stakeholder opinions are expected, it is advisable to forecast the situation and describe the strategy to follow. Options could be to set different COS for different stakeholders or set different COSs for different types of interventions. To illustrate this last option, we can comment on the COMIT'ID project focused on subjective Tinnitus, a condition experienced by 120 million people in the US and EU (Fackrell et al., 2017). For that condition, treatment options remain palliative rather than curative, and judgments about therapeutic benefit typically concern a relative improvement and not simply a binary 'yes/no' or 'present/absent' decision. At the time when a COS was started, there was no consensus on the critically important domains of tinnitus. Following a systematic review of the literature regarding "interventions", the research group realized that pharmacological interventions were mostly experimental and not part of standard clinical practice. The final strategy was then to establish three core outcome domain sets, one for each of the main intervention strategies (sound-, psychology-, and pharmacology-based) and to identify the key outcome domains that are

common across all three interventions (Fackrell et al., 2017). In the field of psychiatry, initiatives are starting clinical trials, including integrated interventions, in which groups of participants receive two or more interventions. Therefore, considering all possibilities, the strategy to tailor a COS for different contexts can make sense, as is common in other medical contexts.

The *risk* of having representatives of just one or two groups of stakeholders, i.e., drug developers and medical advisors, is that the opinions of other stakeholders may be missed. Additionally, listening to other groups allows us to identify outcomes and endpoints that may not be core disease manifestations in patients' daily lives or the ones most bothering for the patients or for the health system, where the new intervention might have an effect.

As proposed in the OMERACT Handbook, a challenging question following a COS project is "*Would it make a difference if there were wider participation?"* (OMERACT, 2018, p. 4, Ch 1),

7.4. Step 4 - Determining "what to measure" as dimensions and concepts of interest

The goal is to identify all that can be relevant to measure to later decide the most relevant ones.

In this section, we will start with an initial extraction and organization of the dimensions considered relevant for a final list as a result of the review done thus far. A description of the main questions related to the review of the existing literature can be found in the corresponding section of this guidance document (*see Section 7.1*).

However, the initial search and review may not produce any results in terms of previous COS activity or any existing *disease impact models* created by other researchers in the field. Disease impact models may exist but quite often are not ultimately published.

In addition, available measurement instruments can also be used to pick up dimensions/concepts and to help with the definition.

A complementary source of "dimensions" can also be found by analyzing existing measurement instruments (COMs) that can be added to the previously identified instruments. The sources of existing and accepted OMIs are as follows:

— Public/private libraries of OMIs, i.e., tests, scales, inventories, questionnaires, such as BiblioPro (IMIM, n.d.), Banco de Instrumentos CIBERSAM (CIBERSAM, n.d.), ePROVIDE (MAPI Research Trust, n.d.), Health and Psychosocial Instruments (HaPI) (EBSCO, n.d.) and SRLab (SRLab, n.d.). All these libraries are available and provide information about the instrument source, available translations and complementary data, as well as the author, domains and COU or psychometric properties. The majority provide information for free and offer additional information under subscription fees or charge registration and/or subscription fees.

— Scientific societies quite often share measurement instruments recognized in the field through their websites (International Parkinson and Movements Disorders Society, (MDS, n.d.), Alzheimer Disease Cooperative Study –ADCS (ADCS, n.d.), etc.).

— Scientific institutions producing instruments from item banks, such as the PROMIS initiative (PROMIS, n.d.).

— Recommendations made by clinician experts in the field are always a source of valuable information on what is being used in the field for clinical purposes.

Last, when no guide exists or what exists is not sufficiently validated, the use of *clinical observation/experience* to help conceptualize the domains will be one of the sources of information.

Additionally, drug developers often work in just *one broad domain* (e.g., cognition, functional impact, or apathy) based on the known/expected drug mechanism of action. Although the work is limited to one domain, the decision can also be complex and require the same or a similar work process to that needed when more domains are included. The goal is not just to determine the effect of an intervention on cognition but to contextualize as much as possible the definition of this global domain.

It is also important to emphasize that core outcome domain development/selection must be clearly separated from core outcome measurement instrument development/selection. Based on experience, the COS group project at COUSIN-CS notes that if there is no clear distinction, the entire process becomes very complicated because it is always mixed up with WHAT and HOW questions.

By always thinking about the HOW, it is impossible to answer the WHAT question appropriately. It is of course always useful to keep in mind how constructs/concepts are measured in a certain field, but this is not helpful for the WHAT.

In psychiatry, if there is a clear list of priority core outcome domains for certain diseases (e.g., schizophrenia), then the definition of the domain already contains whether these domains come from the patient perspective, a biomarker, a physician or a proxy report. This probably also depends very much on the disease, population, and intervention, but it is important to concentrate on the domain/concept as a real first step regardless of how it will eventually be measured.

7.4.1. Defining the method for identifying dimensions/outcome domains

The goal is to review and list all key areas that need to be represented as domains/concepts in the core domain set.

To extract outcomes from the academic literature or existing OMIs, we need to consider that some outcome domains will be identical but described in different ways. Domain extraction is an iterative process of classifying different outcome domains and summarizing similar outcome domains. During the review activity, once the outcomes/dimensions start to collapse into the same ones, the review activity can be considered complete. The activity can continue if more accuracy is feasible but the minimal quality would likely have been reached. Saturation has attained widespread acceptance as a methodological principle in qualitative research. It is commonly taken to indicate that, on the basis of the data that have been collected or analyzed hitherto, further data collection and/or analysis are unnecessary (for specific saturation models *see* Saunders et al., 2018).

With the collected materials, the following activities will improve the overall quality of the COS project:

- Group different definitions of the same outcomes/dimensions together (extracting the wording description verbatim) under the same outcome name.
- Group these outcomes into single outcome domains or constructs that can be used to classify broad aspects of the effects of an intervention.
- Create a final list of outcome domains and define them
- Build a disease impact model to have a full picture of the dimensions and the context of the disease.

In addition to this well-known strategy, there are other approaches that are used in psychiatry and neurology that consist of predefining criteria for the selection of domains that aim to target the intended use of the domains with the domain selection process upon the start of the process *(see Section 0* for more details*).*

7.4.2. Organization and structure of domains

The goal is to organize the preselected domains/concepts of interest in a taxonomy or theoretical model useful to confirming the coverage of relevant aspects of the disorder.

Usually, once the various candidate dimensions are listed, they serve as to compose the starting list of domains to present to the panel.

It is advisable to provide a *structure to the domains* listed to support the conceptualization of the impact of the disease in a comprehensive model. Even if a model of disease impact exists in the literature (*see Section 7.1*), it is always interesting to cross check for consistency. If a disease impact model does not exist, it should be built, as this will ensure that all relevant areas of impact will be covered with the list of domains and concepts, helping identify gaps or duplicities.

There are several models/taxonomies for classifying outcomes to help with this activity. These can also be useful in neuroscience, and there several in the literature (Beaton D et al., 2021; Williamson et al., 2017). The activity here is to select the model from the existing ones that best matches the theoretical assumptions and place the dimensions on the different categories of the model.

Examples of such models are the model proposed by the WHO, which includes three broad health domains (physical, mental and social wellbeing) (WHO, n.d.); the OMERACT group model, which organizes the dimensions into core or more distal to life impact and pathophysiological manifestations; and the Cochrane Review outcome framework. Researchers must choose the model that best matches their interests (*see COMET Handbook Section 2.7.3. for a concise review of the existing models*) (Williamson et al., 2017) and OMERACT Handbook (Beaton D et al., 2021; Beaton et al., 2019).

In the case of PerfO, the "concept of interest" should be aligned with the model proposed by the WHO-ICF provided that the interpretation of the result of the PerfO measure reflects an important aspect of patient functioning (Richardson et al., 2019). However, this translation to "real life" is not always feasible for several reasons (*see* Section 7.5.2.2.1.2 for a more detailed discussion).

In this guidance document, we can recommend one of the latest models developed by Dood (2018), which is useful for all diseases and can be used in the neurosciences. It represents a taxonomy developed for outcomes in medical research to help improve knowledge discovery (Dodd et al., 2018).

In the psychiatric field, we have the NIH Research Domain Criteria (NIH-RDoC) initiative. It developed a research framework for investigating mental disorders with the domains organized in a matrix of elements including a multilevel assessment path for specific domains (Insel et al., 2010).

RDoC definitions provide a taxonomy of six major domains of human functioning and focus on the investigation of targeted biological, physiological, and behavioral elements that comprise mental health (NIMH, n.d.).

- ⓘ This model has been used in a recent project focused on building diagnostic criteria for a specific dimension, i.e., apathy in dementia. The International Society for CNS Clinical Trials Methodology Apathy Work provided a framework for defining apathy as a unique clinical construct in NCD for diagnosis, providing a consistent definition of apathy for further research on drug development (Miller et al., 2021). Future directions for that project will include the operationalization of the consensus-based criteria, validation in both research and clinical settings, and development of new or validation of existing assessment scales. Other examples of RDoC application will come with future advances in drug development using the Research Domain concept for personalized psychiatry.

- ⓘ An illustrative example of its use is the RTOC consortium, which aims to optimize and validate reward processing domains to support drug development tools in schizophrenia and mood disorders (Bilderbeck et al., 2020).

Specific disease taxonomies are also being created for some degenerative disorders and will be ready soon, and these will offer an additional option.

- ⓘ One example of such taxonomies is the IMI-AETIONOMI project (on behalf of the AddNeuroMed Consortium et al., 2021), which looks for molecular characteristics of Alzheimer's disease (AD) and Parkinson's disease (PD) and is contributing to a 'taxonomy' of these conditions and helping move toward a precision-medicine approach.

The final framework/taxonomy used in a specific project will ultimately help in the selection of the COS for each category and define the categories covered. It is advisable use references to better understand of the model to be used.

7.4.3. Starting from predefined domains or concepts of interest already known

The goal is to identify whether a useful strategy already exists to select domains or concepts to be considered for COS selection. This step is not always applicable, but it can be interesting to explore previous assumptions in the disease impact model, instead of starting from scratch.

We can find some past COS projects using a different approach, i.e., they used a strategy to predefine criteria that will ultimately be used to construct the final domain list.

In such projects, the research team agrees on a list of criteria that needs to match the interest of the COS development, specifically its final use. We can find examples of this approach within the cognitive field in drug development.

> ⓘ For instance, in 2004, the NIH-MATRICS working group elaborated (Green, Nuechterlein, et al., 2004) a consensus battery started by identifying dimensions of cognitive performance that were implicated in the past in schizophrenia and relatively independent of each other. For this selection, they identified six separable factors that were replicated in multiple studies of patients with schizophrenia and were appropriate for a consensus cognitive battery for clinical trials: 1. working memory, 2. attention/vigilance, 3. verbal learning and memory, 4. visual learning and memory, 5. reasoning and problem solving, 6. speed of processing and 7. social cognition. The subsequent step of that project was to define a list of criteria to select tests measuring the final list of domains *(see* Section 7.5*)*.

The MATRICS method actually inspired other COS projects that applied a similar methodology to identify outcomes in other conditions, such as neurofibromatosis clinical trials (Walsh et al., 2016). Another example in the field of neurology, using a similar predefined strategy, is the MSOAC, but uses the WHO-ICF classification (WHO, n.d.) of the impact of MS in patients, actually MS-related disability, as a reference. From this, the MSOAC group defined the concepts of interest (COI) for meaningful treatment benefits. This approach helped define a list of domain selection criteria:

Figure 1. List of Selection Criteria for ICF Domains in the MSOAC multiple sclerosis project (LaRocca et al., 2018)

Inclusion criteria	Exclusion criteria
A domain must represent something related to MS that affects a significant proportion of people with MS.	The domain is not thought to relate to activities, functions, or roles that are important to people with MS in their everyday lives.
A domain must be something that can be measured objectively and that does not rely entirely on patient-reported symptoms.	The domain is not commonly affected in people with MS (e.g. hearing).
A domain must be something that can be measured easily, with minimal equipment, and in a reasonable amount of time.	The domain does not change over time or vary depending on MS severity.
A domain must be something that affects a real-life function that is meaningful to a person with MS.	The domain cannot be objectively assessed (e.g. fatigue or pain).
A domain should preferably be one for which accessible data exist from MS clinical trials.	The function related to the domain cannot be quantified or cannot be measured using practical test procedures (e.g. sexual function).

7.4.4. Setting of Candidate Domains

The goal is to build an initial (long) list of outcomes covering each core area, which is the starting point for consensus.

Some *activities* can be completed before presenting the list to the panel to ensure the final result:

a. Decide on the extension of the list of candidate outcomes.

b. Provide a definition of each outcome. Definitions in terms of health vocabulary can be accompanied by sentences from patients' expressions extracted from qualitative searches. A detailed *definition* is needed so that others can clearly understand the domain. It will also serve as the checkpoint for content validity.

c. Sort the list based on priority or level of relevance, which can be done thorough a consensus process, i.e., hierarchically structured outcome domain list. This can be done thorough expert panel meetings or a Delphi survey.

d. Agree on which domains to include in the COS through consensus among panelists using a predefined criteria-strategy.

A *key consideration* is that the usefulness of a core domain set strongly depends on it containing only a small set of core domains since a core domain set aims to capture the minimum number of domains necessary to adequately capture what we want to know, as these will be required in all relevant studies of the benefits, harms and costs of the interventions of interest. There is some evidence that 7 ± 2 individual items are advisable.

Drug developers often work in only one domain (e.g., cognition, functional impact, apathy), which can be complex and require the same or similar work to that required when multiple domains are included.

The COMET Handbook also highlights the relevance of setting the *timing of assessments,* which can be determined by the trialists and the particular context of application. However, in some diseases with clear clinically specific outcomes, it can be relevant to set short- or long-term periods of measurement, e.g., when outcomes are linked to time, as mentioned in regulatory or scientific guidelines.

A risk worth mentioning related to this process is that whenever additional domains appear during the Delphi survey, it is recommended to establish different outcome categories, including "Mandatory or core domains," other "*Important but optional domains*" and domains to be kept in the research agenda, in order to be candidate domains either in the first or in the second group.

7.4.5. Delphi Technique

The goal is to use a method for achieving consensus among the panel of participants regarding the final core domains.

Although other methods exist, the Delphi technique is one of the most popular and efficient methods used to reach consensus among participants/stakeholders. However, the Delphi are generally highly dependent upon the composition of the panel.

The method used to determine consensus and rationale must be predefined. Other methods described within the manuals are the nominal group technique, consensus development conference and semistructured group discussion, or a combination of methods. The Delphi technique is advantageous in that it is anonymous, avoids the effect of dominant individuals, and can be circulated to large numbers with a wide geographic dispersion.

There are some *key* considerations when planning this type of consensus session because it might have an effect on the final decision (adapted from the COMET Handbook) (Williamson et al., 2017):

1. Number of panels
2. Group size
3. Participant information
4. Accepted/non accepted conflicts of interest
5. Number of rounds
6. Structure of the questionnaires
7. Methods of scoring
8. Nature of feedback presented between rounds

9. Criteria for retaining outcomes between rounds
10. Attrition (response bias) between rounds
11. Consensus definitions
12. How the degree of consensus will be assessed
13. Criteria for retaining participants up to consensus meeting
14. Compensation for participation, if any

From a list of candidate domains, the participants *select* those domains they believe to be of critical importance for inclusion in a core domain set, and this process is conducted iteratively over two or, more typically, three rounds until agreement is reached, often with an option in the first round to suggest additional domains. A standard question is "*Please rate how essential you think it is that these outcomes are measured in clinical trials*", and the participant rates each outcome from the list from 0 to 9. The results can be presented as a summary for each stakeholder group.

The participants are asked to score each outcome domain using the GRADE scale of 1–9, where 1 represents least important and 9 represents most important. Options 1–3 indicate that the domain is 'not important', while 7–9 indicate that the domain is 'critically important' in deciding whether a treatment for the specific disorder is effective. Scores of 4, 5, and 6 indicate that the outcome domain is "important but not critical" (from COMIT'D project in Fackrell et al., 2017).

Following prespecified rules, items are retained or dropped between rounds.

The number of domains is predictive of the number of people who withdraw from Delphi surveys; a higher number of items results in a lower response rate; therefore, OMERACT suggests no more than 70 candidate domains in the initial Delphi round.

The background information explaining the rationale for the Delphi should be prepared with consideration that it may be appropriate to tailor information to specific stakeholders (e.g., patients may require more information explaining the study concepts). This is the preconference or pre-Delphi reading material, which should be distributed a minimum of 1 month before the meeting or Delphi round (*see details in OMERACT Handbook* (Beaton D et al., 2021)). A pilot testing of the background information and initial Delphi questions with a small group prior to full implementation of the Delphi is also recommended (*see details on the COMET Initiative Manual in* Williamson et al., 2017).

> ⓘ One example of the application of the Delphi method is described in the CHOICE project paper with a comprehensive description of the method applied by this group to set domains for epilepsy (Crudgington et al., 2019).

7.4.6. Determining the final COS

The goal is to agree on the final COS based on consensus as a final vote from the entire group of panelists.

Within the manuals, it is recommended to consider those domains that reached, for example, 70% (or greater) approval by all stakeholders the core domain set (score 7–9), which ideally would represent all categories considered relevant to the disease impact model taxonomy. There is no reference standard for these consensus rules, and it is recommended that stricter or wider rules be used depending on the purpose. Existing handbooks and FDA Guidance 3 recommend measuring at least the core disease-related concepts. When measuring disease impacts, the FDA recommends targeting disease impacts that result from the core disease-related concepts (*see* FDA, 2020a, pp. 12–13). Additionally, the selection of concepts for a given trial program should be informed by consultation with various stakeholders.

OMERACT group also issued materials to guide and train on the process of consensus on the core domain sets (*See* Maxwell et al., 2019).

Minimal activities: The final consensus process can be reached during a ***consensus meeting*** that includes the most engaged stakeholders. ***Even votes can be stratified and shown separately by stakeholders.***

The process for reaching consensus on a COS will depend on the availability of panelists and on the resources from the research team side. Models and processes can be selected from the existing literature, ranging from minimal to more rounds, panelists and processes.

7.5. Step 5 Determining "how to measure" outcomes

The goal is to identify at least one outcome measurement instrument (OMI) per domain to compose the core outcomes measurement set.

At this stage, existing instruments measuring the target dimensions would have been identified, and some additional instruments would have been suggested by the panelists during the consensus activity. Here, the idea is to group all identified OMIs and analyze the accuracy and feasibility of use in the desired context.

Key considerations: It is advisable to describe the plan for choosing OMIs to cover the evaluation of the agreed-upon COS. A communication plan including all the participants in the final decision is also key to avoid go backwards in the decision process and have all relevant stakeholders/departments updated and with an opportunity to participate on time.

Minimal activities: The relevant activity is to decide upon the instruments based on objective information as psychometric features relevant to our purpose. A guidance document on how to select OMIs has been developed by COMET and COSMIN based on a survey of a large sample of

experts. Through a consensus methodology, they reached agreement on the main steps to follow (Prinsen et al., 2016).

The process can be summarized in 4 steps: 1) conceptual considerations, 2) identification of existing instruments, 3) auditing the candidate instruments, and 4) final decisions. The main features of these steps will be commented upon in the following sections.

Risks: Completed COS projects have used this decision process with a high variability of methods. There are projects that selected OMIs without a deep analysis of the psychometric properties, while others used subjective surveys to reach consensus on the quality of the instruments. It is also a fact that psychometric properties are not always easy to understand for stakeholders without an adequate background.

For that reason, tables summarizing the validity of the instruments can compensate for this weakness by providing a final "ease to read" summary to the reviewers.

Ultimately, by using instruments that are not well validated will add uncertainty to the validation of a new drug. All the information collected about the instrument might be useful for developing a validation strategy, even precompetitive actions, before its use in a real clinical trial.

As a consequence of this analysis, it is quite frequent that no available or validated instruments exist to measure our targeted domains, and more information and more details are provided elsewhere in this guide (see *Section 11).*

7.5.1. Conceptual definitions of dimensions under consensus

The goal is to build an accurate definition of the dimensions that will facilitate the process of finding suitable measurement instruments.

All work done thus far will easily allow us to define the construct to be measured and other important requirements for an ideal measurement tool. These features are basically the relevant requirements for application in the target population and COU, i.e., age, setting, disease stage, etc., but only those affecting the measurement, otherwise being too restrictive risks discarding all existing features.

7.5.2. Identifying existing instruments

The goal is to identify existing validated instruments as candidates to measure the dimensions.

At this point, many of the available and existing instruments might have been identified; however, if this is not the case, there are various methods to follow, as mentioned in *Section* 7.4.

In any case, there are advantages to performing a comprehensive review; however, gathering too much information is not always the best option. In general, the minimal number of instruments that should be retrieved are those known in clinical practice and expert academic groups in the

research field. A comprehensive list of instruments can include instruments out of date or discarded by experts for specific reasons or even due to changes in diagnostic criteria.

However, old instruments have accumulated more background information, with more validation data available for use in various contexts; therefore, an instrument's being old is not a reason to discard it. It can be a good choice if no other extant instrument has been validated.

It is important to ensure that all the most updated tools are included within the option list, since systematic or deep literature reviews will extract an important research using legacy instruments with little attention to innovative tools due to the low impact of citations when the review is done. To produce a list of pros/cons for each outcome and instrument can be useful when making the decision.

7.5.2.1. Defining the method for selecting instruments

The goal is to set the strategies to be used to identify existing OMIs from the potential ones based on their pros and cons.

One of the first questions at this point is whether to conduct a deep literature review or to determine what instruments are most used in the field, supporting the decision on the frequency and experience of use. The decision will depend on the therapeutic field and availability of experts.

In addition to identifying instruments, it will also be necessary to determine their validation properties when the instruments are used in the target population, which will be described in more detail in the next section.

For that reason, with a limited list of OMIs, a literature review focusing on available validation data can be conducted *(see Appendix 3 for proposed search filters).*

The idea is to identify published studies where the OMIs used psychometric property validation data, including in the targeted disease.

If nothing exists to evaluate a specific outcome/domain, another option is to search item data banks. Content validity and other psychometric features for the final items selected also need to be confirmed before its use (*see Section 11).*

7.5.2.2. Quality assessment of the psychometric properties of key instruments overall and in the COI and COU

The goal is to decide which psychometric features will be included in the review of the instruments under evaluation.

To evaluate the quality of an instrument, i.e., measurement properties, it is important to start thinking about the properties that this instrument must have based on its role in the study design, including potential repeated administrations.

In this line, *Table 1* presents a list of measurement properties that can take part in the decision process for a measurement instrument.

The instruments that will be candidates as FFP COA tools are basically content validity, reliability, construct validity and ability to detect change (*see* PFDD (FDA, 2020a)).

There are some features that are always relevant in drug development, such as *content validity*, *reliability*, and *sensitivity to change* (responsiveness), while others might depend on the study design and on how the instrument will be used as an endpoint in the trial. For additional comments on the unitary nature of validity across COAs/OMIs, see online at the COA Group platform.

To start, it is recommended that the relevant properties for the planned clinical trials be considered and that some justification for the others be provided if not needed. The next step will be to collect relevant information regarding the selected properties from published literature, usually validation studies, tabulate them and then make a final decision. For new instruments that are not used frequently in the field, only one validation paper may be available. However, when several publications exist for the same instrument, it is helpful to organize the information based on psychometric properties to summarize the level of evidence to support each validation parameter for the OMI data.

Table 2 is a model proposed to summarize the evidence on the quality of instruments needed for a single OMI, considering that not all the properties listed are always relevant. This table helps consider all the potential features and decide which ones apply to the specific research. Tables 3–4 can be useful for summarizing the psychometric properties when several OMIs are eligible to cover the same construct/dimension.

There are checklists available that can be useful for a deeper qualitative analysis, although the majority address only PRO measurement types. With all the information about a single OMI identified, there are tools available to guide the evaluation of OMI quality. The EMRPO checklist (Valderas et al., 2008) is a tool for the standardized assessment of PROs to assist in the choice of instruments. COSMIN (Mokkink et al., 2010) has also developed a checklist to determine which measurement properties are important and standards for how to evaluate them. These tools will facilitate the selection of the most appropriate PRO measure among competing instruments. In the future, they will also have a tool for non-PRO instruments.

TABLE 2 MODEL LIST OF BASIC INFORMATION RELEVANT TO A CANDIDATE INSTRUMENT TO BE USED IN CLINICAL TRIALS.

Relevant information	*Relevant to the COS or CT Project? Yes/No/NA*	*Relevant comments*
Background		
• Therapeutic indication for development		
• Purpose of its development		
• Population for intended use		
Reliability		
• Internal consistency reliability		
• Test-retest reliability		
• Interrater or inter-interviewer reliability		
Validity		
• Evidence for content validity/Face validity		
• Construct validity		
• Item-scale relationships (e.g., factor analysis, multitrait analysis, item total correlations)		
• Item-response theory analysis (RASH)		
• Floor and ceiling effects		
• Concurrent/convergent validity		
• Divergent/discriminant validity		
• Known group/clinical validity		
Validity for use in interventional clinical trials		
• Ability to detect change/responsiveness		
• Validated MDC, MID or MCID		
• Specificity/sensitivity in the context of use (if available), e.g., stratification factor, selection criteria, etc.		

**If the information is not required or not available it is desirable to state the reason.*

eCOA, electronic Clinical Outcome Assessment, MDC, Minimal Detectable Change, MID, Minimal Important Difference, as relevant change intrasubject between time-points, MCID Minimal Clinically Significant Difference between treatments arms (within groups) that has clinical relevance in the patient's management, widely recognized as a key concept in differentiating among outcomes of treatments.

TABLE 3 PROPOSED LIST OF INSTRUMENT FORMATS

Proposed list of potential formats requiring validation before its use	
OMI format description	***Validation***
• eCOA format migration *(examples provided by FDA are for instance IVR, web-based platform, mobile apps, pen, tablet)*	
• Remote assessment	
• Centralized administration	
• Validated structured/semi structured interviews	
• Other formats available (telephonic administration, clinician read-administered PRO at site, bed side format, smartphone, website platform, etc.)	
• Availability of modules specific for different age ranges or different reporters (patient, proxies, observers, etc.)	

7.5.2.2.1. Special Recommendations for content validity

7.5.2.2.1.1. Content validity definition for most COAs

Content validity is considered the most important measurement quality because it represents what the OMIs actually measure (i.e., concept of interest). According to Prinsen et al. (2016), this is true to the extent that if the content validity is not clear or is not confirmed in the therapeutic field, the assessment of other measurement properties is not valuable. Additionally, the FDA states that this feature should be established prior to evaluating other measurement properties (FDA, 2020a).

The COSMIN group mentions three aspects of content validity: (1) relevance regarding the construct of interest within a specific population and COU, (2) comprehensiveness (no key aspects of the construction should be missed) and 3) comprehensibility (the items should be understood by patients as intended).

Following FDA guidance, the adequacy of OMI content validity has a direct impact on the evaluation of the accuracy of a medical product's labeling claim based on that COM. This should be supported by evidence obtained from qualitative and quantitative studies and/or published literature (FDA, 2020a).

The evaluation of these features is the most challenging. At this point, we need to distinguish between PROs and PerfOs or other types of measurement. We can use the same three aspects and apply the definition to other types of OMIs, although this will not match completely. The next section discusses the special case of cognitive measurement and its content validity.

Specifically, for PROs, the COSMIN group developed a methodology to systematically evaluate the content validity, based on the quality and results of the PROM development process. They provide a system of criteria and tools that can be easily applied when selecting and comparing OMIs (Terwee et al., 2018). The COSMIN approach is very good and is primarily designed for PROs. This means that many aspects simply do not apply to other types of OMIs or need to be adjusted for clinical outcomes, such as simple scores and classifications, among others.

7.5.2.2.1.2. Content Validity for Cognitive Measures (PerfOs)

The evaluation of the content validity of cognitive measures is even more challenging, as this type of performed-based evaluation matches the criteria mentioned in the previous section, i.e., relevance and comprehensiveness. However, although patients may perfectly understand the requested activity, they will rarely be completely aware of the construct that is being measured.

Similarly, the clinical meaningfulness of cognitive measures, which relates to the clinical benefit perceived by the patient, is also challenging. Judging the clinical meaningfulness of cognitive measures and neuropsychological tests is quite challenging. Patients may be unaware of or misinterpret their deficits based on the nature of these deficits (Rudick et al., 2014). Neuropsychological tests may have psychometrical validity but are not associated with perceived patient benefit.

Below are some illustrative examples:

- In the case of MS, for the selection of the "concept of interest", the ICF model from the WHO (WHO, n.d.) was used; specifically, the "*disability on cognition*" was used as a way to define important components of the constructs to be measured. A clear definition of the concepts was needed so that the scores obtained with the instrument can be mapped back (match) to their definition in the WHO-ICF.

  From this point, researchers need to search for research works showing the relationship between cognitive measures and specific cognitive skills in daily life. In addition, for the MSOAC project, once the individual components of a disability measure were identified, they leveraged existing data sets to see how well a composite performs as a disability measure in longitudinal studies.

  In this line, a section related to MSOAC and showing how specific cognitive subtests selected in clinical trials for specific neurological disorders justify its content validity can be cited directly from the EMA website (EMA, n.d.-b; FDA, n.d.-b):

  > *"The SDMT presents a key, consisting of nine abstract symbols. Each symbol is paired with a number ranging from 1 to 9. The test consists of 120 abstract symbols presented in random order. PwMS are asked to associate the symbols with the correct corresponding number, as shown in the key. PwMS respond orally as quickly as possible, and the number of correct responses is recorded. Processing speed is a basic, elemental cognitive function. A systematic review of the literature revealed one cognitive measure, the SDMT, as being particularly sensitive to the slowed processing of information that is commonly seen in MS. Published evidence supports the reliability and validity of this test, its relevance to daily activities, and recently has supported a responder definition of a change in the SDMT score as approximately 4 points or 10% in magnitude."*

- Another example is the NIH-MATRICS project in schizophrenia, where researchers presented a review of the literature showing correlations between cognitive tests and functional factors such as cognition and community outcomes in schizophrenia. This information provided a rationale for psychopharmacological interventions for cognitive deficits in schizophrenia (Green, Kern, et al., 2004).

It is worth commenting that the ISPOR task force has been formed to prepare manuscripts on good practices for PerfO development, validation and implementation. Additionally, an updated FDA guidance is expected in this area.

7.5.2.2.2. Summarizing the quality of existing OMIs

The goal is to tabulate the relevant psychometric features of the instruments summarizing the evidence accumulated thus far in various validation studies.

For newly developed instruments, there will be little information published (papers, posters, etc.), so it will be easy to express in a single table. However, for well-known instruments with widespread use, a number of validation studies may be available and difficult to summarize. Once all the references available are retrieved, it is often difficult to summarize all the information to evaluate the convenience of the COS measurement.

When *several publications are available for a single COM*, summary strategies such as the one shown in Table 2 can be useful to organize all the evidence in one place with a GRADE system setting the level of evidence for each publication. Columns can be named according to the relevant psychometric features, as decided in the previous section.

This can be the case for instruments such as cognitive batteries from RBANS, MoCA, PD-CRS for Parkinson disease, scales to measure psychiatric symptoms (HAD), depression (HAMD), anxiety (GAS); disability (the Sheehan disability scale (SDS)), or quality of life (SF-36). The greater the increase in the use of the scale in clinical trials, the more validation papers are generated in different conditions and COUs.

TABLE 3 PROPOSED TEMPLATE FOR SUMMARIZING OMI PSYCHOMETRIC PROPERTIES FOR A SINGLE INSTRUMENT – COLORS AND "+" SYMBOLS ARE USED AS EXAMPLES TO ILLUSTRATE THE CODES USED.

OMI Name	Content Validity	Construct Validity	Internal Consistency	Test Retest Reliability	Inter-Rater Reliability	Convergent-Divergent Validity	Known Group Validity	Ability to Detect Change (Responsiveness)	Sensitivity to Change
Validation Study 1									
Validation Study 2		++	++	++	NA		++	++	
Validation Study 3									
Validation Study 4							*(Validated but not conclusive)*		
TOTAL AVAILABLE STUDIES	#	#	#	#	#	#	#	#	#
SUMMARY OF VALIDATION STATUS					NA				
Instrument Overall Rating & Additional Comments									

Legend for Colors and Symbols
Green Cell = Validated, Ambar Cell = Validated but not conclusive, ++=Strong evidence, +=Some evidence
Summary method from Adapted from OMERACT project (ref Omeract Handbook, 2019)

TABLE 4 PROPOSED TEMPLATE FOR SUMMARIZING SEVERAL INSTRUMENTS IDENTIFIED FOR ONE DIMENSION: COMPARISON OF PSYCHOMETRIC PROPERTIES ACROSS RELEVANT INSTRUMENTS.

Measurement Instruments	Content Validity	Construct Validity	Internal Consistency	Test Retest Reliability	Inter-Rater Reliability	Convergent-Divergent Validity	Known Group Validity	Ability to Detect Change (Responsiveness)	Sensitivity to Change Parameters	Summary Psychometric Propierties
OMI-1										
OMI-2										
OMI-3										

Another common situation is there being several validated OMIs, and the analysis has to focus on determining which OMI to select among the existing validated options. In such situations, a summary table such as Table 3 can also be useful for summarizing the existing evidence regarding the psychometric features of each candidate.

In addition to the tabulation of information corresponding to several instruments, we have the example of the ImproveLTO project, which created a single psychometric data sheet for each candidate instrument, summarizing all available information. This resource is available and free for use via the study resources website www.ImproveLTO.com (Needham DN, n.d.).

Based on objective information collected from completed studies, a final decision needs to be made. Not all existing COS projects have reached the point of providing psychometric information to support final decisions. It is desirable to ensure that the eligible OMIs are at least sufficiently validated to accomplish their purpose.

This piece of information is especially complex to evaluate for panel participants not familiar with psychometric information. However, efforts should be made to illustrate the validation features (weaknesses and strengths) of each instrument. This will ensure adequate decision-making based on objective information rather than subjective information. Not all past COS projects have managed to complete the psychometric evaluation step or include only usability scores provided by clinicians to decide on the best instrument. The panel members can gain a fair understanding of the measurement properties by including an expert on the topic in the panel. This professional will be able to summarize the validation status to the other members of the panel before making a final decision regarding the COA instrument. There are some resources describing the most relevant statistical terms for measurement properties (Allied Health Professions (AHP) Outcome & Measures UK Working Group, n.d.); specifically, the WHO issued a "*Checklist for allied health professionals*" that describes the key questions to ask when selecting outcome measures.

Following the FDA statement (*see* FDA, 2020a), the sponsor task a risk when proceeding with a COA in pivotal trials without evaluating its measurement properties (i.e., at least content validity, reliability, constructs validity and ability to detect meaningful change should be evaluated). Typically, in early clinical trials, a number of COAs may be piloted for exploratory purposes.

7.5.3. Summary of the Coverage of OMIs Regarding the Measurement of Relevant Dimensions: Gap analysis

The goal is to map the dimensions that are covered/uncovered by the existing and selected instruments.

Once instruments have been selected for the measurement of the COS, there remains the important activity of mapping all OMIs and their dimensions, allowing the observation of any overlap or an otherwise existing gap of measurement among the relevant dimensions included in the COS. Table 5 is a template proposed to cover and document this activity.

It is recommended that a single OMI be selected for each outcome in a COS whenever possible, as it will allow us to increase the comparability of clinical trials (Prinsen et al., 2016). However, this is not always possible since multidimensional instruments are often used in neurology and psychiatry.

In certain areas of neuroscience, there are a high number of evaluations per patient, including evaluations addressed to caregivers or relatives (ObsROs) or even proxys (ProxyROs). For that reason, it is important to consider the risk of an excessive overlap between OMIs that may impact patient burden, i.e., the total time of evaluation at each study visit should be considered.

A given patient may be asked about a single domain several times, i.e., when using different OMIs that contain items measuring similar domains, which can lead the patient to lose confidence with the research staff.

Additionally, in cognitive evaluation (PerfO), the patient may be requested to perform a similar subtest in multiple test batteries, i.e., patients may be requested to learn different lists of words for memory tests in the same evaluation session, promoting proactive interference between memory lists.

For this activity, it is advisable that a summary of OMIs and dimensions be produced to identify potential overlaps and gaps in the battery of evaluations.

TABLE 5 TEMPLATE TO SUMMARIZE THE DIMENSION COVERAGE OF OMI FINALIST INSTRUMENTS

Outcome Measurement Instruments (OMI)	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Summary of OMI Dimensions coverage
OMI-1						
OMI-2						
OMI-3						
Dimension Covered Vs Non-Covered						

7.5.4. Alternative Methods for Endpoint Measurement

7.5.4.1. Digital Health Monitoring Technology

The dramatic growth of digital technology use in our daily lives, such as the widespread use of smartphones with embedded sensors, has created opportunities to develop novel approaches and tools for clinical outcome assessment (Coran et al., 2019; Inan et al., 2020; Perry et al., 2018). Regulators have signaled their receptivity to new digital health technology tools (DHTTs) to help understand patients' functioning and how it is affected by various diseases and treatments (see Digital Health Center Excellence (FDA, 2021b)). This section describes the range of potential DHHTs for use as clinical trial end points and the emerging regulatory landscape.

DHTTs can help capture patients' experiences during the course of their daily lives in natural settings. These tools include "wearables" and mobile sensor-based recordings of movement and physiological factors (e.g., wrist-band actigraphy recordings of activity patterns), stationary sensors (e.g., home-based motion sensors to detect gait patterns or falls), and even ingestible/implantable sensors. Data can be captured either passively, such as via continuous automatic recordings that do not require patient responses, or actively, such as via intermittently cued patient-reported experiences (e.g., mood ratings) or performance measures (e.g., cognitive tasks) throughout the course of a day.

A key potential benefit of DHHTs is their unique capacity to capture rich, ecologically valid data about how patients feel and function in real-world settings. They may also have benefits in terms of operational efficiency and minimizing barriers to clinical trial participation. For example,

DHHTs can be administered and accessed remotely, thereby reducing site (e.g., decreased study visits), patient (e.g., decreased travel costs), and on-site data monitoring costs. In addition, they may help recruit and retain geographically diverse and difficult-to-reach participants.

Regarding regulatory considerations, the emerging FDA evidentiary standards for DHTT-based clinical outcome assessments are broadly similar to those of conventional COAs (see Dashiell-Aje B, Kovacs S and Sacks L, Regulatory Perspective: Digital Health Technology Tools Use in Clinical Investigations to Evaluate Clinical Benefit in Patients) (Dashiell-Aje B et al., 2019).

DHHT endpoints must be well defined and focus on a concept of interest that is clinically meaningful for the proposed study population, ideally incorporating input from key stakeholders throughout the measure development process. Endpoints must also demonstrate acceptable psychometrics properties for clinical trial endpoints, including adequate reliability, validity, and ability to detect change.

In Europe, there are also useful resources issued by EMA for DHTT to better understand the legal environment (*see* Questions and answers: Qualification of digital technology-based methodologies to support approval of medicinal products, EMA, 2020) (EMA, n.d.-c).

Key considerations: While DHTTs show strong potential to serve as informative new COAs, they raise some unique practical and scientific considerations. Regarding instrumentation, it is important to select an instrument that is reliably calibrated, acceptably validated, and practical (e.g., has sufficient battery life) for measuring the intended concept of interest. Many DHTTs generate a very large volume of data, which raises practical data transmission, storage, and processing considerations. Furthermore, complex multivariate statistical algorithms are often used to derive DHTT-based endpoints; it is important to define, a priori, how data will be aggregated to generate a clinically meaningful endpoint. It is also important to consider potentially relevant characteristics of the population of interest, such as technical aptitude or cognitive/physical limitations, which may impact the usability of certain DHTTs. Finally, approaches to ensure data authenticity, integrity, and confidentiality require careful consideration.

Minimal activities: The development and validation of DHTTs is a rapidly emerging area. Experts have begun to develop recommendations and resources to increase the quality and efficiency of DHTT development. For example, the Clinical Trials Transformation Initiative (CTTI) (CTTI, n.d.-a)has issued best practices that sponsors can use to develop DHTTs for FDA submission and inspection. The FDA recommends that sponsors engage with regulators early in the DHTT development process to develop these novel COAs.

7.5.4.2. Goal Attainment Scaling (GAS) Methodology

For disorders with high clinical heterogeneity among patients, traditional methods of endpoint measurement usually face significant limitations. For this specific therapeutic context, other methods are proposed to avoid failure due to methodological issues. This section describes the basic features of the goal attainment scaling (GAS) method, which has undergone substantial development for use in clinical trials but has not been fully validated in all contexts of use.

GAS is a methodology based on individual patient progress in relation to a treatment, similar to how it occurs in clinical practice (Kings College London. The North West London Hospitals NHS Trust, 2009; Kiresuk & Sherman, 1968). This means that in the process of evaluating the effects of the intervention, an individualized outcome objective is created for each patient on the basis of patient-specific goals (Urach et al., 2019).

Minimal activities: The method is based on setting goals for each participant. The goal definition comes from a structured interview between the patient and clinician in a way that will allow valid comparisons of treatment progress across patients and treatment modalities. The selected goals need to allow for the evaluation of "change" in a specific patient. Patients or caregivers typically provide a statement of intention rather than a goal. Thus, a critical task during the baseline interview is to *shape* intention statements into measurable goals and help participants *choose* among multiple potential aims and goals, selecting those that meet the standards for research conduct. The process of setting goals in consensus with the goal attainment scaling assessor ensures that *measurable*, *equidistant* anchors are established across all levels of performance, with *equal* levels of difficulty for patients to ensure comparability (Opler M et al., 2018).

Each goal is classically defined by 5 levels of attainment, i.e., five anchors from + 2 (much better than expected) to − 2 (much worse than expected), which define the GAS attainment levels as suggested by the original authors (Kiresuk & Sherman, 1968). Regarding scaling, it is critical to assign equal steps or spaces between scale anchor points. During a clinical trial, the scale is then used at designated time intervals to assess the degree of progress toward goal attainment, and following the trial intervention, independent assessment is conducted to evaluate the level of attainment for each goal.

Additionally, the goals can be weighted based on an optional approach using the relative importance to the subject. For this process, patients choose the weights of the goals to differentiate between goals of different relevance (Urach et al., 2019).

Key considerations: It is important that the goals chosen will potentially be affected by the intervention; otherwise, the method can result in a substantial loss of power. Additionally, if weights are applied, they must be correlated with treatment effects.

The method by which data are analyzed is also relevant, and while the outcome measure is unique to each subject, standardized scoring needs to be applied to allow for statistical analysis, requiring a specific statistical approach developed for some therapeutic fields by expert groups (Urach et al., 2019).

This method has been proposed to cover the measurement gaps in clinical trials for heterogeneous groups of patients, rare diseases with heterogeneous and small patient samples (for instance, mitochondrial DNA and Duchenne muscular dystrophy)(Gaasterland et al., 2019), and chronic diseases with geriatric or rehabilitation trials. However, it has not been completely validated for all disease areas. For instance, it has been mentioned elsewhere that this method is potentially less useful for acute, episodic or unpredictable diseases (Urach et al., 2019). Care must be taken when considering whether to apply goal attainment scaling for a given disease or treatment.

Some of the challenges and *risks* associated with using this method are the lack of experience of use in the targeted therapeutic area and the training and time required for goal setting. Before using this method, it is advisable to create a comprehensive manual with the goal of guiding users and bringing consistency to the goal setting process. Furthermore, its validation will follow many of the same steps as a standard COA instrument (*see* Table 1).

- ⓘ An example of validating the number of goals to define in controlled Alzheimer's disease clinical trials can be found at (McGarrigle & Rockwood, 2020).

- ⓘ A second example of GAS goal validation was given in the field of mood disorders by Opler M., et al. 2018 based on data collected from a real open label trial to analyze its validation properties. In that case, the GAS methodology was applied as the primary endpoint in a phase 4 open label study for the treatment of major depression (NCT02972632, n.d.). The primary endpoint was set as the percentage of participants with an achieved GAS score of ≥50 at a specific time frame (equivalent to an individual unadjusted score of "+ 1"). In that study, we can also find how two types of goals were set: one type was defined naturalistically at the subject level, and another type was selected using a domain-defined approach.

In addition to these examples, a full manual on how to use this methodology for depressive disorders and related conditions is available from (Opler M et al., 2018) and may be used as an example of following the GAS methodology in other contexts of use.

In early drug development, the GAS method may also serve as a useful tool to guide the identification and selection of concepts of interest. In addition, this approach will also allow incorporation of "the voice of the patient" (e.g., ensure patient-centric drug development).

7.5.5. Additional information from *post hoc* analysis or RWE databases

The goal is to plan an analysis of existing data sets to fill the existing gaps related to the psychometric features of the instruments or dimensions.

Once a COS and the corresponding OMIs have been established, there are some resources that can be used to further explore additional properties of the instruments, such as gender differences or therapeutic response, to optimize the measurement of the outcome, such as *patient registries* and completed *trial data sets* (Arnerić et al., 2018). This has been particularly successful in the MS field through the C-PATH MSOAC project.

The specific types of databases are as follows:

- Datasets from existing large consortia on a specific disorder (searchable database from EU-funded projects, such as CORDIS from the European Commission (EU European Commission, n.d.)).
- Data from single completed clinical trials with specific features (blinded or unblinded).
- Pulled data from different clinical trial data sets, either with the same or different study treatments (considering the risk of mixing patient samples, even from the same disorder).

Exploring the efficiency of COAs in completed RCTs

The case of existing completed *trial data sets* provides a valuable opportunity to explore the efficiency of a COA before making the final decision or once decided to be sure about its performance in the context of RCTs.

Resources of this type are becoming more available to researchers from various initiatives. The Yale Open Data Access (YODA) initiative is worth mentioning as an illustrative example through which data holders can share their clinical research data responsively and researchers can request access to clinical trial data from multiple data sets in a number of indications (YODA, n.d.).

Several factors can be crosschecked with real RCT data. In addition to psychometric properties specific to the context, other features, such as exploring factors influencing *placebo arm* response to treatment, are useful for better determining the risk of having a high placebo response making it difficult to obtain a positive result.

The efficiency shown by the instrument in real trials confirms its properties in the corresponding COU, as well as its sensitivity to change and MCID properties in the context of the study population and study design. All this information can inform sample size calculations for new study protocols.

From post hoc analysis of completed clinical trials or RWE databases, the information that can be obtained can be summarized as follows:

- Confirm the psychometric properties of the OMIs selected for the COS.
- Confirm the factors influencing the placebo arm response
- Responsiveness, sensitivity to change and MCID
- Divergent and convergent validity with other measurements used in the same trial

ⓘ As an example, confirmation of convergent validity in schizophrenia between the CGI-SCH and PANSS was conducted in a sample of inpatients and outpatients from existing data sets. The concordance shown between COAs included the divergent pattern regarding depression measured by HDRS. The results supported the use of the multidimensional CGI-SCH specific for schizophrenia instead of a generic and unidimensional CGI in future clinical trials (Haro JM & Domenech C, 2019).

Data from completed trials also allow for quality data analysis as follows:

- Identify monitoring data checks to confirm the quality of the measurements during ongoing clinical trials or for auditing purposes
- Confirm factors affecting the quality of the data collected

ⓘ Examples of consistency checks for data monitoring can be have been given by ISCTM working groups for MADRS (Rabinowitz et al., 2019) and PANSS (Rabinowitz et al., 2017), with a post hoc analysis of the inconsistencies found, and applying flags to ratings may improve the reliability of ratings and validity of trials.

ⓘ Other works have explored the question of whether the data quality of COA measurements may have affected the results, usually in negative trials. In that case, sensitivity analyses at the patient and site level of poor performing sites (flagged) can provide additional information for future use of the same COA tools. The troubling question is whether the failure of studies may have been due to or confounded by a substantial amount of low-quality assessments rather than the lack of a pharmacological effect of the study drug (Umbricht et al., 2020).

One of the risks of using existing data sets is limited external validity, which is usually limited to the study population included in the set. For that reason, to support phase III designs based on the results of previous phase II trials, the risks of bias toward the selected sample must be considered. A comprehensive discussion section can be advisable to mitigate any risk when other researchers use the data obtained regarding COS *post hoc* results.

The extent of this approach depends mainly on the type of data available for appropriate hypothesis-driven analysis.

The *risk* associated with this analysis is undermining the external validity of the results. Post hoc analyses are usually limited to the study features from which the data are obtained.

Exploring the study design improvements

We can find several examples of this type of *post hoc* analysis of data sets, which are usually hypothesis driven, to answer methodological questions.

- ⓘ As an example, in the area of treating negative symptoms of schizophrenia, a recent publication analyzed data from completed clinical trials to answer some questions to improve the development of new trials. *(1) How can placebo effects be minimized? (2) Should global measures of negative symptoms be included? (3) Should a new drug targeting negative symptoms be tested in a monotherapy design or in an add-on design? (4) Can new information from negative symptom trials inform the selection of COAs for future trials?* (Marder et al., 2020).

- ⓘ An example of this type of reuse of data to learn from experience is seen in the work of Umbricht et al., 2020, which assesses the impact of erratic ratings on drug-placebo response separation using data from multiple completed clinical trials with the same compound. From this analysis, the need to flag erratic ratings was shown as a need for enhanced quality control in clinical trials.

Exploring the effect of creating shorter versions of full instruments

Shortening the length of gold standard COAs is also an option to optimize legacy instruments to reduce the burden of evaluations in clinical trials when a number of outcomes need to be measured.

ⓘ Younis et al., 2020 aimed to improve the efficiency of clinical trials for schizophrenia interventions; the authors analyzed existing data sets from 32 placebo-controlled RCTs corresponding to 8 atypical antipsychotics. They analyzed the potential to streamline the design of schizophrenia trials by shortening the gold standard PANSS to the 19 items most informative as endpoints and the duration of the randomization period.

Exploring the effect of study procedures to improve study efficiency

During the planning and execution phase of clinical trials, decisions made can also have an effect on the final study result. For instance, the availability of raters at study sites can be seen as a limiting factor, as raters might change for a given patient for several operational reasons. Interrater reliability can be one of the psychometric features to consider when selecting instruments, and the effect of changing raters during study visits is something that can be explored outside the context of a clinical trial by using existing data sets.

The impact of the effect of rater change across patient study visits on the measurement of outcomes has also been analyzed in psychiatry.

ⓘ The effect of rater change on the PANSS scores of within-subject variability has been explored to improve the efficiency of the COAs of composite score, dimension score and item level of PANSS, for example (Crittenden-Ward K et al., 2020; Kott A et al., 2020).

7.6. Step 6 - Final generic recommendations of OMIs for COS measurement in the COU

The goal is to draw conclusions on the recommended COS and suggested OMIs for use in clinical trials for specific drug development as primary and secondary outcomes.

Once the basic domains are identified, i.e., specifying the minimal COS for the condition, the next step is to clarify the criteria that the corresponding OMIs should have to be appropriate for measuring the specific outcomes. Then, the final work consists of picking between 1 and 3 exemplar OMIs that are either validated or not but that require further validation steps.

For this recommendation, we need to distinguish two situations:

- When working in a *new full COS project*, it is advisable to avoid selecting specific tests for a given condition, as setting OMIs as gold standards might stifle innovation. This is because pharmaceutical companies are reluctant to use newer, possibly better measures. The

recommendation of 1 to 3 OMIs can be a good strategy to leave the final decision to future users of the COS.

- When working in a *specific clinical trial*, the selection of the OMI for the primary endpoint should ideally be based on completely validated instruments. This means that for instruments that are not validated, i.e., FFP, a research program addressed to reach the highest validation status possible can be designed.

It is important to keep in mind that in a clinical trial, the main goal is to validate the new compound in a given indication in a specific COU. For this goal, the ideal scenario is to use a validated instrument; otherwise, in the case of negative results, we cannot conclude whether the negative result is due to the intervention or to the measurement instrument. We should keep in mind a meaningful statement by Dough Altman (1995): "Absence of evidence is not evidence of absence" (Altman & Bland, 1995).

Regardless of the specific utility, the comparability of findings across trials should be maximized to facilitate meta-analysis and comparative and effectiveness research.

Additional factors to consider for a final OMI decision

- ***Deciding the timing and frequency of measurements:*** As noted in Section 7.1, setting *what* and *how* is not always enough. Therefore, it is important to consider the variation regarding the use of measurement instruments when reviewing the literature for clinical trials (Lange et al., 2021). For a new COS in a new indication, to allow meta-analysis or qualitative comparison of evidence from multiple trials, it is not enough to set the core outcome domains and measurement instruments for meaningful evidence synthesis. The COS should go further and at least standardize the timing of measurements, the specific analysis metric and the method of aggregation.

- ***Confirm the best practices with a specific instrument***. A clear example of apprising instruments in use is provided by the recently published review work of Rabinowitz et al., 2021 in the harmonization of the use of the Personal and Social Performance Scale (PSP), a well-known instrument used in schizophrenia clinical trials. The authors propose methods to run consistency checks during ongoing studies (known as "central data surveillance" or "central review of ratings"). Additionally, for another well-known instrument, the PANSS, consistency checks were developed to improve the reliability and validity of clinical trials in schizophrenia (Rabinowitz et al., 2017).

- ***Confirm the efficiency of the measurement method using recent trials and learn from recent applications:*** Another example of analyzing data sets to increase the efficiency of clinical trials, in this case for schizophrenia interventions, is given in (Younis et al., 2020), which includes data from 32 placebo-controlled RCTs of 8 atypical antipsychotics. The

authors analyzed the potential to streamline the design of schizophrenia trials by shortening the gold standard PANSS to the 19 items most informative as endpoints and the duration of the randomization period. Other attempts to shorten the PANSS to 6 items were obtained from existing databases, such as the CATIE cohort (Stroup et al., 2003).

- ***Seek standardization efforts for data management with the selected measures:*** For further standardization efforts, the *CDISC Initiative* (*Clinical Data Interchange Standards Consortium)*(CDISC, n.d.) is making efforts to minimize this situation, but it is just a first step. They rely on only how the data collected should be structured but do not specify what data should be collected or how to conduct clinical trials, assessments or endpoints. All efforts made to prevent variation in measurement in advance will have a clear impact on how the field will progress in the future.

- ***Ensure the standardization of the administration procedures for the selected measure***: Usually, under administration procedures, directions are included not only for administration but also for recording the measurements and scoring the instrument. One of the most common issues with OMIs is the lack of administration instructions, resulting in low reliability when administered by different raters in multicentric trials. This type of complementary OMI material is relevant to ClinROs and PerfOs and to PROs to some extent and is more complex than a list of questions and answers. During recent decades, a number of OMIs used for clinical practice have shown reliability issues when applied to clinical trials. For that reason, administration tools have been developed for traditional instruments with the aim of improving the reliability and consistency of ratings. Clear administration instructions and occasional improvement of anchor score descriptions help ensure reliability across clinicians applying OMIs and across assessments. Examples of such complementary tools developed for legacy instruments are, for instance, the Semistructured Interview Guide for the Hamilton Depression Rating Scale (SIGH-D) for HAM-D (J. B. Williams, 1988; J. B. W. Williams et al., 2008), Interview Guide for the Hamilton Anxiety Rating Scale SGH-A for HAM-A (Shear et al., 2001) and the structured interview guide for the Montgomery Depression Rating Scale (SIGMA) for MADRS (J. B. W. Williams & Kobak, 2008). In the field of neurology, it is worth mentioning the effort of the Movements Disorder Society to standardize the administration and scoring of the UPDRS into the MDS-UPDRS (Goetz et al., 2008) revised version of this legacy instrument. Upon the selection of an OMI, efforts have to be made to find instructions for administration, recording and scoring. When nothing exists, this type of materials can be generated by clinicians with the expertise in the use of the instruments.

7.6.1. Considerations regarding drug labeling

The goal is to ascertain whether the final COS will be accepted by regulators to support labeling claims and identify the risks and next steps.

COAs are often endpoints in clinical trials that are used to support drug approval and labeling claims or other communications regarding clinical benefits. FDA primarily uses COAs to determine whether a drug has been shown to provide clinical benefits to patients. The severity of side effects or treatment burden can also be measured by COAs or by other objective endpoints (lab, imaging, etc.). In the process of obtaining a label claim, regulators will review the totality of data and conduct a risk-benefit analysis.

The evaluation of risks can create a road map for a COS development plan (s*ee Section 5*).

Future steps for proceeding to validate COS for drug labeling purposes

Any labeling claim requires the collection of measures with validated instruments regarding outcomes and its context of use *(see Section 2)*. There are some guidance documents from regulatory agencies to take into account; however, available documents do not encompass all COA types. As an example, existing PRO guidance documents for industry (as elaborated by FDA) are from the 2009 "Patient-Reported Outcome Measures: Use on Medical Product Development to Support Labelling Claims" report (FDA, 2009).

Another source of guidance includes the consensus documents issued by Task Force groups at the Existing Scientific Societies, which can be used in support of the COA strategy for the early stages of drug development. This might be useful for building the background of the COU, domains to measure, target population or subpopulations to support the labeling claim.

For new instruments, consider qualification steps in the context of the FDA, EMA, etc.

The US FDA has a specific office focused on supporting the qualification of new tools for use in drug development research, named the CDER DDT COA Qualification Program (FDA, 2020b) (Drug Development Tools). However, in the past, few researchers or sponsors followed the path, in part due to the uncertainty of outcome and the anticipated lengthy research process required to establish evidence discouraging initially interested researchers. However, applications that are public at certain stages can guide the path for other researchers and drug developers (FDA, n.d.-a). "Qualification" represents the determination that the drug development tool (DDT), within a specific context of use, can be relied upon to have a specific interpretation and application in drug development and regulatory review. Once qualified, the DDT that has successfully completed the process can generally be included in IND, NDA, or BLA submission without the need for the FDA to reconsider and reconfirm its suitability.

Existing public databases of DDT-FDA applications/approvals show that this specific path is used by academic research groups at hospitals, clinical research organizations and consortia/working groups at Critical Path Institute (C-Path), as the sponsors that have thus far submitted

instruments or innovative endpoints via this strategy of qualification process. Overall, the submitted documents, which are public thorough the FDA website, can be considered a reference and inspire new researchers undertaking new COA projects.

COA developers might have an interest in submitting their COA development plans (see Section 5) to gain formal recognition of the validation activity and to have a review from technical experts on the validation plans. All this work can be performed in the preparation to offer codevelopment with industries interested in a specific COA in a specific COU for a specific disorder.

However, pros and cons must be considered to select this strategy based on the time and available resources required for drug development, which is the main goal for drug makers. A quicker and more feasible path is the series of meetings with FDA CDER teams, which may or may not include COA teams, but eventually, all COA plans will be evaluated by the same team of experts. A meeting with regulators is the most common strategy used by the pharma industry (*Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products Guidance for Industry*, 2017).

In Europe, the EMA publishes opinions on the qualification of innovative development methods and letters of support for novel methodologies that have been shown to be promising in the context of research into and development of pharmaceuticals. Opinions were given by the EMA's CHMP on the basis of recommendations by the Scientific Advice Working Party (EMA EWP, n.d.). Examples of such qualifications in the neurosciences are MOASC in MS and Alzheimer's disease and novel data-driven methods for monitoring disease progression.

7.7. Step 7- Conclusions of Gap Analysis and Future Steps

The goal is to specify the next steps based on the analysis conducted at the end of COS selection.

The application of Tables 4 and 5 will produce a list of potential overlaps and gaps in measurement that will need further work.

Overlapping dimensions/items should be identified, and it is recommended that such overlap be avoided to prevent asking the patient about the same issue in multiple evaluations. IRT methods can help minimize redundancy across instruments in the evaluations (*see Section 11.1* for details about the method).

Regarding dimension measurements, the selected instruments do not cover the dimensions or concepts of interest, therefore requiring a special plan to develop new instruments (*see* Creation of a COA development plan on Section 5) or to add sections to existing instruments, among the options.

In the field of neuroscience, there are several examples of instruments repositioned to measure similar dimensions in conditions other than that initially targeted.

- ⓘ An example is the SAPS-PD, the adaptation of the legacy instrument for Psychosis SAPS (Andreasen NC, 1984), which is used to measure the most common and relevant features of psychosis in Parkinson's disease (T. S. Voss et al., 2010), and a short version of the adapted instrument has been validated (Voss et al., 2013).

- ⓘ Another example is the CDR® Plus NACC FTLD, adapted from the CDR legacy instrument as an outcome measure for clinical trials in mild-symptomatic frontotemporal lobar degeneration (FTLD) or frontotemporal dementia (FD) (Miyagawa et al., 2020).

7.7.1. Selecting the most suitable format for existing OMIs

Special mention should be given to the formats of the instrument available for use in clinical trials. The format of the instrument that will be used in a clinical trial also needs to be adapted to the context of the evaluation, e.g., whether it will be at a site of the clinical study visit or patient follow-up as a remote evaluation.

Gaps might also result from not only the dimension coverage of the instruments but also the availability of formats of the instrument adapted to the study design. Adaptation of existing OMIs to specific study requirements will also require specific validation and testing for usability.

7.7.1.1. Availability of electronic formats of the OMIs (eCOAs)

Depending on the type of the COM selected, a key consideration for clinical trial implementation, particularly for large-scale trials, is whether the measure is amenable to administration in an electronic format. Many traditional PROs, ClinROs, ObsROs, and even PerfOs that involve pen-and-paper or specialized equipment (e.g., cognitive tasks that involve object manipulation) can potentially be adapted to electronic COAs (eCOAs). Tablet-based eCOAs are increasingly common, although electronic assessments can be presented in a variety of formats, including a web-linked computer, hand-held formats, digital pens, and interactive voice response systems. Although there are many potential benefits to using eCOAs over conventional approaches, several factors require careful consideration during the adaptation process.

There are many scientific, practical, and regulatory advantages of eCOAs over conventional methods. The benefits have been most clearly described for PROs (Coons et al., 2009, 2015). As has been reviewed by Coons and colleagues, ePROs can produce more accurate and complete data, provide greater protocol compliance, and result in reduced data entry and scoring errors, lower participant and administrative burden, reduced sample size requirements, and cost savings. The growing evidence of increased accuracy and integrity of ePRO data collection in clinical trials has been accompanied by regulatory acceptance, including FDA guidance documents (FDA, 2021b). Electronic formats offer similar benefits for other types of COAs. For example, electronic ClinROs can enhance standardization of administration by systematically guiding raters through structured interviews and by rapidly highlighting atypical ratings (e.g., between related items) to raters for quality control. Similarly, electronic cognitive PerfOs can standardize administration of instructions and time limits and automate scoring (e.g., Atkins et al., 2017).

The potential benefits of eCOAs will not be realized if a conventional measure is not properly adapted to this format. Investigators need to ensure that the adapted measure validly assesses the relevant items/constructs in the same way as the original modality. Adaptation to an electronic format has the potential to introduce various response biases (e.g., impacting content, meaning, or interpretation) and negatively affect psychometric properties. The extent of modification required can vary substantially across different COAs. For example, some types of PROs require little modification to migrate to electronic format, whereas some PerfO measures that involve specialized materials (e.g., block design type tasks) would require substantial modification or may not be suitable for adaptation.

Typically, the more modification is required during the migration process, the higher the level of evidence will be required to demonstrate that the change did not introduce response bias, did not negatively affect psychometric properties, or did not alter construct validity. In the area of ePROs, consensus-based guidelines and evidentiary standards have been developed for demonstrating equivalence across modalities that involve different levels of modification (Coons et al., 2009). For

example, a minor modification to a PRO, such as a nonsubstantive change from circling responses on paper to touching a response on a screen, would require usability testing (whether respondents from a given population are able to use the software and device properly) plus cognitive debriefing (exploring how members of the target population understand, mentally process, and respond to items). A moderate modification, such as changing from visual item presentation on paper to aural electronic presentation, would require usability testing plus equivalence testing (e.g., comparing the modalities with randomized parallel groups or a crossover design). Finally, a substantial modification, such as changes in item response options or wording, would require usability testing plus full psychometric testing (evaluation as if it were a new measure, including content validity, internal consistency, test-retest reliability, and construct validity).

It can never be simply assumed that administering a COA as an electronic modality is equivalent to the original measure. While clear guidance has been developed for PROs, adaptation guidelines are less clearly established for other types of measures, particularly PerfO measures. Although the guiding principles developed for PROs are generally applicable to other types of COAs, efforts to systematically adapt and validate conventional measures are relatively rare (e.g., Atkins et al., 2017).

7.7.1.2. Availability of remote administration formats of COA (rCOA/reCOA)

Another key consideration for COAs is whether they are suitable for remote administration. The past decade has witnessed growing interest in decentralized or "site-less" clinical trials (Apostolaros et al., 2020; CTTI, n.d.-b). This patient-centric approach aims to shift clinical trial assessments from in-person visits to partly, or in some cases entirely, remotely conducted assessments with participants in their own natural environments. Enabled by advances in digital technology, remote assessments offer many potential benefits in terms of clinical trial access and efficiency. However, incorporation of this approach into clinical trials has been slow owing to factors such as immature digital infrastructure, limited experience with this approach, and the perception of regulatory barriers. Notably, in the very recent context of the COVID-19 global pandemic, utilization of remote assessment approaches dramatically increased in clinical trials (Xue et al., 2020), as investigators sought to minimize data loss associated with lockdowns and social distancing requirements, which greatly complicated in-person assessments. Adaptation during this crisis period was facilitated by issuance of regulatory guidelines that signaled openness to this approach under certain circumstances (EMA, 2021; FDA, 2021c). Indeed, COVID-19 has, in many ways, catalyzed the adoption of "virtual" clinical trial technologies.

In the context of a decentralized trial approach, remote COAs capitalize on data collection technology to virtually administer assessments to participants in their home and/or daily life environments. Relevant technologies include telehealth and video conferencing interfaces (e.g., administering a clinical interview via Zoom) or the use of devices to administer supervised or unsupervised eCOAs (e.g., electronic questionnaires or performance measures administered via smartphone). As such, participants can be enrolled almost anywhere, which has numerous

potential benefits. These include decreased burden for patients through lower travel costs and time loss, as well as the capacities to recruit faster, engage more diverse and generalizable samples (e.g., participants residing far from a medical center or unable to travel), and facilitate retention (which can reduce sample size requirements).

As with all new technological advances, remote assessments involve a variety of challenges and risks. First, the issue of demonstrating equivalence between in-person assessments conducted in a controlled research setting versus remote assessments conducted "in the wild" is paramount. Similar to the considerations for demonstrating equivalence between conventional and electronic COAs (*see* Section 7.7.1.1), the more modification is required for adaptation to a remote administration format, the higher the evidentiary standards will be to provide compelling demonstration of equivalence with in-person administration. Notably, there is evidence that even relatively complex neuropsychological tasks can be validly administered in a remote context (Bilderbeck et al., 2020). Second, patients vary in their access to, and comfort with, technology. Trial devices can be provided to participants or a "bring your own device" (BYOD) approach can be used. While provisioned devices have the benefit of ensuring access to standardized hardware, participants may be less familiar with these devices. On the other hand, while a BYOD approach has the benefits of maximizing participant familiarity with their own devices and decreasing the study costs, there is less hardware standardization across participants (e.g., the screen size or operating system), which creates scientific and practical challenges. Third, when using unsupervised assessments, the quality of data collection is unknown. For example, household members may help or substitute for the patient, or the participant may be multitasking and suboptimally attentive while completing assessments. Fourth, key issues related to participant privacy and data security/storage require careful consideration.

The development of COAs is growing rapidly, and rCOAs will likely play an increasingly larger role in the post-COVID era. While there are key challenges and risks to consider, guidance to mitigate risks is emerging. For example, the Clinical Trials Transformation Initiative (Apostolaros et al., 2020; CTTI, 2015, 2018) assembled data from diverse stakeholders to develop recommendations to address the most prevalent legal, regulatory, and practical issues. While regulators have increasingly signaled openness to rCOAs, the regulatory landscape is emerging in this area. It is critical to seek regulatory guidance early in the process of developing remote assessment measures.

8. Consensus on measurement, appraisal, prescription and discussion

The goal is to compile all information collected during the analysis and make it ready for a broad dissemination in a larger audience, allowing us to accelerate future research in the same field.

A list of topics that can be included in this activity are as follows:

- Start a discussion section on the recommended COS and OMIs suggested as results of the project.
- Main prescriptions for OMI selection in the close context of use for various intervention options.
- Identify all lessons learned for future COS users and COS developers.

9. Cultural considerations in the translation of OMIs within a specific condition

For multicenter international research, translation activities are an important part of the starting activities. Local versions of instruments are known as linguistic validated versions of the source COA, usually in English from the US, since psychometric validation is not required. A number of guidance documents exist for producing local versions ensuring concept equivalence to the source instrument (Wild et al., 2005).

An initial translatability analysis of the OMI is always recommendable to identify sections of the instruments that may require different adaptation processes and sections or items that could require deeper work and discussions during the cultural adaptation process.

Preparatory work includes the translatability analysis, followed by the confirmation of constructs and item conceptual content, which quite often is included within the instrument materials, i.e., test manual or published validation papers.

The linguistic validation process required for international drug development follows quite a straightforward sequential path starting with forward translations (usually twice), reconciliation, and back translation (usually once), followed by another review processes and proof-reading steps.

The process can vary depending on the type of instrument and format. Description of this process with examples can be found at (Hall et al., 2018) and a summary of experiences in CNS clinical trials at Krishna et al (in press).

10. Licensing and relevant operational considerations

The goal is to use the selected OMIs appropriately with respect to the source documents, available translations, and existing related information and eventually perform due diligence to obtain permissions from copyright owners for use of their materials in clinical trials for a specific number of participants and uses.

The key items to cover when considering a specific existing instrument are

1. Confirmation of source version and original language
2. Identify the copyright owner and reference statements to include in the materials (copyright notice)
3. Understanding conditions of use or information about the entity in charge of licensing the use of the OMI (if different from the copyright owner)
4. Estimation of the cost of the use of the instrument and terms for the cost estimation (number of administrations, , number of languages required, etc.)
5. Information about the author review and approval process and associated timelines
6. Samples of all necessary materials related to the instrument administration, recording and scoring including but not limited to the following: administration instructions, record forms (rater), response forms (subject), scoring materials (templates), scoring instructions and software, stimulus book (e.g., pictures, words), stimulus kit materials (toys, goods, cubes, etc.), manual (including instructions), or stopwatch, etc,.
7. Role of the licensor when changes are introduced in terms of review or merely providing information.
8. Application of appropriate guidelines when electronic adaptation is needed (eCOA for tablet, web-based, App)
9. Existing validation information related to remote administration of the instrument when needed.
10. Utilization of existing translations and understanding the Linguistic validation process (usually described within the corresponding Certification of Translations, COT) and any specific request around potential development of translations.

Related to this section is also the need to define the level of education or certification required to administer and rate the instruments to be used as primary endpoints and to ensure the proper use of the secondary and exploratory endpoints.

For the certification and training of raters, although there are no standards defined other than in existing guidelines (Wise-Rankovi A et al., 2014), there are common practices that the majority of industry follows.

11. Roadmap for innovation on OMIs in the area

Following the decision on the COS, there may be no instruments developed to measure the dimension(s) of interest. The option is to develop new instruments either from scratch or from existing ones, completing necessary validation steps.

Typically, in early clinical trials, a number of COAs may be piloted for exploratory purposes. According to FDA guidelines, exploratory studies (early medical product development) are an opportune time to examine COA measurement properties and performance a priori when initiating confirmatory clinical trials; stand-alone noninterventional studies are another option. The goal of pilot testing COAs is to select and/or refine a COA to be carried forward into registration trials to establish product effectiveness (FDA, 2020a). The bottom-up approach involves designing PoC/Phase II trials including new instruments and use the data to demonstrate its validation properties.

However, there is a risk to proceeding in this way since the sample selected for an experimental-exploratory design might not be fully representative of the population suffering from the condition, which poses a risk when planning pivotal trials.

There are several options for building new instruments. In addition to the option to start from scratch, there are other alternatives, such as the use of *item data banks* or the reuse of dimensions included in existing validated instruments. In any case, the new instrument will require a full validation process for the new intended use.

The next section briefly describes the process of building instruments from item data banks.

11.1. Building a new COA using Item Data Banks supported by Item Response Theory (IRT)

As commented in *Section* 7.5.2.1, if nothing exists to evaluate a specific outcome/domain, a feasible approach is to build item data banks. Item banks were developed by the PROMIS initiative after exhaustive identification of items measuring the targeted outcomes in a way that ensures efficiency, flexibility and precision in the measurement of commonly studied PROs (Cella et al., 2010).

PROMIS item banks, as prime examples, are one of the most used sources for PROM outcome/domains to measure generic symptoms and functional outcomes. The COAs generated by them address a number of conditions, including pain, fatigue, emotional distress, and sleep/wake disturbances, and are related to various types of populations. In addition, item banks are enlarged for specific subdomains and levels of measurement that are needed for assessing specific patient populations and patients with rare diseases. This approach is very well aligned with the high standards of content validity required by regulatory agencies.

PROMIS represents a model to be followed by sponsors aiming to conduct COA projects based on existing materials, i.e., items/scales. As described by the authors, the main steps to following this approach start with a *systematic review* to identify existing questionnaires to collect the core item pool and then *six phases of item development*: identification of existing items; item classification and selection; item review and revision; focus group input to confirm domain coverage; cognitive interviews with individual items; and a final revision of items before exposing them to field testing.

When nothing exists or is applicable to the new COA, the *new generation of items* to existing item pool data banks is driven by qualitative research with the target patient population thorough focus groups. Qualitative methodology allows the identification of the most relevant disease domains and even subdomains, therefore increasing measurement accuracy and supporting content validity when the new instrument will be used for labeling purposes (for a systematic review of qualitative studies, refer to *Section 7.1.2*).

Once identified as candidates COA items, item style and response options also require additional work for consistency; to an extent, uniformity needs to be favored when legacy items are used. Therefore, item review and revision steps are conducted to harmonize the question style, ease the literacy requirements of respondents, and apply a consistent set of response options and time frames.

In the last phase, those *items successfully screened or generated* through the abovementioned processes and harmonized in style (question and response format) are sent for field testing and then subject to scale construction procedures based on item response theory (IRT) methods (DeWalt et al., 2007).

In this line, the testing phase consists of a quantitative evaluation and calibration of the final items by collecting patient-reported data. The testing plan is designed to reach five objectives: to obtain item calibrations for each domain, to estimate profile scores for disease(s) population(s), to create linking metrics to legacy questionnaires (construct validity), to confirm the factor structure of the domains and to conduct item and bank analyses. Additionally, in addition to these objectives, the validity of longitudinal clinical research has also been confirmed (Cella et al., 2010). Further information about quantitative analysis can be found at (Reeve et al., 2007). Importantly, IRT methods allow for accurate measurement linking across populations by providing a common COA metric.

The quantitative method of analysis allows for the development of short forms of an instrument using a specific analysis plan, by which few items (approximately 7–8) can provide information comparable to that provided by "legacy measures" containing more items, as is the case for PROMIS instruments developed for anxiety, depression and anger (Pilkonis et al., 2011).

During qualitative steps, field testing with the target population is critical to confirm the vocabulary and thinking patterns of the target population and, more importantly, to identify gaps in domain measurement (DeWalt et al., 2007). Focus groups can include medical specialists attending the targeted patient population and patients (Pilkonis et al., 2011).

Since an item pool is obtained from various sources, it is very important to carefully review the items to ensure readability and resolve any intellectual property issues for all items. The lineage of the items needs to be documented from inclusion in the preliminary item data bank and only included if permission from the copyright holder is provided.

Additionally, as the PROMIS item pool has been developed in English for the US (sometimes also in Spanish for the US), it is critical to explore the translatability of the items as part of the project to ensure its use outside the US in international trials (e.g., cultural context, relevance and validity).

11.2. Creation of new COAs

For the development of new measurement instruments, the technology readiness level (TRL) system is a good reference system to guide the validation steps; the system is used within the technology area in the US and accepted by the EU Commission, providing a common language across stakeholders in the innovation field (European Comission, n.d.).

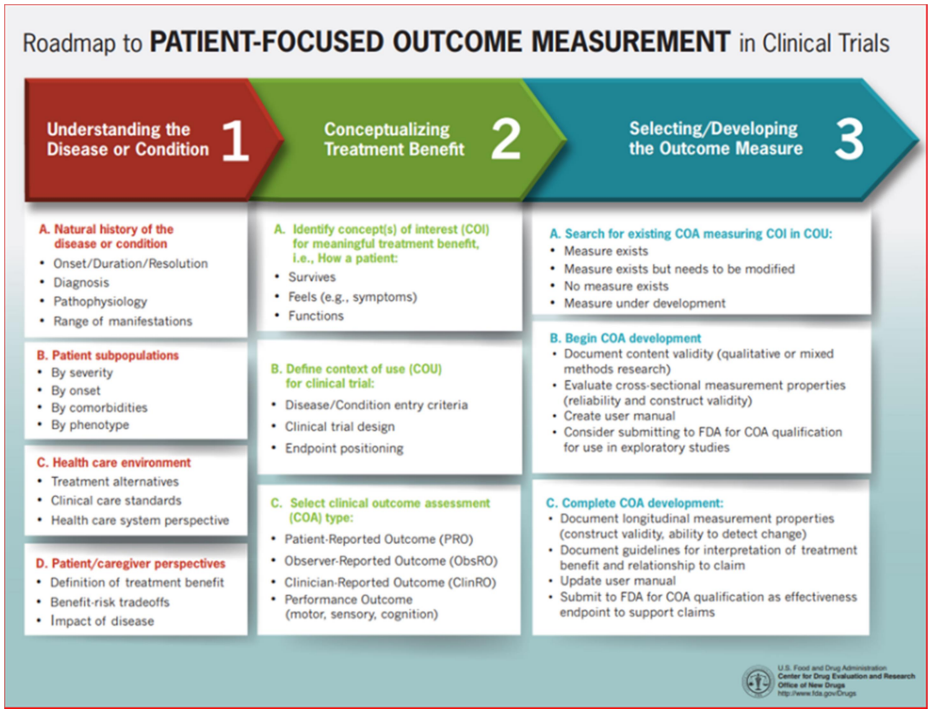
The TRL system classifies the levels of maturation, from the idea (level 1) to the final product (level 9), within the validation path according to the distance to the final use in the intended setting. The levels are organized such that levels 1 to 4 correspond to laboratory research, levels 5 to 6 correspond to simulation, and levels 7 to 9 correspond to validation in real life or a real context.

A new OMI is seen as a new technology so that it can start from the initial levels and then progress to the next levels until the research reaches its final use. Additionally, when using an existing validated OMI (level 9), the TRL system allows us to return to earlier levels depending on the validation work needed up to the new final use (new COU). Therefore, for a new use of a validated OMI, the maturity level will return to confirm the psychometric properties in the new environment or in a new format of the existing instrument.

ⓘ An example of using the TRL system for the validation of new formats of existing OMIs is the migration of PROs in paper-based formats to electronic supports such as tablets and web-based formats. In this context, a PRO migrated to an ePRO format will require returning to TRL 4 and confirming that the questionnaire migrated into the electronic platform remains faithful to the original paper version of the questionnaire and that it is appropriate for use in clinical trials. It might be necessary to complete cognitive interviews with patients suffering from the targeted disorder to confirm their understanding of the electronic version of the questionnaire (TRL 5–6) before validating the usability of the questionnaire via the IT device (TRL 7–9).

The FDA has also previously suggested a roadmap for the development of patient-focused outcome measurement tools in clinical trials (Figure 2) (FDA, n.d.-d).

Regarding PROs, the TRL system combines well with the FDA Roadmap; specifically, TRLs are related to the validation process proposed at stages 3rd B and C in Figure 2. The FDA roadmap shows all related materials to be produced in the specific context of the development and validation of OMIs for clinical trials.



12. Review and Approval of the COS Project Protocol

It is advisable to obtain feedback from a group independent of the COS research project to have a complementary view and agree on the basics.

At the ECNP, an external technical advisory board with members of the COA ECNP subgroup at the experimental medicine working group will be formed and will review and provide written comments on the methods described in the sponsor's *COA project protocol* upon request and before starting the project to ensure that the minimal guidelines are followed and the suggestions for improvement can be considered in the final version.

The list of members of the board will be disclosed and continuously updated according to the members' availability. Specific defined procedures for the independent review will be detailed upon starting the activity of the board.

13. Dissemination of a Protocol for COS Research

It is advisable to make the protocol publicly available by adding it to public databases (COMET-Initiative) or publishing it in a journal.

The purpose of this protocol is to ensure that, as these standard core sets of clinical outcome assessments are developed, the identified concepts, COAs, and endpoints reflect what is most important and relevant to patients and support regulatory and potentially other stakeholder

decision-making (ref FDA COS program). For that reason, it is important to start dissemination activities as soon as the COS protocol is defined, allowing potential users in the field to follow up on how the protocol achievements are evolving.

COS developers should be encouraged to register their project in a free-to-access, unrestricted public repository, such as the COMET database. The COMET registration page collects all relevant information regarding included COS research projects, either completed or ongoing (COMET, n.d.).

For *systematic review of the literature*, it is advisable to register the project in the PROSPERO registry (NIHR, n.d.).

13.1. Dissemination of the COS Protocol and Results

The final report of a *COS research project* is an internal comprehensive report, and it is advisable to produce a document summarizing the main methods and results that can be presented as a written communication (poster) at scientific congress-events or as a manuscript to be publishing in a journal. Williamson et al., 2012 suggested a checklist of items that should be considered when reporting the development of a COS, which can also be useful for reporting the final results. (Kirkham et al., 2017).

14. Available Grants for COS, COA or OMI Development

Due to the relevance of the research on COS in challenging disorders, there are some institutions that offer resources to boost this type of project.

- FDA COS

Pilot grant program since 2019 - https://www.fda.gov/drugs/development-approval-process-drugs/cder-pilot-grant-program-standard-core-clinical-outcome-assessments-coas-and-their-related-endpoints

- FDA - Contract Opportunities

FDABAA-21-00123 includes, among other research topics, “Tools Stimulate Innovation in Clinical Evaluations and Personalized Medicine to Improve Product Development and Patient Outcomes.” https://sam.gov/opp/f8da49472dd54ea0939f06a690f1f592/view

15. References

Acosta, M. T. (2013). Challenges of cognitive research in neurofibromatosis type 1. *The Lancet Neurology*, *12*(11), 1040–1041. https://doi.org/10.1016/S1474-4422(13)70179-0

ADCS. (n.d.). *Alzheimer Disorder Cooperative Study- Instruments*. https://www.adcs.org/instruments/

Allied Health Professions (AHP) Outcome & Measures UK Working Group. (n.d.). *Key questions to ask when selecting outcome measures: A checklist for allied health professionals*. https://www.rcslt.org/wp-content/uploads/media/docs/selecting-outcome-measures.pdf?la=en&hash=12ECB2CFDA0B2EFB1979E592A383D24E792AB9DD

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485–485. https://doi.org/10.1136/bmj.311.7003.485

Andreasen NC. (1984). *The Scale for the Assessment of Positive Symptoms (SAPS).* Iowa City, IA: The University of Iowa.

Apostolaros, M., Babaian, D., Corneli, A., Forrest, A., Hamre, G., Hewett, J., Podolsky, L., Popat, V., & Randall, P. (2020). Legal, Regulatory, and Practical Issues to Consider When Adopting Decentralized Clinical Trials: Recommendations From the Clinical Trials Transformation Initiative. *Therapeutic Innovation & Regulatory Science*, *54*(4), 779–787. https://doi.org/10.1007/s43441-019-00006-4

Arnerić, S. P., Kern, V. D., & Stephenson, D. T. (2018). Regulatory-accepted drug development tools are needed to accelerate innovative CNS disease treatments. *Biochemical Pharmacology*, *151*, 291–306. https://doi.org/10.1016/j.bcp.2018.01.043

Atkins, A. S., Tseng, T., Vaughan, A., Twamley, E. W., Harvey, P., Patterson, T., Narasimhan, M., & Keefe, R. S. E. (2017). Validation of the tablet-administered Brief Assessment of Cognition (BAC App). *Schizophrenia Research*, *181*, 100–106. https://doi.org/10.1016/j.schres.2016.10.010

Beaton, D. E., Maxwell, L. J., Shea, B. J., Wells, G. A., Boers, M., Grosskleg, S., Bingham, C. O., Conaghan, P. G., D'Agostino, M. A., de Wit, M. P., Gossec, L., March, L. M., Simon, L. S., Singh, J. A., Strand, V., & Tugwell, P. (2019). Instrument Selection Using the OMERACT Filter 2.1: The OMERACT Methodology. *The Journal of Rheumatology*, *46*(8), 1028–1035. https://doi.org/10.3899/jrheum.181218

Beaton D, Maxell L, Grosskleg S, Shea B, & Tugwell B. (2021). *The OMERACT Handbook*. https://omeract.org/handbook/

Bilderbeck, A. C., Raslescu, A., Hernaus, D., Hayen, A., Umbricht, D., Pemberton, D., Tiller, J., Søgaard, B., Sambeth, A., van Amelsvoort, T., Reif, A., Papazisis, G., Pérez, V., Elices, M., Maurice, D., Bertaina-Anglade, V., Dawson, G. R., & Pollentier, S. (2020). Optimizing Behavioral Paradigms to

Facilitate Development of New Treatments for Anhedonia and Reward Processing Deficits in Schizophrenia and Major Depressive Disorder: Study Protocol. *Frontiers in Psychiatry*, *11*, 536112. https://doi.org/10.3389/fpsyt.2020.536112

Busner J,. (2021). Clinician and patient reported endpoints in CNS orphan drug clinical trials: ISCTM position paper on best practices for endpoint selection, validation, training, and standardization: A position paper of the ISCTM Rare Disease/Orphan Drug Development Working Group . *Innovations in Clinical Neuroscience*.

Carr, M. (2010). The Small- and Medium-sized Enterprises Office (SME Office) at the European Medicines Agency. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *53*(1), 20–23. https://doi.org/10.1007/s00103-009-0989-z

CDISC. (n.d.). *CDISC Initiative – Therapeutic Area Standards*. https://wiki.cdisc.org/display/PUB/Introduction+to+Therapeutic+Area+Standards

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., … Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Children's Tumor Foundation (CTF). (2019). *FDA Listening Session on Neurofibromatosis (NF) Summary Report.* https://www.ctf.org/images/uploads/documents/Childrens_Tumor_Foundation_FDA_Listening_Session_June_1319.pdf

CIBERSAM. (n.d.). *BIBLIOPRO Cibersam*. https://bi.cibersam.es/busqueda-de-instrumentos

COMET. (n.d.). *COMET Initiative*. https://www.comet-initiative.org/Studies

COMET-Initiative. (2017). *COMET Proect 606* [Public Database]. COMET-Initiative. https://www.comet-initiative.org/Studies/Details/606

COMET-Initiative, Project 722. (n.d.). *Achieving consensus for clinical trials: The REiNS International Collaboration* [Public Database]. COMET-Initiative. https://www.comet-initiative.org/Studies/Details/722

Coons, S. J., Eremenco, S., Lundy, J. J., O'Donohoe, P., O'Gorman, H., & Malizia, W. (2015). Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. *The Patient - Patient-Centered Outcomes Research*, *8*(4), 301–309. https://doi.org/10.1007/s40271-014-0090-z

Coons, S. J., Gwaltney, C. J., Hays, R. D., Lundy, J. J., Sloan, J. A., Revicki, D. A., Lenderking, W. R., Cella, D., & Basch, E. (2009). Recommendations on Evidence Needed to Support Measurement Equivalence between Electronic and Paper-Based Patient-Reported Outcome (PRO) Measures:

ISPOR ePRO Good Research Practices Task Force Report. *Value in Health*, *12*(4), 419–429. https://doi.org/10.1111/j.1524-4733.2008.00470.x

Coran, P., Goldsack, J. C., Grandinetti, C. A., Bakker, J. P., Bolognese, M., Dorsey, E. R., Vasisht, K., Amdur, A., Dell, C., Helfgott, J., Kirchoff, M., Miller, C. J., Narayan, A., Patel, D., Peterson, B., Ramirez, E., Schiller, D., Switzer, T., Wing, L., … Doherty, A. (2019). Advancing the Use of Mobile Technologies in Clinical Trials: Recommendations from the Clinical Trials Transformation Initiative. *Digital Biomarkers*, *3*(3), 145–154. https://doi.org/10.1159/000503957

Crittenden-Ward K, Verma P, Micaleo M, Laird R, Hong B, Feaster T, Tackett ZC, Machizawa S, Zhou W, Daniel DG, & Kott A. (2020, February). *Evaluating the Impact of Rater Change on PANSS Score Change: An Item Level Analysis.* [Poster ISCTM]. https://isctm.org/public_access/Autumn2020/Posters/Verma_2020_Autumn_Poster.pdf

Crudgington, H., Rogers, M., Bray, L., Carter, B., Currier, J., Dunkley, C., Gibbon, F. M., Hughes, D., Lyle, S., Roberts, D., Tudur Smith, C., Gringras, P., Pal, D. K., & Morris, C. (2019). Core Health Outcomes in Childhood Epilepsy (CHOICE): Development of a core outcome set using systematic review methods and a Delphi survey consensus. *Epilepsia*, *60*(5), 857–871. https://doi.org/10.1111/epi.14735

CTTI. (n.d.-a). *Clinical Trials Transformation Initiative*. https://www.ctti-clinicaltrials.org

CTTI. (n.d.-b). *CTTI 1: Clinical Trials Transformation Initiative. Program: Mobile Clinical Trials (MCT). Clinical Trials Transformation Initiative.* https://ctti-clinicaltrials.org/our-work/digital-health-trials/original-digital-health-trials-projects/

CTTI. (2015). *Approaches to mitigate risks by The Clinical Trials Transformation Initiative .* https://ctti-clinicaltrials.org/type/news/ctti-releases-new-recommendations-toolkit-to-help-organizations-implement-quality-by-design/

CTTI. (2018). *CTTI recommendations: Decentralized clinical trials*. https://ctti-clinicaltrials.org/topics/mobile/new-ctti-recommendations-offer-path-forward-for-decentralized-clinical-trials/

Dashiell-Aje B, Kovaks S, & Sacks L. (2019). *Regulatory Perspective: Digital Health Technology Tools Use in Clinical Investigations to Evaluate Clinical Benefit in Patients. Dia Global Forum. June 2019*. Dia Global Forum. https://globalforum.diaglobal.org/issue/june-2019/regulatory-perspective-digital-health-technology-tools/

de Roos, P., Bloem, B. R., Kelley, T. A., Antonini, A., Dodel, R., Hagell, P., Marras, C., Martinez-Martin, P., Mehta, S. H., Odin, P., Chaudhuri, K. R., Weintraub, D., Wilson, B., & Uitti, R. J. (2017). A Consensus Set of Outcomes for Parkinson's Disease from the International Consortium for Health Outcomes Measurement. *Journal of Parkinson's Disease*, *7*(3), 533–543. https://doi.org/10.3233/JPD-161055

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of Item Candidates: The PROMIS Qualitative Item Review. *Medical Care*, *45*(5), S12–S21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2

Dodd, S., Clarke, M., Becker, L., Mavergames, C., Fish, R., & Williamson, P. R. (2018). A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of Clinical Epidemiology*, *96*, 84–92. https://doi.org/10.1016/j.jclinepi.2017.12.020

EBSCO. (n.d.). *Health and Psychosocial Instruments (HaPI)* [Private Database Instruments]. https://www.ebsco.com/products/research-databases/health-and-psychosocial-instruments-hapi

Eichler, H.-G., & Sweeney, F. (2018). The evolution of clinical trials: Can we address the challenges of the future? *Clinical Trials*, *15*(1_suppl), 27–32. https://doi.org/10.1177/1740774518755058

EMA. (n.d.-a). *EMA Registry*. https://www.clinicaltrialsregister.eu/ctr-search/search

EMA. (n.d.-b). *Qualification Opinion MS COA*. https://www.ema.europa.eu/en/documents/other/qualification-opinion-multiple-sclerosis-clinical-outcome-assessment-mscoa_en.pdf

EMA. (n.d.-c). *Questions and answers: Qualification of digital technology-based methodologies to support approval of medicinal products.* https://www.ema.europa.eu/en/documents/other/questions-answers-qualification-digital-technology-based-methodologies-support-approval-medicinal_en.pdf

EMA. (2005). *EMA SME office*. https://www.ema.europa.eu/en/human-regulatory/overview/supporting-smes

EMA. (2020a). *Innovation in Medicine*. https://www.ema.europa.eu/en/human-regulatory/research-development/innovation-medicines#ema's-innovation-task-force-(itf)-section

EMA. (2020b). *Qualification of Novel Methodologies for Medicine Development by EMA (access 13 June 2021)* [Regulatory]. Qualification of Novel Methodologies for Medicine Development. https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance/qualification-novel-methodologies-medicine-development-0

EMA. (2021). *Guidance on the Management of Clinical Trials During the COVID-19 (Coronavirus) Pandemic.* https://ec.europa.eu/health/sites/default/files/files/eudralex/vol-10/guidanceclinicaltrials_covid19_en.pdf

EMA EWP. (n.d.). *Opinions and letters of support on the qualification of novel methodologies for medicine development*. https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance/novel-methodologies-biomarkers/opinions-letters-support-qualification-novel-methodologies-medicine-development

EU European Comission. (n.d.). *CORDIS Research Results* [Public Resource]. https://cordis.europa.eu/

European Comission. (n.d.). *Technology readiness levels (TRL)*. https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf

Fackrell, K., Smith, H., Colley, V., Thacker, B., Horobin, A., Haider, H. F., Londero, A., Mazurek, B., & Hall, D. A. (2017). Core Outcome Domains for early phase clinical trials of sound-, psychology-, and pharmacology-based interventions to manage chronic subjective tinnitus in adults: The COMIT'ID study protocol for using a Delphi process and face-to-face meetings to establish consensus. *Trials*, *18*(1), 388. https://doi.org/10.1186/s13063-017-2123-0

FDA. (n.d.-a). *Clinical Outcome Assessments (COA) Qualification Program Submissions*. Retrieved 17 December 2020, from https://www.fda.gov/drugs/clinical-outcome-assessment-coa-qualification-program/clinical-outcome-assessments-coa-qualification-program-submissions

FDA. (n.d.-b). *Clinical Outcome Assessments (COA) Qualification Program: Symbol Digit Modalities Test Qualification Plan*. https://www.fda.gov/media/139873/download

FDA. (n.d.-c). *FDA COMPENDIUM*. https://www.fda.gov/media/130138/download

FDA. (n.d.-d). *ROADMAP to Patient-Focused Outcomes Instruments*. https://www.fda.gov/media/87004/download

FDA. (2009). *Guidance for industry:patient-reported outcome measures: Use in medical product development to support labeling claims.* http://www.fda.gov/dowloads/Drugs/

FDA. (2020a, June). *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. Guidance 3* [Public Health Regulatory]. https://www.fda.gov/media/139088/download

FDA. (2020b). *Clinical Outcome Assessments (COA) Qualification Program Resources*. https://www.fda.gov/drugs/clinical-outcome-assessment-coa-qualification-program/clinical-outcome-assessments-coa-qualification-program-resources

FDA. (2021a). *Fit-for-Purpose Initiative*. https://www.fda.gov/drugs/development-approval-process-drugs/drug-development-tools-fit-purpose-initiative

FDA. (2021b, July 9). *Digital Health Center of Excellence*. https://www.fda.gov/medical-devices/digital-health-center-excellence

FDA. (2021c). *FDA Guidance on Conduct of Clinical Trials of Medical Products During the COVID-19 Public Health Emergency*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/fda-guidance-conduct-clinical-trials-medical-products-during-covid-19-public-health-emergency

FDA-NIH Biomarker Working Group. (2021). *BEST (Biomarkers, EndpointS, and other Tools)*. Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US); https://www.ncbi.nlm.nih.gov/books/NBK338448/

for the COS-STAP Group, Kirkham, J. J., Gorst, S., Altman, D. G., Blazeby, J. M., Clarke, M., Tunis, S., & Williamson, P. R. (2019). Core Outcome Set-STAndardised Protocol Items: The COS-STAP Statement. *Trials*, *20*(1), 116. https://doi.org/10.1186/s13063-019-3230-x

Formal Meetings Between the FDA and Sponsors or Applicants of PDUFA Products Guidance for Industry. (2017). 20.

Gaasterland, C. M. W., van der Weide, M. C. J., Roes, K. C. B., & van der Lee, J. H. (2019). Goal attainment scaling as an outcome measure in rare disease trials: A conceptual proposal for validation. *BMC Medical Research Methodology*, *19*(1), 227. https://doi.org/10.1186/s12874-019-0866-x

Geraci, J., Wong, B., Ziauddin, Jalal, Jain, Siddhant, Leonczky, Paul, & Bishop, Kim. (2018). *Classical and Quantum Machine Learning Applied To Predicting Placebo Response For Clinical Trials In Bipolar Disorder: Recent Results*. ISCTM, Washington. https://isctm.org/public_access/Autumn2018/PDFs/Geraci-Poster.pdf

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., … LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Green, M. F., Kern, R. S., & Heaton, R. K. (2004). Longitudinal studies of cognition and functional outcome in schizophrenia: Implications for MATRICS. *Schizophrenia Research*, *72*(1), 41–51. https://doi.org/10.1016/j.schres.2004.09.009

Green, M. F., Nuechterlein, K. H., Gold, J. M., Barch, D. M., Cohen, J., Essock, S., Fenton, W. S., Frese, F., Goldberg, T. E., Heaton, R. K., Keefe, R. S. E., Kern, R. S., Kraemer, H., Stover, E., Weinberger, D. R., Zalcman, S., & Marder, S. R. (2004). Approaching a consensus cognitive battery for clinical trials in schizophrenia: The NIMH-MATRICS conference to select cognitive domains and test criteria. *Biological Psychiatry*, *56*(5), 301–307. https://doi.org/10.1016/j.biopsych.2004.06.023

Hall, D. A., Zaragoza Domingo, S., Hamdache, L. Z., Manchaiah, V., Thammaiah, S., Evans, C., Wong, L. L. N., & On behalf of the International Collegium of Rehabilitative Audiology and TINnitus Research NETwork. (2018). A good practice guide for translating and adapting hearing-related questionnaires for different languages and cultures. *International Journal of Audiology*, *57*(3), 161–175. https://doi.org/10.1080/14992027.2017.1393565

Harmer, C. J., O'Sullivan, U., Favaron, E., Massey-Chase, R., Ayres, R., Reinecke, A., Goodwin, G. M., & Cowen, P. J. (2009). Effect of Acute Antidepressant Administration on Negative Affective Bias in

Depressed Patients. *American Journal of Psychiatry*, *166*(10), 1178–1184. https://doi.org/10.1176/appi.ajp.2009.09020149

Haro JM & Domenech C. (2019, April). Agreement of the CGI SCH scale with the Marder's PANSS positive and negative score. *European Psychiatry*, *56 Supplemenet*, S1–S662.

IMIM. (n.d.). *Biblioteca Virtual de PROs*. BiblioPRO. https://www.bibliopro.org

Inan, O. T., Tenaerts, P., Prindiville, S. A., Reynolds, H. R., Dizon, D. S., Cooper-Arnold, K., Turakhia, M., Pletcher, M. J., Preston, K. L., Krumholz, H. M., Marlin, B. M., Mandl, K. D., Klasnja, P., Spring, B., Iturriaga, E., Campo, R., Desvigne-Nickens, P., Rosenberg, Y., Steinhubl, S. R., & Califf, R. M. (2020). Digitizing clinical trials. *Npj Digital Medicine*, *3*(1), 101. https://doi.org/10.1038/s41746-020-0302-y

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Iorio, A., Skinner, M. W., Clearfield, E., Messner, D., Pierce, G. F., Witkop, M., Tunis, S., & the coreHEM panel. (2018). Core outcome set for gene therapy in haemophilia: Results of the coreHEM multistakeholder project. *Haemophilia*, *24*(4), e167–e172. https://doi.org/10.1111/hae.13504

Katiri R, H. D. (2020). Developing an international core outcome set for SSD interventions. *Hearing J*, *73*(4), 41–43.

Kings College London. The North West London Hospitals NHS Trust,. (2009). *Goal Attainment Scaling (GAS) in Rehabilitation: A Practical Guide. . 2009*.

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4*(6), 443–453. https://doi.org/10.1007/BF01530764

Kirkham, J. J., Davis, K., Altman, D. G., Blazeby, J. M., Clarke, M., Tunis, S., & Williamson, P. R. (2017). Core Outcome Set-STAndards for Development: The COS-STAD recommendations. *PLOS Medicine*, *14*(11), e1002447. https://doi.org/10.1371/journal.pmed.1002447

Kott A, Wang X, & Daniel D. (2020, February). *Exploring the association of PANSS rater change with extreme within-subject variability in schizophrenia clinical trials.* [Poster]. https://isctm.org/public_access/Feb2019/Posters/Kott-poster.pdf

Kottner, J., Jacobi, L., Hahnel, E., Alam, M., Balzer, K., Beeckman, D., Busard, C., Chalmers, J., Deckert, S., Eleftheriadou, V., Furlan, K., Horbach, S. E. R., Kirkham, J., Nast, A., Spuls, P., Thiboutot, D., Thorlacius, L., Weller, K., Williams, H. C., … International Cochrane Skin Group Core Outcome Set Initiative (CSG-COUSIN) group. (2018). Core outcome sets in dermatology: Report from the second meeting of the International Cochrane Skin Group Core Outcome Set Initiative. *British Journal of Dermatology*, *178*(4). https://doi.org/10.1111/bjd.16324

Langdon, D., Amato, M., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hämäläinen, P., Hartung, H.-P., Krupp, L., Penner, I., Reder, A., & Benedict, R. (2012). Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *Multiple Sclerosis Journal*, *18*(6), 891–898. https://doi.org/10.1177/1352458511431076

Lange, T., Kottner, J., Weberschock, T., Hahnel, E., Apfelbacher, C., Brandstetter, S., Dreher, A., Datzmann, T., Burden-Teh, E., Rogers, N. K., Spuls, P., Grainge, M. J., Jacobi, L., Williams, H. C., & Schmitt, J. (2021). Outcome assessment in dermatology clinical trials and cochrane reviews: Call for a dermatology-specific outcome taxonomy. *Journal of the European Academy of Dermatology and Venereology*, *35*(2), 523–535. https://doi.org/10.1111/jdv.16854

LaRocca, N. G., Hudson, L. D., Rudick, R., Amtmann, D., Balcer, L., Benedict, R., Bermel, R., Chang, I., Chiaravalloti, N. D., Chin, P., Cohen, J. A., Cutter, G. R., Davis, M. D., DeLuca, J., Feys, P., Francis, G., Goldman, M. D., Hartley, E., Kapoor, R., … for the Multiple Sclerosis Outcome Assessments Consortium (MSOAC). (2018). The MSOAC approach to developing performance outcomes to measure and monitor multiple sclerosis disability. *Multiple Sclerosis Journal*, *24*(11), 1469–1484. https://doi.org/10.1177/1352458517723718

Lucchetta R C. (n.d.). *Core outcome set for relapsing multiple sclerosis in adults* [Public Database]. COMET-Initiative. https://www.comet-initiative.org/Studies/Details/1559

MAPI Research Trust. (n.d.). *EPPROVIDE* [Public Database]. https://eprovide.mapi-trust.org/about/about-proqolid

Marder, S. R., Davidson, M., Zaragoza, S., Kott, A., Khan, A., Wang, X., Velligan, D. I., Umbricht, D., Luthringer, R., & Daniel, D. (2020). Issues and Perspectives in Designing Clinical Trials for Negative Symptoms in Schizophrenia: Consensus Statements. *Schizophrenia Bulletin Open*, *1*(1), sgz001. https://doi.org/10.1093/schizbullopen/sgz001

Marder, S. R., & Fenton, W. (2004). Measurement and Treatment Research to Improve Cognition in Schizophrenia: NIMH MATRICS initiative to support the development of agents for improving cognition in schizophrenia. *Schizophrenia Research*, *72*(1), 5–9. https://doi.org/10.1016/j.schres.2004.09.010

Martin, S., Olayinka-Amao, O., Henke, C., Kamudoni, P., & Gnanasakthy, A. (2017). Conceptual Model and Instrument Review in Multiple Sclerosis. *Value in Health*, *20*(9), A726–A727. https://doi.org/10.1016/j.jval.2017.08.1967

Maxwell, L. J., Beaton, D. E., Shea, B. J., Wells, G. A., Boers, M., Grosskleg, S., Bingham, C. O., Conaghan, P. G., D'Agostino, M. A., de Wit, M., Gossec, L., March, L., Simon, L. S., Singh, J. A., Strand, V., & Tugwell, P. (2019). Core Domain Set Selection According to OMERACT Filter 2.1: The OMERACT Methodology. *The Journal of Rheumatology*, *46*(8), 1014–1020. https://doi.org/10.3899/jrheum.181097

McGarrigle, L., & Rockwood, K. (2020). The responsiveness of goal attainment scaling using just one goal in controlled clinical trials: An exploratory analysis. *Journal of Patient-Reported Outcomes*, *4*(1), 35. https://doi.org/10.1186/s41687-020-00196-8

MDS. (n.d.). *MDS Rating Scales*. https://www.movementdisorders.org/MDS/MDS-Rating-Scales.htm

Miller, D. S., Robert, P., Ereshefsky, L., Adler, L., Bateman, D., Cummings, J., DeKosky, S. T., Fischer, C. E., Husain, M., Ismail, Z., Jaeger, J., Lerner, A. J., Li, A., Lyketsos, C. G., Manera, V., Mintzer, J., Moebius, H. J., Mortby, M., Meulien, D., … Lanctôt, K. L. (2021). Diagnostic criteria for apathy in neurocognitive disorders. *Alzheimer's & Dementia*, alz.12358. https://doi.org/10.1002/alz.12358

Miyagawa, T., Brushaber, D., Syrjanen, J., Kremers, W., Fields, J., Forsberg, L. K., Heuer, H. W., Knopman, D., Kornak, J., Boxer, A., Rosen, H., Boeve, B., & on Behalf of the ARTFL/LEFFTDS Consortium. (2020). Use of the CDR® plus NACC FTLD in mild FTLD: Data from the ARTFL/LEFFTDS consortium. *Alzheimer's & Dementia*, *16*(1), 79–90. https://doi.org/10.1016/j.jalz.2019.05.013

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, *19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Nagy, M., Radakovich, N., & Nazha, A. (2020). Machine Learning in Oncology: What Should Clinicians Know? *JCO Clinical Cancer Informatics*, *4*, 799–810. https://doi.org/10.1200/CCI.20.00049

NCT02972632. (n.d.). https://www.clinicaltrials.gov/ct2/show/NCT02972632?term=Goal+attainment+scaling&type=Intr&phase=123&fund=2&draw=2&rank=4

Needham, D. M., Sepulveda, K. A., Dinglas, V. D., Chessare, C. M., Friedman, L. A., Bingham, C. O., & Turnbull, A. E. (2017). Core Outcome Measures for Clinical Research in Acute Respiratory Failure Survivors. An International Modified Delphi Consensus Study. *American Journal of Respiratory and Critical Care Medicine*, *196*(9), 1122–1130. https://doi.org/10.1164/rccm.201702-0372OC

Needham DN. (n.d.). *ImproveLTO*. www.improveLTO.com

NIHR. (n.d.). *PROSPERO International prospective register of systematic reviews*. https://www.crd.york.ac.uk/prospero

NIMH. (n.d.). *National Institute of Mental Health. Development and definitions of the RDoC domains and constructs.* [Public Health]. Research Domain Criteria. Retrieved 5 August 2021, from https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/index.shtml

on behalf of the AddNeuroMed Consortium, Birkenbihl, C., Westwood, S., Shi, L., Nevado-Holgado, A., Westman, E., Lovestone, S., & Hofmann-Apitius, M. (2021). ANMerge: A Comprehensive and Accessible Alzheimer's Disease Patient-Level Dataset. *Journal of Alzheimer's Disease*, *79*(1), 423–431. https://doi.org/10.3233/JAD-200948

Opler M, Ruble L, & McGrew J. (2018). *D. Goal Attainment Scaling for Depression. Technical Manual & Training Guide.*

Perry, B., Herrington, W., Goldsack, J. C., Grandinetti, C. A., Vasisht, K. P., Landray, M. J., Bataille, L., DiCicco, R. A., Bradley, C., Narayan, A., Papadopoulos, E. J., Sheth, N., Skodacek, K., Stem, K., Strong, T. V., Walton, M. K., & Corneli, A. (2018). Use of Mobile Devices to Measure Outcomes in Clinical Research, 2010–2016: A Systematic Literature Review. *Digital Biomarkers*, *2*(1), 11–30. https://doi.org/10.1159/000486347

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item Banks for Measuring Emotional Distress From the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, Anxiety, and Anger. *Assessment*, *18*(3), 263–283. https://doi.org/10.1177/1073191111411667

Plotkin, S. R., Blakeley, J. O., Dombi, E., Fisher, M. J., Hanemann, C. O., Walsh, K. S., Wolters, P. L., & Widemann, B. C. (2013). Achieving consensus for clinical trials: The REiNS International Collaboration. *Neurology*, *81*(Issue 21, Supplement 1), S1–S5. https://doi.org/10.1212/01.wnl.0000435743.49414.b6

Powers, J. H., Patrick, D. L., Walton, M. K., Marquis, P., Cano, S., Hobart, J., Isaac, M., Vamvakas, S., Slagle, A., Molsen, E., & Burke, L. B. (2017). Clinician-Reported Outcome Assessments of Treatment Benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. *Value in Health*, *20*(1), 2–14. https://doi.org/10.1016/j.jval.2016.11.005

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials*, *17*(1), 449. https://doi.org/10.1186/s13063-016-1555-2

PROMIS. (n.d.). *PROMIS Initiative*. https://www.healthmeasures.net/search-view-measures

Rabinowitz, J., Opler, M., Rabinowitz, A. A., Negash, S., Anderson, A., Fu, D. J., Williamson, D., Kott, A., Davis, L. L., & Schooler, N. R. (2021). Consistency checks to improve measurement with the Personal and Social Performance Scale (PSP). *Schizophrenia Research*, *228*, 529–533. https://doi.org/10.1016/j.schres.2020.11.040

Rabinowitz, J., Schooler, N. R., Anderson, A., Ayearst, L., Daniel, D., Davidson, M., Khan, A., Kinon, B., Menard, F., Opler, L., Opler, M., Severe, J. B., Williamson, D., Yavorsky, C., & Zhao, J. (2017). Consistency checks to improve measurement with the Positive and Negative Syndrome Scale (PANSS). *Schizophrenia Research*, *190*, 74–76. https://doi.org/10.1016/j.schres.2017.03.017

Rabinowitz, J., Schooler, N. R., Brown, B., Dalsgaard, M., Engelhardt, N., Friedberger, G., Kinon, B. J., Lee, D., Ockun, F., Mahableshwarkar, A., Tsai, J., Williams, J. B. W., Sauder, C., & Yavorsky, C. (2019). Consistency checks to improve measurement with the Montgomery-Asberg Depression Rating Scale (MADRS). *Journal of Affective Disorders*, *256*, 143–147. https://doi.org/10.1016/j.jad.2019.05.077

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J., & Cella, D. (2007). Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for

the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Retzer, A., Sayers, R., Pinfold, V., Gibson, J., Keeley, T., Taylor, G., Plappert, H., Gibbons, B., Huxley, P., Mathers, J., Birchwood, M., & Calvert, M. (2020). Development of a core outcome set for use in community-based bipolar trials—A qualitative study and modified Delphi. *PLOS ONE*, *15*(10), e0240518. https://doi.org/10.1371/journal.pone.0240518

Richardson, E., Burnell, J., Adams, H. R., Bohannon, R. W., Bush, E. N., Campbell, M., Chen, W. H., Coons, S. J., Papadopoulos, E., Reeve, B. R., Rooks, D., & Daniel, G. (2019). Developing and Implementing Performance Outcome Assessments: Evidentiary, Methodologic, and Operational Considerations. *Therapeutic Innovation & Regulatory Science*, *53*(1), 146–153. https://doi.org/10.1177/2168479018772569

Rose T. (2020, September 22). *Including families in protocol design, protocol simulation, identification of meaningful clinical aspects, with patients* [Presentation]. ISCTM AUtumn 2020. https://isctm.org/public_access/Autumn2020/Presentation/Traceann_Rose_slides.pdf

Rudick, R. A., LaRocca, N., Hudson, L. D., & MSOAC. (2014). Multiple Sclerosis Outcome Assessments Consortium: Genesis and initial project plan. *Multiple Sclerosis Journal*, *20*(1), 12–17. https://doi.org/10.1177/1352458513503392

Sanders, S. J., Sahin, M., Hostyk, J., Thurm, A., Jacquemont, S., Avillach, P., Douard, E., Martin, C. L., Modi, M. E., Moreno-De-Luca, A., Raznahan, A., Anticevic, A., Dolmetsch, R., Feng, G., Geschwind, D. H., Glahn, D. C., Goldstein, D. B., Ledbetter, D. H., Mulle, J. G., … Bearden, C. E. (2019). A framework for the investigation of rare genetic disorders in neuropsychiatry. *Nature Medicine*, *25*(10), 1477–1487. https://doi.org/10.1038/s41591-019-0581-5

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, *52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8

Shear, M. K., Vander Bilt, J., Rucci, P., Endicott, J., Lydiard, B., Otto, M. W., Pollack, M. H., Chandler, L., Williams, J., Ali, A., & Frank, D. M. (2001). Reliability and validity of a structured interview guide for the Hamilton Anxiety Rating Scale (SIGH-A). *Depression and Anxiety*, *13*(4), 166–178. https://doi.org/10.1002/da.1033

SRLab. (n.d.). *The Rehabilitation Measure Database*. Https://Www.Sralab.Org/Rehabilitation-Measures/Database.

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., McGee, M. F., Simpson, G. M., Stevens, M. C., & Lieberman, J. A. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Project: Schizophrenia Trial Design and Protocol Development. *Schizophrenia Bulletin*, *29*(1), 15–31. https://doi.org/10.1093/oxfordjournals.schbul.a006986

Sullivan EJ. (2012). *2012 Clinical Investigator Course—Clinical Trial End Points* [FDA Presentation Course]. https://www.fda.gov/media/84987/download

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research*, *27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Tunis, S. R., Maxwell, L. J., Graham, I. D., Shea, B. J., Beaton, D. E., Bingham, C. O., Brooks, P., Conaghan, P. G., D'Agostino, M. A., de Wit, M. P., Gossec, L., March, L. M., Simon, L. S., Singh, J. A., Strand, V., Wells, G. A., & Tugwell, P. (2017). Engaging Stakeholders and Promoting Uptake of OMERACT Core Outcome Instrument Sets. *The Journal of Rheumatology*, *44*(10), 1551–1559. https://doi.org/10.3899/jrheum.161273

Umbricht, D., Kott, A., & Daniel, D. G. (2020). The Effects of Erratic Ratings on Placebo Response and Signal Detection in the Roche Bitopertin Phase 3 Negative Symptom Studies—A Post Hoc Analysis. *Schizophrenia Bulletin Open*, *1*(1), sgaa040. https://doi.org/10.1093/schizbullopen/sgaa040

Urach, S., Gaasterland, C., Posch, M., Jilma, B., Roes, K., Rosenkranz, G., Van der Lee, J., & Ristl, R. (2019). Statistical analysis of Goal Attainment Scaling endpoints in randomised trials. *Statistical Methods in Medical Research*, *28*(6), 1893–1910. https://doi.org/10.1177/0962280218777896

Valderas, J. M., Ferrer, M., Mendívil, J., Garin, O., Rajmil, L., Herdman, M., & Alonso, J. (2008). Development of EMPRO: A Tool for the Standardized Assessment of Patient-Reported Outcome Measures. *Value in Health*, *11*(4), 700–708. https://doi.org/10.1111/j.1524-4733.2007.00309.x

van der Doef, T. F., Zaragoza Domingo, S., Jacobs, G. E., Drevets, W. C., Marston, H. M., Nathan, P. J., Tome, M. B., Tamminga, C. A., van Gerven, J. M. A., & Kas, M. J. H. (2018). New approaches in psychiatric drug development. *European Neuropsychopharmacology*, *28*(9), 983–993. https://doi.org/10.1016/j.euroneuro.2018.06.006

Voss, T., Bahr, D., Cummings, J., Mills, R., Ravina, B., & Williams, H. (2013). Performance of a shortened Scale for Assessment of Positive Symptoms for Parkinson's disease psychosis. *Parkinsonism & Related Disorders*, *19*(3), 295–299. https://doi.org/10.1016/j.parkreldis.2012.10.022

Voss, T. S., Brocht, A. F. D., & Ravina, B. (2010). Performance of the scale for assessment of positive symptoms in Parkinson's disease psychosis: Letters to the Editor. *Movement Disorders*, *25*(1), 124–125. https://doi.org/10.1002/mds.22575

Walsh, K. S., Janusz, J., Wolters, P. L., Martin, S., Klein-Tasman, B. P., Toledo-Tamula, M. A., Thompson, H. L., Payne, J. M., Hardy, K. K., de Blank, P., Semerjian, C., Gray, L. S., Solomon, S. E., Ullrich, N., & For the REiNS International Collaboration. (2016). Neurocognitive outcomes in neurofibromatosis clinical trials: Recommendations for the domain of attention. *Neurology*, *87*(7 Supplement 1), S21–S30. https://doi.org/10.1212/WNL.0000000000002928

WHO. (n.d.). *International Classification of Functioning, Disability and Health (ICF) (access 14 June 2021)*. https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Willgoss, T., Cassater, D., Connor, S., Krishnan, M. L., Miller, M. T., Dias-Barbosa, C., Phillips, D., McCormack, J., Bird, L. M., Burdine, R. D., Claridge, S., & Bichell, T. J. (2021). Measuring What Matters to Individuals with Angelman Syndrome and Their Families: Development of a Patient-Centered Disease Concept Model. *Child Psychiatry & Human Development*, *52*(4), 654–668. https://doi.org/10.1007/s10578-020-01051-z

Williams, J. B. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry*, *45*(8), 742–747. https://doi.org/10.1001/archpsyc.1988.01800320058007

Williams, J. B. W., & Kobak, K. A. (2008). Development and reliability of a structured interview guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA). *British Journal of Psychiatry*, *192*(1), 52–58. https://doi.org/10.1192/bjp.bp.106.032532

Williams, J. B. W., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., Olin, J., Pearson, J., & Kalali, A. (2008). The GRID-HAMD: Standardization of the Hamilton Depression Rating Scale: *International Clinical Psychopharmacology*, *23*(3), 120–129. https://doi.org/10.1097/YIC.0b013e3282f948f5

Williamson, P. R., Altman, D. G., Bagley, H., Barnes, K. L., Blazeby, J. M., Brookes, S. T., Clarke, M., Gargon, E., Gorst, S., Harman, N., Kirkham, J. J., McNair, A., Prinsen, C. A. C., Schmitt, J., Terwee, C. B., & Young, B. (2017). The COMET Handbook: Version 1.0. *Trials*, *18*(S3), 280. https://doi.org/10.1186/s13063-017-1978-4

Williamson, P. R., Altman, D. G., Blazeby, J. M., Clarke, M., Devane, D., Gargon, E., & Tugwell, P. (2012). Developing core outcome sets for clinical trials: Issues to consider. *Trials*, *13*(1), 132. https://doi.org/10.1186/1745-6215-13-132

Wise-Rankovi A, Daniel D, Opler M, & Kalali A. (2014). Consensus Recommendations on Rater Training and Certification. Link to document: *Innov Clin Neurosci.*, *11(11–12)*, 10–13.

Xue, J. Z., Smietana, K., Poda, P., Webster, K., Yang, G., & Agrawal, G. (2020). Clinical trial recovery from COVID-19 disruption. *Nature Reviews. Drug Discovery*, *19*(10), 662–663. https://doi.org/10.1038/d41573-020-00150-9

YODA. (n.d.). *Yale Open Data Access (YODA)*. https://yoda.yale.edu/

Younis, I. R., Gopalakrishnan, M., Mathis, M., Mehta, M., Uppoor, R., Zhu, H., & Farchione, T. (2020). Association of End Point Definition and Randomized Clinical Trial Duration in Clinical Trials of

Schizophrenia Medications. *JAMA Psychiatry*, *77*(10), 1064. https://doi.org/10.1001/jamapsychiatry.2020.1596

APPENDIX 1 COA SELECTION WORKFLOW

Proposed Standard Protocol for Developing a COS Project

Standard Flow Chart

1. **Step 1 – Disease Model and Background**
    1.1. Review of existing literature
    1.2. Review of existing qualitative studies or disease models
    1.3. Defining the unmet need (or gaps to fill)
    1.4. Describing "drug targets" and "drug actions": actual/desired
2. **Step 2 – Defining the Scope of Use for the Clinical Outcomes Set**
    2.1. Health condition, condition stage, target patient population
    2.2. Defining the intervention
    2.3. Defining setting
    2.4. COS application in a "*nutshell*"
3. **Step 3 - Stakeholder Involvement**
    3.1. Key stakeholders
    3.2. Balancing stakeholders
4. **Step 4 - Determining "*What to Measure*" as Dimensions**
    4.1. Defining the method for identifying dimensions
    4.2. Building a disease impact model
    4.3. Setting of candidate domains
    4.4. Delphi technique
    4.5. Determining the COS
5. **Step 5 - Determining "*How to Measure*" Outcomes**
    5.1. Conceptual definitions of dimensions under consensus
    5.2. Identifying existing instruments
        5.2.1. Defining the method to select instruments (literature review vs frequency of use)

APPENDIX 2 PROPOSED DATA EXTRACTION FROM LITERATURE REVIEW

For COAs the following measurement properties can be reported depending on their use in the research plan; additionally, the following key words can be used for reviews of the literature.

- Purpose of development
- Therapeutic indication of development
- Population of use
- Type of COA
- Mode of administration
- Data collection mode
- Domains
- Response scales/scoring and cut-off scores
- Number of items
- Recall period
- Time of administration
- Reliability
    - Internal consistency reliability
    - Test-retest reliability
    - Inter-rater or inter-interviewer reliability
- Validity
    - Face/content validity
    - Construct validity
- Item-scale relationships (e.g., factor analysis, multitrait analysis, item total correlations)
- Floor and ceiling effects
- Concurrent/convergent validity
- Divergent/discriminant validity
- Known-group/clinical validity
    - Ability to detect change/responsiveness
    - MCID/MID
    - Specificity/sensitivity

APPENDIX 3 ITEMS TO INCLUDE WHEN REPORTING RESULTS FROM COS PROJECTS

Table 1 Checklist of the items that groups should consider when reporting the development of a COS of domain concepts (that is, 'what' to measure)

From: Developing core outcome sets for clinical trials: issues to consider

Section/topic	#	Checklist item
Title	1	Identify the report as a study to develop a COS.
Structured summary	2	Provide a structured summary including, as applicable: background, objectives, data sources, participant eligibility criteria, study methods, results, limitations, conclusions, and implications of key findings.
Rationale	3	Describe the rationale for the development of a COS in the context of what is already known. This may include a review of outcomes in previous trials or systematic reviews.
Objectives	4	Provide an explicit statement of questions being addressed with reference, as applicable, to: health condition, population, and types of intervention(s).
Protocol and registration	5	Indicate if a study protocol exists, and where it can be accessed (for example, web address)
Eligibility criteria	6	Specify participant eligibility criteria, including stakeholder group, the rationale for involving them, and how participants were identified and sampled.
Information sources	7	Describe all information sources (for example, systematic review, databases with dates of coverage, contact with study authors) provided to participants before the start of and during the consensus process. If no information on previously measured outcomes is provided, this should be clearly stated together with details of the method for obtaining information on outcomes of importance from the participants.
Consensus process	8	Describe method to determine consensus and the rationale. A checklist for reporting Delphi methods applied to the development of COS has previously been recommended [29].
Outcome scoring	9	Describe how outcomes will be scored during the consensus exercise, and how scores will be summarized across participants during each stage of the consensus process.
Definition of consensus	10	Clearly describe any pre-determined definition of consensus. Describe procedure for determining how outcomes will be included or excluded from consideration at each stage of the consensus process.
Participants	11	Give the total number of participants invited and the number involved in each aspect of the study. Give the proportion of each type of participant from the various stakeholder groups involved. Present any data collected on participant characteristics.
Results of the consensus process	12	As a minimum, provide a comprehensive list of all the outcomes that participants agreed should be included in the core set. Describe a measure of group response and distribution of response for each outcome considered during the process.
Summary of evidence	13	Summarize the main findings regarding the level of consensus and the content of the COS. Consider its relevance to key groups e.g. patients and the public, healthcare providers, and policy makers, and any potential barriers to implementation.
Limitations	14	Discuss limitations in terms of stakeholder and geographical coverage. Describe methods used for assessing risk of bias, in relation to information provided to participants beforehand, attrition, any lack of anonymity, etc.
Conclusions	15	Provide a general interpretation of the results in the context of other evidence, and implications for future research.
Funding	16	Describe sources of funding and the role of the funder in the study.
Conflicts of interest	17	Describe any conflicts of interest within the study team, for example researchers who have developed an outcome measurement instrument applicable to the scope of the COS.

From: Williamson et al. Trials 2012, 13:132 Page 5 of 8, http://www.trialsjournal.com/content/13/1/132

APPENDIX 4 LIST OF MEMBERS OF THE ECNP COA SUBGROUP

Last updated on September 1st, 2021

	Name	Affiliation
1.	Alexandra Atkins	VeraSci, US
2.	Annamarie Vogt	Roche, EU
3.	Alan Kott	Bracket, EU
4.	Antonella Mollica	Syneos Health, EU
5.	Anna-Karin Berger	Lundbeck, EU
6.	Anzalee Khan	Nathan S. Kline Institute, US
7.	Bill Horan	VeraSci, US
8.	Brian Harel	Takeda
9.	Butlen Florence	EMA, EU
10.	Chris Edgar	Cogstate, Ltd., EU
11.	Chris Yavorsky	Valis Bioscience, US
12.	Daniel Umbricht	Independent consultant, EU
13.	David Daniel	Signant Health, US
14.	Daniela Tinoco	PPD, EU
15.	Gerry Dowson	p1vital, EU
16.	Laura Moreno	University of Cambridge, UK
17.	Inés del Cerro	Bellvitge, EU
18.	John Harrison	Metis Cognition, EU
19.	Josep Maria Haro Abad	Hospital Sant Joan de Déu, EU
20.	Judith Jaeger	Cognitionmetrics, US
21.	Kim Bishop	Globalpharmaconsultancy, US
22.	Kiri Granger	Cambridge Cognition, UK
23.	Larry Alphs	Newron, US
24.	Luke Allen	Cambridge Cognition, UK
25.	Jean-Pierre Lindenmayer	New York University Sch. of Medicine, US
26.	Jordi Alonso (*)	IMIM, EU
27.	Jan Kottner	Charité-Universitätsmedizin Berlin, EU
28.	Mark Opler	Medavante-Prophase, US
29.	Maria T. Acosta	NHGRI, US
30.	Michael Davidson	Nicosia University
31.	Stephen R. Marder	UCLA, US
32.	Monika Vance	Santium, CA
33.	Montserrat Ferrer Forès (*)	IMIM, EU
34.	Pavel Balabanov	EMA, EU
35.	Paul Maruff	Cogstate, Ltd., AU
36.	Peter Annas	Lundbeck, EU
37.	Silvia Zaragoza Domingo (*)	Neuropsynchro
38.	Simon Collinson	Savonix, APAC
39.	Richard Keefe	VeraSci, US
40.	Rohini Sen	Takeda, US
41.	Vanitha Krishna	IQVIA Biotech, US