Published in final edited form as:
Alzheimers Dement. 2023 February ; 19(2): 708–720. doi:10.1002/alz.12773.

Why a clinical trial is as good as its outcome measure: A framework for the selection and use of cognitive outcome measures for clinical trials of Alzheimer's disease

Roos J. Jutten[1], **Kathryn V. Papp**[1,2], **Suzanne Hendrix**[3], **Noel Ellison**[3], **Jessica B. Langbaum**[4], **Michael C. Donohue**[5], **Jason Hassenstab**[6], **Paul Maruff**[7,8], **Dorene M. Rentz**[1,2], **John Harrison**[9,10,11], **Jeffrey Cummings**[12], **Philip Scheltens**[11], **Sietske A. M. Sikkes**[11,13]

[1]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

[2]Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

[3]Pentara Corporation, Salt Lake City, Utah, USA

[4]Banner Alzheimer's Institute, Phoenix, Arizona, USA

[5]Alzheimer's Therapeutic Research Institute, Keck School of Medicine, University of Southern California, San Diego, California, USA

Correspondence Dr. Roos J. Jutten, Department of Neurology, Massachusetts General Hospital & Harvard, Medical School, 149 13th St, Boston, MA, 02129, USA. rjutten@mgh.harvard.edu.

This paper is an initiative that resulted from the Alzheimer's Association International Conference 2019 Featured Research Session (FRS) 'Composite endpoints for cognition: endpoints in biomarker-confirmed AD' (FRS panellists: Sikkes, Papp, Langbaum, Jutten, Hendrix, Ellison).

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

CONFLICTS OF INTEREST
R.J.J., P.M., and S.A.M.S. report no relevant conflicts of interests. K.V.P. has served on scientific advisory boards for Cogstate; she has consulted for Biogen, Eli Lilly, Novoic, and Digital Cognition Technologies. S.H. is owner and employee of Pentara Corporation. N.E. is a contractor for Pentara Corporation. Pentara consults for several dozen pharmaceutical, biotech, non-profit and academic groups in the Alzheimer's disease space. J.L. has served on a scientific advisory board for Biogen and has consulted for Alector. M.D. has served on scientific advisory boards for Biogen, Eli Lilly, and Neurotrack; and has consulted for Roche. His spouse is a full-time employee of Janssen. Jason Hassenstab has served as a paid consultant for Roche, Lundbeck, Eisai, and Parabon Nanolabs. John Harrison reports receipt of personal fees in the past 2 years from Actinogen, AlzeCure, Aptinyx, Astra Zeneca, Athira Therapeutics, Axon Neuroscience, Axovant, Bial Biotech, Biogen Idec, BlackThornRx, Boehringer Ingelheim, Brands2life, Cerecin, Cognito, Cognition Therapeutics, Compass Pathways, Corlieve, Curasen, EIP Pharma, Eisai, G4X Discovery, GfHEU, Heptares, Ki Elements, Lundbeck, Lysosome Therapeutics, MyCognition, Neurocentria, Neurocog, Neurodyn Inc, Neurotrack, the NHS, Novartis, Novo Nordisk, Nutricia, Probiodrug, Prothena, Recognify, Regeneron, reMYND, Rodin Therapeutics, Samumed, Sanofi, Signant, Syndesi Therapeutics, Takeda, Vivoryon Therapeutics and Winterlight Labs. Additionally, he holds stock options in Neurotrack Inc. and is a joint holder of patents with My Cognition Ltd. J.C. has provided consultation to AB Science, Acadia, Alkahest, AlphaCognition, ALZPathFinder, Annovis, AriBio, Artery, Avanir, Biogen, Biosplice, Cassava, Cerevel, Clinilabs, Cortexyme, Diadem, EIP Pharma, Eisai, GatehouseBio, GemVax, Genentech, Green Valley, Grifols, Janssen, Karuna, Lexeo, Lilly, Lundbeck, LSP, Merck, NervGen, Novo Nordisk, Oligomerix, Ono, Otsuka, PharmacotrophiX, PRODEO, Prothena, ReMYND, Renew, Resverlogix, Roche, Signant Health, Suven, Unlearn AI, Vaxxinity, VigilNeuro pharmaceutical, assessment, and investment companies. P.S. has received consultancy fees (paid to the institution) from AC Immune, Brainstorm Cell, EIP, ImmunoBrain Checkpoint, Genentech, Novartis, Novo Nordisk. He is PI of studies with AC Immune, FUJI-film/Toyama, UCB, and Vivoryon. He is a part-time employee of Life Sciences Partners Amsterdam. Author disclosures are available in the supporting information.

[6]Knight Alzheimer Disease Research Center, Department of Neurology, Washington University in St. Louis, St. Louis, Missouri, USA

[7]Cogstate Ltd., Melbourne, Victoria, Australia

[8]The Florey Institute of Neuroscience and Mental Health, Melbourne, Victoria, Australia

[9]Metis Cognition Ltd., Kilmington, UK

[10]Department of Psychiatry, Psychology & Neuroscience, King's College London, UK

[11]Alzheimer Center Amsterdam, Department of Neurology, Amsterdam UMC, location VUmc, VU University, Amsterdam, The Netherlands

[12]Chambers-Grundy Center for Transformative Neuroscience, Department of Brain Health, School of Integrated Health Sciences, University of Nevada Las Vegas (UNLV), Las Vegas, Nevada, USA

[13]Department of Clinical, Neuro and Developmental Psychology, Faculty of Movement and Behavioral Sciences, VU University, Amsterdam, The Netherlands

Abstract

A crucial aspect of any clinical trial is using the right outcome measure to assess treatment efficacy. Compared to the rapidly evolved understanding and measurement of pathophysiology in preclinical and early symptomatic stages of Alzheimer's disease (AD), relatively less progress has been made in the evolution of clinical outcome assessments (COAs) for those stages. The current paper aims to provide a benchmark for the design and evaluation of COAs for use in early AD trials. We discuss lessons learned on capturing cognitive changes in predementia stages of AD, including challenges when validating novel COAs for those early stages and necessary evidence for their implementation in clinical trials. Moving forward, we propose a multi-step framework to advance the use of more effective COAs to assess clinically meaningful changes in early AD, which will hopefully contribute to the much-needed consensus around more appropriate outcome measures to assess clinical efficacy of putative treatments.

Keywords

Alzheimer's disease; clinical outcome assessments; clinical trial; cognition; preclinical; prodromal

1 | INTRODUCTION

The success of a clinical trial depends on many factors, including the therapeutic intervention, the target population and the trial protocol.[1] A critical aspect of the protocol is that the outcome measure(s) to evaluate treatment efficacy are appropriate. That is, a treatment can be shown to be effective only if the outcome measures can capture the treatment effect.[2] *Clinical* efficacy is usually evaluated using a report by a clinician, the participant, or a non-clinician observer (i.e., study partner), which all fall under the umbrella term "clinical outcome assessments" (COAs). Common guidance for COAs holds that the selected instrument(s) should be reliable, valid, and sensitive to clinically meaningful change in the target population[3]. In the context of Alzheimer's disease (AD), however, these guiding principles raise unique challenges and are not always followed carefully.

In AD clinical trials, a slowing or halting of cognitive decline is usually the main clinical goal, and the COAs for assessing this typically include a performance outcome assessment (PerfO) of cognition accompanied by a clinician-reported outcome (ClinRO), a report by the individual themselves (i.e., a patient-reported outcome, PRO), or observation by a study partner or clinician (ObsRO) to demonstrate the impact on everyday functioning.[4] Historically, the clinical instrumentation for AD trials was defined in the early 1990s by trials of tacrine, the first agent approved for the treatment of AD.[5,6] Patients were selected using the Mini–Mental State Examination (MMSE) developed in 1975,[7] and COAs to establish clinical efficacy included the Clinical Global Impression of Change (CGIC) published in 1975[8] and the Alzheimer's Disease Assessment Scale (ADAS) designed in 1984.[9] The approval process was based on draft guidelines from the U.S. Food and Drug Administration (FDA) provided in 1990,[10] which required that anti-dementia agents show improvement on the core symptoms of AD (i.e., memory and global cognition) and that the effect be clinically meaningful as shown on a global or a functional rating. This approach to mild-to-moderate AD clinical trials and approval remains highly influential, as trial participants are still typically defined by MMSE score range and evaluated by the ADAS – cognitive subscale (ADAS-Cog) and a global assessment such as the CGIC.

In the past decade, AD clinical trials have increasingly focused on earlier, predementia stages of AD,[11,12] as intervening with AD pathology before the onset of dementia may be more efficacious for certain interventions, and slowing the disease while the individual is still highly functional is an important goal[1]. This shift toward earlier disease stages has led to the need for outcome measures that can capture the more subtle cognitive changes that occur prior to individuals being classified clinically with dementia.[13,14] The US Food and Drug Administration (FDA) updated their draft guidance for early-stage AD trials in 2018, including recommendations for different COAs by clinical stage of AD (Table 1).[4] This guidance stated that many of the assessment tools typically used in patients with overt dementia may not be suitable for use in earlier symptomatic or preclinical stages of AD, which is consistent with studies indicating that these measures exhibit ceiling effects in individuals with mild cognitive impairment (MCI) and even mild AD dementia.[15–17] Despite these updates, however, most of the recent large Phase 3 AD clinical trials for prodromal to mild AD (e.g., in references[18–22]) have incorporated the ADAS-Cog and the MMSE as primary or secondary endpoints. Applying these later disease-stage targeted measures in these early trials, however, could have serious consequences for trial outcomes, such as false positive futility analyses (i.e., the incorrect conclusion of futility) or underestimated treatment effects, and may thereby, in part, explain both recently failed trials and tentative successes in prodromal and mild AD.[19,22,23]

Several efforts have been undertaken in an attempt to improve the measurement of cognitive changes in early symptomatic stages of AD (i.e., FDA Stage 3, Table 1), for example by selecting the parts that have been shown to be sensitive to early AD change and combine them in novel composite tools, for example, the Alzheimer's disease Composite Score (ADCOMS),[24] the Integrated Alzheimer's disease Rating Scale (iADRS),[25] the Neuropsychological Test Battery (NTB) and the Cognitive-Functional Composite (CFC).[26,27] However, these endeavors have yet to be broadly accepted as primary outcomes in clinical trials for symptomatic populations. For preclinical stages of

AD (i.e., FDA Stage 1–2, Table 1), novel, multi-domain cognitive composites combining sensitive neuropsychological tests have been designed, such as the Preclinical Alzheimer's Cognitive Composite (PACC)[28] and the Alzheimer's Prevention Initiative Preclinical Cognitive Composite (APCC) for preclinical AD,[29] but it remains unclear as to whether these composites will have adequate sensitivity to detect a clinically meaningful treatment effect in (secondary) prevention trials.[30]

Given the lack of a single well-accepted and validated 'gold standard' COA to assess meaningful cognitive changes in early stages of AD, a more critical use of COAs in trials of AD will be crucial given the recent developments in pharmacotherapy and increased focus towards earlier disease stages.[1,31] To facilitate this, we aimed to provide a benchmark for the design and evaluation of appropriate COAs for use in early, predementia stages of AD with a focus on PerfOs of cognition. We will start with summarizing the most important lessons learned from historical data on capturing cognitive changes in preclinical and prodromal stages of AD. Next, we will discuss theoretical and psychometric challenges when developing and validating novel COAs to evaluate cognitive changes in these early stages, followed by a consideration of what validation evidence is needed to advance the implementation of novel measures in clinical trials.

2 | CAPTURING COGNITIVE CHANGES IN AD: WHAT WE (COULD) HAVE LEARNED

2.1 | Commonly used cognitive tests are differentially sensitive to change

A necessary condition for capturing cognitive changes is the sensitivity to change of the measurement instrument, also referred to as responsiveness.[34] This may include responsiveness to decline (in the placebo group) as well as responsiveness to improvement (in the treatment group), although the first seems most relevant in the context of current disease modifying treatments that aim to demonstrate slowing or halting of cognitive decline in the treated group. Commonly used COAs selected for AD clinical trials tend to demonstrate acceptable levels of reliability as well as face validity (see Section 3), but these characteristics may not necessarily represent sufficient conditions for detecting change over time or treatment benefit. Thus, investigating the sensitivity to change as a separate measurement characteristic is crucial for assessing cognitive change.

A substantial body of information on the sensitivity to decline of widely used cognitive tests is readily available from observational studies (e.g., in references[13,32–36]), which, for example, have consistently pointed out specific tests that seem to be sensitive to decline in the preclinical to prodromal stage of AD, such as the Free and Cued Selective Reminding test (FCSRT), Category Fluency Test (CFT), Digit Symbol Substitution Test (DSST), and Controlled Oral Word Association Test (COWAT).[32,33,35] Aligned with these findings, recent clinical trials have provided promising results regarding the sensitivity to treatment effects of those measures, as exemplified by the use of the CFT, COWAT, and the DSST in some Phase 2 MCI and mild AD dementia trials.[37–39] Employing those tests in confirmatory Phase 3 trials could represent the much-needed sensitive measures of cognition,[4] especially regarding cognitive domains such as executive functioning that are not well (or not at all) measured by traditional instruments such as the ADAS-Cog. To date, there have been no systematic reviews addressing the question of which specific cognitive tests are successful in detecting changes. This is regrettable, as it would be helpful to know how these measures have performed with respect to treatment mode of action, and disease severity, biomarker and susceptibility gene status, and so on. However, a risk in the reporting of positive effects with specific instruments is that failed or discontinued trials may have used the same or similar measures but were not reported.[40]

2.2 | The magnitude of cognitive change in clinical trials differs from observational studies

Data from observational studies can help set expectations for how much cognitive decline is expected to occur over a specific time course. In clinical trials of symptomatic AD, however, individuals on placebo might decline *less* than expected even though they were similar to observational cohorts at baseline.[41] There are several potential reasons for this, including placebo effects, which may influence both cognitive test performance and self- and proxy-report outcomes in AD trials and thus seem to occur regardless of the type of COA used.[42,43] Another potential explanation is the likelihood of enrolling "healthier" individuals (i.e., due to the in- and exclusion criteria prohibiting commonly occurring medical conditions or medication that would not be exclusionary for an observational study[44]) and monitoring individuals more closely. Whether these phenomena occur in preclinical AD trials is not yet fully known. Only one trial in mostly preclinical AD participants has been completed (Dominantly Inherited Alzheimer Network – Treatment Unit [DIAN-TU-001][45]) and several more are ongoing.[46] Early reports from DIAN-TU-001 indicate that the trial data did not meet the assumptions based on the observational data used to model the trial. In fact, the placebo group did not show as much decline as expected.[47]

Another striking difference between observational studies and clinical trials regards the frequency of study assessments. Many observational studies administer cognitive assessments no more than once a year, whereas in clinical trials, most participants are tested every 3–6 months.[48] This may result in larger, unanticipated practice effects (PE), also referred to as learning or retest effects. PE refer to the phenomenon that performance on cognitive tests improves with retesting as participants become familiar with test content, test strategies, and the testing environment. The occurrence of PE is a vexing issue that makes the detection of cognitive changes more difficult, and if not accounted for, can mask cognitive decline[49,50] and decrease statistical power to detect treatment effects.[51,52] An observation of particular relevance for clinical trials in early AD is that PE appear to be most salient on measures of episodic memory.[49,53,54] Although PE seem to be most apparent at the second time of testing, it has been shown that retest gains are stronger with more frequent exposure,[55,56] meaning that the frequency of assessments, in addition to alternate test versions, in clinical trials should be seriously considered. Finally, placebo effects may interact with and possibly magnify PE in intervention studies.[57]

2.3 | Individual trajectories of cognitive decline are heterogeneous

Another issue observed in longitudinal data is that individuals with AD show substantial heterogeneity in terms of disease progression rates. That is, even when individuals are matched on disease severity levels at baseline, those individuals may have variable times to dementia[58] as well as in rates of cognitive decline.[59] This inter-individual variation in cognitive change is an obstacle for measuring therapeutic effects using cognitive outcome measures. A recent study has shown that natural variability in disease progression may lead to significant group differences on clinical outcomes that are not due to the therapy administered but, instead, reflect individual differences in their rate of decline.[60] This observed inter-individual heterogeneity may be inflated by the small signal-to-noise ratio observed on the conventional COAs used, which may be influenced by contextual factors (i.e., within-subject sources of variability) such as day-to-day variability in mood, effort, and concentration, educational influences, and/or inter-rater variability. This further emphasizes the need for refined COAs that tap into all relevant cognitive processes in early stages of AD while reducing error as much as possible.[60] To accomplish this, the individual cognitive tests included in the COAs should be carefully selected.

3 | EVALUATING COGNITIVE TESTS FOR COA SELECTION: CHALLENGES AND CONSIDERATIONS

3.1 | Validity aspects

Perhaps the most challenging aspect of evaluating a cognitive test is to establish its validity; the degree to which the test indexes the construct it purports to measure.[61] In 2005, the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) initiative was founded with the overall aim to improve the selection of health-related outcome measurement instruments. As part of this initiative, a framework including the taxonomy, terminology, and definitions of measurement properties was developed by an international multidisciplinary team of researchers.[62] The COSMIN methodology was initially designed for PROs, but most principles also apply to other types of outcome measurement instruments including ClinROs and PerfOs.[63] COSMIN provides tools to help researchers and clinicians assess the methodological quality of studies reporting on measurement instruments[64] as well as a database that includes over 2000 systematic reviews on measurement properties of health-related measurement instruments for various constructs (e.g., quality of life, pain, functional status, see: https://database.cosmin.nl). Remarkably, a systematic review of COAs to assess treatment efficacy of cognitive enhancing drugs in AD showed that only 50% of the measures used as primary outcome measures had information about their validity,[65] which underlines the challenges that the concept of validity is still facing in our field. To improve the validity of COAs for use in AD research and clinical trials, we apply the COSMIN methodology for validity aspects that are most relevant to performance-based cognitive tests. We start with basic prerequisites for validity, including content validity, followed by a discussion of the interpretability of test scores, often referred to in our field as clinical meaningfulness.

3.1.1 | Content validity—The first and most important component of any measure is to consider whether the content of an instrument is an adequate reflection of the construct to be measured, referred to as “content validity.”[61] This entails that (1) all the items of a measurement instrument are relevant for measuring the construct of interest (relevance); (2) all key concepts are included (comprehensiveness); and (3) all items, response options, and instructions are understood by the target population as intended (comprehensibility).[62] In the context of AD trials, this would mean that a COA covers all cognitive domains one would theoretically expect to decline because of the disease process or improve as a result of a treatment. It is important to consider this element even before statistical considerations, as an agnostic statistical approach might result in including tests that are not directly relevant to the disease process but are simply more sensitive to change in general by exhibiting wider measurement ranges (e.g., reaction time measures). One important component of content validity is face validity, which refers to the 'clinical impression', or the degree to which an instrument looks as though it is an adequate reflection of the construct to be measured.[62] This is a subjective experts' impression, and there are no clear standards with regard to how it should be assessed and quantified.[66] As a result, "face validity" alone is insufficient to confirm an instrument's content validity, although it might be crucial to the implementation of the instrument. Therefore, it is recommended to test other content aspects (i.e., relevance, comprehensiveness, comprehensibility), by cognitive debriefing and qualitative methods.[67]

3.1.2 | Criterion validity—Criterion validity refers to the degree to which the scores of a measure are an adequate reflection of a 'gold' standard. In the context of cognitive assessment in AD, this gold standard would be "cognitive decline due to AD" when the purpose is measuring disease progression. However, measuring AD-specific cognitive decline and distinguishing it from other sources of cognitive changes is challenging, since the cognitive trajectories of incipient symptomatic AD are subtle and heterogeneous.[60] One way cognitive tests may be deemed to have good criterion validity is by their ability to predict cognitive deficits later in the course of the disease, referred to as predictive validity.[4,54] For example, across three cohorts of clinically unimpaired older adults with elevated amyloid, those with subtle cognitive decline over 3 years on the PACC were 5.47 times more likely to receive a diagnosis of MCI,[68] and although this is not yet a gold standard for cognition, this could serve as a "copper standard," or an indirect reflection of the criterion validity. A limitation of this approach, however, is that it may require extended follow-up of clinical trial participants.

It could be argued that autopsy-based verification of AD pathology and in vivo imaging, or in the future even fluid AD biomarkers, can be considered the gold standard for AD pathology and can as such be used as a criterion for the validity of cognitive tests. However, we believe it is important to disentangle the disease process as reflected with biomarkers from the clinical symptoms as reflected with cognitive tests, because AD pathology does not always fully correlate to performance on cognitive tests (e.g., in extreme aging[69]). Moreover, in theory a gold standard consists of a perfectly reliable and valid measure across populations, which is not yet the case for all AD biomarkers. In addition, it would be difficult to hypothesize what the strength of the relationship between the instrument scores and criterion scores should be given the heterogeneity in clinical-pathological relationships,[60] which may, in part, be due to factors such as resilience and cognitive reserve.[70] We therefore remain cautious with implying that a gold standard for AD pathology is also a gold standard for "cognition" and suggest that validating clinical measures against biomarkers is rather a way to establish construct and convergent validity as will be discussed in the next section.

3.1.3 | Construct validity—

The main validity component to which we must defer in the absence of a gold standard, is construct validity. This refers to whether the test score is an adequate reflection of the underlying (latent) construct and its expected relationship with related and unrelated constructs.[62] One aspect of construct validity is determining whether the scoring algorithm (resulting in a single total score, weighted or unweighted composite score) is an adequate reflection of the dimensionality (i.e., one or multiple cognitive domains) being measured. To support this, a factor analysis should be performed to demonstrate adequate model fit and score use (single score or domain scores) (also referred to as "structural validity" in the COSMIN methodology). Subsequently, the test score (or domain scores) should be investigated against relevant clinical (e.g., everyday functioning, quality of life) and/or biological measures (e.g., biomarkers of neurodegeneration, amyloid, or phosphorylated tau). A distinction can be made between convergent and divergent validity, both of which could support construct validity. For example, in the context of cognitive tests for use in AD, the test score would ideally be more strongly related to AD biomarker burden than to age, sex, level of education, or mood. Both convergent and divergent relationships should be in line with hypotheses, which can be formulated as range of correlations rather than based on statistical significance.[61]

3.1.4 | Cross-cultural validity—

Cross-cultural validity refers to the degree to which the performance of the items on a translated or culturally adapted measurement instrument are an adequate reflection of the performance of the items of the original version of the measurement.[62] Cognitive outcome measures are used in a variety of countries and cultures and an often-overlooked issue is how these measures are inherently culturally dependent. A broad literature on cross-cultural neuropsychology has shown that many features of assessment cannot be assumed to be comparable across cultures, including assumptions that tests measure similar aspects of brain function in all people[71] and that test concepts apply equally to all levels of education and quality of education.[72,73]

In general, there is a fallacy that cognitive tests can be "culture-free" by emphasizing non-verbal content or by localizing test content to match familiar items. These approaches have been shown to fail repeatedly when comprehensively studied.[74,75] In addition, there are classes of measures that cannot be adapted across cultures to provide directly comparable data due to cultural differences and reliance on specific aspects of language. Many cultures do not use primarily phonemic alphabets and their written language is logographic in nature (e.g., Japanese, Chinese, Korean).[72] Another extremely complicated issue for global studies is establishing the validity necessary to assume metric equivalence across cultures. Great care must therefore be taken in selecting cognitive outcomes that have undergone cross-cultural validation procedures, and this is particularly relevant for global AD trials and observational studies. To optimize comprehensibility, a focus on cross-cultural adaptations should be given just as much or more priority than a quality translation of test instructions and item content.[76] Best practices for selecting cognitive measures appropriate for global AD studies should include considerations regarding adaptation and translation capabilities, selecting measures that assess meaningful constructs in each culture where data will be collected, and consideration of local normative data.[77]

3.1.5 | Interpretability—

Finally, an important aspect of special relevance to cognitive test scores is interpretability, which revolves around the question of whether a score can be interpreted as being “meaningful” and translatable to everyday life performance.[62] Cognitive tests which require individuals to draw lines between numbers and letters, learn lists of unrelated words, and draw figures may not immediately appear analogous to the cognitive skills required to function in everyday life. One example of demonstrating the interpretability of a test score could be whether test performance is associated with behaviors in real-world settings, referred to as ecological validity.[78] In symptomatic stages of AD, there is a clear association between deficits in performance on cognitive testing and deficits in everyday functioning.[79–81] However, the direct link between poorer cognitive test performance and decrements in everyday functioning is less immediately apparent in the preclinical and prodromal stages of AD, where, by definition, functional symptoms are minimal, cognitive decrements exceedingly subtle, and cognitive decline is thought to precede functional decline.[82,83] One means to enhance ecological validity in the context of subtle cognitive change is to include elements that are associated with behavior in everyday life. For example, a memory measure that requires learning and recall of a grocery list may confer greater ecological validity in contrast with a list of unrelated words, as the former is a cognitive exercise required in everyday life.[84] Additional approaches to enhancing ecological validity may include performance-based assessments of function (e.g., ability to navigate through a website or telephone decision tree[85]) or PROs measuring subtle everyday cognitive problems.[13,86,87] Integrating cognitive testing more closely into everyday life (e.g., burst-based cognitive assessments on an individual’s own smartphone) may be a means of increasing the ecological validity of a cognitive assessment.[88]

Of importance to clinical trials is not only the interpretability of a single test score, but the question of when a *change* in score is clinically meaningful. Treatment effects that have clinical relevance to the patient and caregiver are more acceptable to regulators and payors.[4] AD secondary prevention trials targeting Stage 1 and 2 individuals have been placed in a conundrum whereby a cognitive-only outcome (in contrast with the usually used co-primary outcome of cognition and function in trials of Stage 3 participants) is likely the most sensitive measure to determine a treatment effect, but the clinical meaningfulness of a measured cognitive change is not well-defined and there is no consensus, as of yet, on what represents a clinically meaningful treatment effect.[30,89] Mixed methods approaches, combining the perception of change scores by patients, caregivers, and healthcare professionals with quantitative characteristics of tests (such as the smallest detectable change), might be necessary to determine the clinical meaningfulness of cognitive test scores. In epidemiological studies, several approaches to determine the “minimal important change” have been evaluated for patient reported outcome measures.[90,91] The field of AD and related dementias can learn from these precedents.

3.2 | Composite outcome measures

Since cognitive and functional composite measures have garnered interest as primary endpoints for AD trials, we will now discuss challenges and considerations when evaluating composite outcome measures for AD trials. Composite measures summarize multiple individual tests into a single score but can be constructed using multiple approaches (e.g., data-driven vs. theoretical, single vs. multi-domain, explicitly weighted vs. implicitly weighted, etc.). This infers that the “optimal” composite measure can differ by disease stage, since the nature as well as the rate of cognitive and functional decline differs across the continuum from preclinical to symptomatic AD.[4,92–94] When adapted to the disease stage of interest, composites may improve reliability and sensitivity by refining the signal-to-noise ratio. That is, whereas a single test might under- or overestimate decline given natural within-person variability in performance, multi-domain composites are more capable of capturing “robust” decline.[95] A recent study demonstrated this phenomenon in unimpaired late middle-age adults where a series of theoretically derived composites exhibited lower within-individual variability and stronger age and amyloid associations compared with scores from single tests.[96] Likewise, composites may improve sensitivity to change over time, especially when the individual measures are selected based on showing the steepest decline in a population of interest.[97] That is, composites may help mitigate the psychometric shortcomings of individual measures in the composite (e.g., by reducing in less floor or ceiling effects). For example, the MMSE may be included in a composite because of its face validity, but the limitations of its range restrictions in scoring, especially ceiling effects in cognitively normal populations, could be attenuated by inclusion among other measures that are more normally distributed in that population. Moreover, multi-domain composites may improve detection of decline in a heterogeneous disease such as AD. For example, some individuals may exhibit relatively greater decline in executive functions versus memory or vice versa,[98] implying that an individual test or composite focused on a single test or domain may underestimate cognitive change in participants who don’t have the most typical presentation of symptoms.[99] However, it should be noted that a composite score remains a summary measure determined by its parts and entails the potential risk of having detection of subtle cognitive changes on some measures compromised by other unchanged measures included in the composite score. This further emphasizes that the selection of individual tests for a composite measure will influence the composite’s ability to capture disease progression. The validity considerations and recommendations in Table 2 can be applied to advance the optimal selection of tests included in a composite and thereby its overall quality as a COA for use in AD trials.

3.2.1 | Cognitive composite measures for AD Stage 1 and 2 trials—The need for more sensitive measures of neuropsychological performance in Stage 1 and 2 studies[4] (Table 1), has encouraged further development of cognitive composites.[97] As discussed in the validity section, it is important to base the selection of measures on sound theoretical foundations. However, the disadvantage of a purely theoretical approach is that the items or tests selected might lack the ability to capture changes due to limited measurement properties. To overcome this, items for composites are generally selected based on good sensitivity to progression over time. Several large longitudinal datasets are available for deriving composite outcome measures over time. Empirical derivation of cognitive outcome measures based on historical data allows selection of items and weighting of items based on one of several approaches: (1) optimize the sensitivity of items to track progression over time, (2) optimize the cross-sectional discrimination between populations such as healthy normal and MCI or AD dementia participants, (3) optimize the separation between healthy normal and MCI or AD dementia participants in progression over time, or (4) optimize the separation between active treatment and placebo treatment participants in progression over time. Each of these approaches reflects a different goal for use of the composite, but often, they can result in somewhat similar optimal composite scores.[100] As highlighted in Section 2, inter-individual variation in disease progression makes it difficult to create optimized composite measures that are externally valid and perform well prospectively.[101] Thus, optimization of composite scores based on historical datasets may result in overfitting and the optimal combination for progression over time might not generalize to other cohorts. To control for these biases, cross-validation methods should be used to estimate the bias of the specific method of derivation that is used, so that the performance can be estimated after correcting for this bias.[28] Furthermore, future public-private partnered trials will hopefully allow for broader data sharing to do these types of analyses on actual clinical trial data rather than observational studies.

3.2.2 | Composites combining cognition and function for AD Stage 3 trials

—For trials targeting Stage 3 individuals, combining cognitive with functional outcome measures may provide an effective approach to establish the meaningfulness of treatments, and this has been recommended by regulatory agencies.[4,89] Accumulating evidence indicates that instrumental activities of daily living (IADL) begin to deteriorate before the onset of dementia,[102] and that an MCI diagnosis accompanied by more IADL problems leads to increased risk of developing dementia.[103] Previous studies have shown the added clinical value of combining cognitive with IADL measures, as illustrated by the fact that associations between the overall composite combining cognitive and functional abilities and other clinical measures such as quality of life and caregiver burden were mostly driven by the functional component of the composite in individuals with MCI and mild AD dementia[17]; likewise the detection of change over time in individuals with MCI on the composite was most apparent on the instrumental ADL (IADL) component.[34] Together, these and other findings emphasize the importance of the inclusion of a sensitive IADL measure to capture clinically meaningful decline in prodromal stages of AD.[102] Thus, outcome measures that combine cognitive and functional items may be useful for measuring the totality of disease progression in Stage 3 participants. A limitation of combining both constructs in one composite is that change on only one of both aspects can be masked in the composite score. Therefore, it is recommended to analyze the cognitive and functional scores separately, for example in a secondary or sensitivity analysis.

4 | THE IMPLEMENTATION OF NOVEL MEASURES IN CLINICAL TRIALS

The pharmaceutical industry, particularly in neurology, has exhibited substantial conservatism in adopting new instrumentation, especially with respect to cognitive and functional outcome measures.[104] This has led to situations in which tests were chosen based on historical precedents, rather than on a hypothesized link with the mode of action or target of the specific compound. To improve trial methodology, we believe that progress must be bidirectional. On the one hand, test developers are encouraged to design and validate tests more carefully. This requires a stepwise process in order to demonstrate that a test has acceptable levels of (1) reliability, such as test-retest reliability and internal consistency; (2) various forms of validity that together lead to converging evidence for the validity of the proposed measure (Table 2); and (3) sensitivity to change in the target population. On the other hand, end-users should more critically (re)consider their selection of tests for COAs for their trials. That is, based on the hypothesized mode of action and presumed cognitive change in the target population, the specific tests/items should be selected based on the constellation of evidence for their validity and measurement quality (Figure 1).

Several promising endeavors have been undertaken to provide evidence for the validity of novel COAs for use in AD trials. For example, the PACC, theoretically designed and originally investigated in a multicohort study[28], was subsequently cross validated[101] and studied in a replication cohort confirming its ability to capture cognitive decline during the preclinical stages of AD.[105] Later, an optimized version of the PACC (i.e., the PACC5) was designed to improve the sensitivity to amyloid-related cognitive decline in preclinical stages of AD.[106] Examples of validation work on composites that were empirically derived include the ADCOMS and APCC, which were both externally validated in multiple datasets.[24,26,27] Another example of a step-wise validation process includes the Catch-Cog study on the CFC, which was first investigated for its content validity and reliability,[107] followed by a longitudinal construct validation study in an independent cohort investigating the CFC against other clinical and biological measures of disease severity[17,108] and sensitivity to change over time compared to traditional measures such as the ADAS-Cog-13 and CDR-SB.[34] Other examples of recently developed COAs for early AD trials are provided in a recent overview,[97] pointing out the differences regarding the body of validation evidence available for different measures and highlighting that validity is an ongoing, multi-step process rather than an end-product.

5 | CONCLUSION

Compared to the rapidly evolved understanding and consensus of measuring pathophysiology in early stages of AD, relatively less progress has been made in the evolution of COAs for use in the early stages of the disease. We believe that this is a missed opportunity impeding advances in evaluation of new therapies, particularly for trials in predementia stages of AD. We discussed several barriers to the development and selection of more effective cognitive outcome measures in early AD trials. These include unique challenges related to assessing cognition in predementia stages such as the detection of very subtle and heterogeneous decline across individuals, evaluating and demonstrating the validity of measures of cognition, and reservations from pharmaceutical sponsors towards adopting new COAs. To overcome this, we propose a multi-step framework to advance the selection and implementation of more effective COAs in clinical trials of early AD (Table 2, Figure 1). This framework can be applied when selecting existing tests or items for a composite measure to include as a COA as well as when developing entirely novel measurement instruments for cognition to implement as a COA in future trials of AD.

We recognize that addressing all validity aspects mentioned in Table 2 will be an effortful endeavor, but we believe that setting this high standard to collect converging evidence for the validity of a cognitive test will be crucial for the long overdue and much-needed improvement of outcome measures for cognition in our field. However, different validity aspects may arguably deserve different weights depending on the type of the

COA and intended target population for which the COA is being evaluated and selected. According to the COSMIN framework, the content validity and internal structure (including structural validity and cross-cultural validity) are considered most important, followed by the remaining measurement properties, such as hypothesis testing for construct validity. The latter might differ per disease stage, for example, information about associations with a certain biomarker will be more relevant for presymptomatic disease stages, whereas associations with quality of life measures might be more relevant for clinically advanced disease stages. Furthermore, recommendations to establish the various aspects of validity may evolve in response to developments in the field, for example, when novel types of measurement instruments (e.g., using passive technology or virtual reality) are being developed or novel prognostic (bio)markers for AD are discovered. Together, this will hopefully contribute to the much-needed consensus and use of more appropriate outcome measures to assess clinical efficacy, and thereby increase the chance of successful clinical trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank dr. L.B. Mokkink for her assistance in interpreting and applying the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) framework.

REFERENCES

1. Cummings J, Feldman HH, Scheltens P. The “rights” of precision drug development for Alzheimer’s disease. Alzheimers Res Ther. 2019;11(1):76. [PubMed: 31470905]
2. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun. 2018;11:156–164. [PubMed: 30112460]
3. Prinsen CAC, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a “Core Outcome Set” – a practical guideline. Trials. 2016;17(1):449. [PubMed: 27618914]
4. FDA, Early Alzheimer’s Disease: Developing Drugs for Treatment. Guidance for Industry. 2018.
5. Davis KL, Thal LJ, Gamzu ER, et al. A double-blind, placebo-controlled multicenter study of tacrine for Alzheimer’s disease. The Tacrine Collaborative Study Group. N Engl J Med. 1992;327(18):1253–1259. [PubMed: 1406817]
6. Farlow M. A controlled trial of tacrine in Alzheimer’s disease. The Tacrine Study Group. JAMA. 1992;268(18):2523–2529. [PubMed: 1404819]
7. Folstein MF, Folstein SE, Mchugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189–198. [PubMed: 1202204]
8. Guy W. Clinical Global Impressions. ECDEU Assessment Manual for Psychopharmacology—Revised. 1976: Rockville, MD: U.S. Department of Health, Education, and Welfare; Public Health Service, Alcohol; Drug Abuse, and Mental Health Administration; National Institute of Mental Health; Psychopharmacology Research Branch; Division of Extramural Research Programs. p. 218–222.
9. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer’s disease. Am J Psychiatry. 1984;141(11):1356–1364. [PubMed: 6496779]

10. Leber P. Guidelines for the Clinical Evaluation of Antidementia Drugs. First draft. Rockville, MD: U.S. Food and Drug Administration; 1990.

11. Sperling RA, Aisen PS, Beckett LA, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7(3):280–292. [PubMed: 21514248]

12. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011:7(3):270–279. [PubMed: 21514249]

13. Weintraub S, Carrillo MC, Farias ST, et al. Measuring cognition and function in the preclinical stage of Alzheimer's disease. Alzheimers Dement (N Y). 2018;4:64–75. [PubMed: 29955653]

14. Snyder PJ, Kahle-Wrobleski K, Brannan S, et al. Assessing cognition and function in Alzheimer's disease clinical trials: do we have the right tools? Alzheimers Dement. 2014;10(6):853–860. [PubMed: 25458309]

15. Cano SJ, Posner HB, Moline ML, et al. The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. J Neurol Neurosurg Psychiatry. 2010;81(12):1363–1368. [PubMed: 20881017]

16. Karin A, Hannesdottir K, Jaeger J, et al. Psychometric evaluation of ADAS-Cog and NTB for measuring drug response. Acta Neurol Scand. 2014;129(2):114–122. [PubMed: 23763450]

17. Jutten RJ, Harrison JE, Lee Meeuw Kjoe PR, et al. Assessing cognition and daily function in early dementia using the cognitive-functional composite: findings from the Catch-Cog study cohort. Alzheimers Res Ther. 2019;11(1):45. [PubMed: 31092277]

18. Honig LS. Trial of solanezumab for mild dementia due to Alzheimer's disease. N Engl J Med. 2018;378(4):321–330. [PubMed: 29365294]

19. Wessels AM, Tariot PN, Zimmer JA, et al. Efficacy and safety of Lanabecestat for treatment of early and mild Alzheimer disease: the AMARANTH and DAYBREAK-ALZ randomized clinical trials. JAMA Neurol. 2020;77(2):199–209. [PubMed: 31764959]

20. von Hehn C, Rosenstiel PV, Tian Y, et al. Baseline Characteristics from ENGAGE and EMERGE: Two Phase 3 Studies to Evaluate Aducanumab in Patients with Early Alzheimer's Disease (P4. 1–001). AAN Enterprises; 2019.

21. Vandenberghe R, Rinne JO, Boada M, et al. Bapineuzumab for mild to moderate Alzheimer's disease in two global, randomized, phase 3 trials. Alzheimers Res Ther. 2016;8(1):18. [PubMed: 27176461]

22. Egan MF, Kost J, Voss T, et al. Randomized trial of verubecestat for prodromal Alzheimer's disease. N Engl J Med. 2019;380(15):1408–1420. [PubMed: 30970186]

23. Knopman DS, Jones DT, Greicius MD. Failure to demonstrate efficacy of aducanumab: An analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. Alzheimers Dement. 2021;17(4):696–701. [PubMed: 33135381]

24. Wang J, Logovinsky V, Hendrix SB, et al. ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. J Neurol Neurosurg Psychiatry. 2016;87(9):993–999. [PubMed: 27010616]

25. Wessels AM, Siemers ER, Yu P, et al. A combined measure of cognition and function for clinical trials: the integrated Alzheimer's Disease Rating Scale (iADRS). J Prev Alzheimer Dis. 2015;2(4): 227–241.

26. Ayutyanont N, Langbaum JBS, Hendrix SB, et al. The Alzheimer's prevention initiative composite cognitive test score: sample size estimates for the evaluation of preclinical Alzheimer's disease treatments in presenilin 1 E280A mutation carriers. J Clin Psychiatry. 2014;75(6):652–660. [PubMed: 24816373]

27. Langbaum JB, Hendrix SB, Ayutyanont N, et al. An empirically derived composite cognitive test score with improved power to track and evaluate treatments for preclinical Alzheimer's disease. Alzheimers Dement. 2014;10(6):666–674. [PubMed: 24751827]

28. Donohue MC, Sperling RA, Salmon DP, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol. 2014;71(8):961–970. [PubMed: 24886908]

29. Langbaum JB, Ellison NN, Caputo A, et al. The Alzheimer's Prevention Initiative Composite Cognitive Test: a practical measure for tracking cognitive decline in preclinical Alzheimer's disease. Alzheimers Res Ther. 2020;12:1–11.

30. Rentz DM, Wessels AM, Annapragada AV, et al. . Building clinically relevant outcomes across the Alzheimer's disease spectrum. Alzheimers Dement (N Y). 2021;7(1):e12181.

31. Cummings J, Lee G, Ritter A, et al. Alzheimer's disease drug development pipeline: 2019. Alzheimers Dement (N Y). 2019;5:272–293. [PubMed: 31334330]

32. Mura T, Proust-Lima C, Jacqmin-Gadda H, et al. Measuring cognitive change in subjects with prodromal Alzheimer's disease. J Neurol Neurosurg Psychiatry. 2014;85(4):363–370. [PubMed: 23840054]

33. Mortamais M, Ash JA, Harrison J, et al. Detecting cognitive changes in preclinical Alzheimer's disease: a review of its feasibility. Alzheimers Dement. 2017;13(4):468–492. [PubMed: 27702618]

34. Jutten RJ, Harrison JE, Brunner AJ, et al. The Cognitive-Functional Composite is sensitive to clinical progression in early dementia: longitudinal findings from the Catch-Cog study cohort. Alzheimers Dement (N Y). 2020;6(1):e12020.

35. Harrison JE, Rentz DM, Brashear HR, Arrighi HM, Ropacki MT, Liu E. Psychometric evaluation of the neuropsychological test battery in individuals with normal cognition, mild cognitive impairment, or mild to moderate Alzheimer's Disease: results from a longitudinal study. J Prev Alzheimers Dis. 2018;5(4):236–244. [PubMed: 30298182]

36. Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): modifications and responsiveness in pre-dementia populations. a narrative review. J Alzheimers Dis. 2018;63(2):423–444. [PubMed: 29660938]

37. Maher-Edwards G, De'ath J, Barnett C, Lavrov A, Lockhart A. A 24-week study to evaluate the effect of rilapladib on cognition and cerebrospinal fluid biomarkers of Alzheimer's disease. Alzheimers Dement (N Y). 2015;1(2):131–140. [PubMed: 29854933]

38. Gilman S, Koller M, Black RS, et al. Clinical effects of Abeta immunization (AN1792) in patients with AD in an interrupted trial. Neurology. 2005;64(9):1553–1562. [PubMed: 15883316]

39. Winblad B, Gauthier S, Scinto L, et al. Safety and efficacy of galantamine in subjects with mild cognitive impairment. Neurology. 2008;70(22):2024–2035. [PubMed: 18322263]

40. Yilmaz T, Jutten RJ, Santos CY, Hernandez KA, Snyder PJ. Discontinuation and nonpublication of interventional clinical trials conducted in patients with mild cognitive impairment and Alzheimer's disease. Alzheimers Dement (N Y). 2018;4:161–164. [PubMed: 29955660]

41. Berres M, Monsch AU, Spiegel R. Using historical data to facilitate clinical prevention trials in Alzheimer disease? An analysis of longitudinal MCI (mild cognitive impairment) data sets. Alzheimers Res Ther. 2021;13(1):1–12. [PubMed: 33397495]

42. Schneider LS, Sano M. Current Alzheimer's disease clinical trials: methods and placebo outcomes. Alzheimers Dement. 2009;5(5):388–397. [PubMed: 19751918]

43. Benedetti F, Carlino E, Piedimonte A. Increasing uncertainty in CNS clinical trials: the role of placebo, nocebo, and Hawthorne effects. Lancet Neurol. 2016;15(7):736–747. [PubMed: 27106073]

44. Franzen S, Smith JE, van den Berg E, et al. Diversity in Alzheimer's disease drug trials: the importance of eligibility criteria. Alzheimers Dement. 2021;18(4):810–823. [PubMed: 34590409]

45. Bateman RJ, Benzinger TL, Berry S, et al. The DIAN-TU Next Generation Alzheimer's prevention trial: adaptive design and disease progression model. Alzheimers Dement. 2017;13(1):8–19. [PubMed: 27583651]

46. Insel PS, Donohue MC, Sperling R, Hansson O, Mattsson-Carlgren N. The A4 study: β-amyloid and cognition in 4432 cognitively unimpaired adults. Ann Clin Transl Neurol. 2020;7(5):776–785. [PubMed: 32315118]

47. Salloway S, Farlow M, McDade E, et al. A trial of gantenerumab or solanezumab in dominantly inherited Alzheimer's disease. Nat Med. 2021;27(7):1187–1196. [PubMed: 34155411]

48. Tariot PN, Lopera F, Langbaum JB, et al. The Alzheimer's Prevention Initiative Autosomal-Dominant Alzheimer's Disease Trial: a study of crenezumab versus placebo in preclinical PSEN1 E280A mutation carriers to evaluate efficacy and safety in the treatment of autosomal-dominant

Alzheimer's disease, including a placebo-treated noncarrier cohort. Alzheimers Dement (N Y). 2018;4:150–160. [PubMed: 29955659]

49. Hassenstab J, Ruvolo D, Jasielec M, Xiong C, Grant E, Morris JC. Absence of practice effects in preclinical Alzheimer's disease. Neuropsychology. 2015;29(6):940–948. [PubMed: 26011114]

50. Machulda MM, Pankratz VS, Christianson TJ, et al. Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. Clin Neuropsychol. 2013;27(8):1247–1264. [PubMed: 24041121]

51. Wang G, Kennedy RE, Goldberg TE, Fowler ME, Cutter GR, Schneider LS. Using practice effects for targeted trials or sub-group analysis in Alzheimer's disease: how practice effects predict change over time. PLoS One. 2020;15(2):e0228064.

52. Jacobs DM, Ard MC, Salmon DP, Galasko DR, Bondi MW, Edland SD. Potential implications of practice effects in Alzheimer's disease prevention trials. Alzheimers Dement (N Y). 2017;3(4):531–535. [PubMed: 29124111]

53. Lievens F, Reeve CL, Heggestad ED. An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. J Appl Psychol. 2007;92(6):1672–1682. [PubMed: 18020804]

54. Calamia M, Markon K, Tranel D. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol. 2012;26(4):543–570. [PubMed: 22540222]

55. Mccaffrey RJ, Westervelt HJ, Haase RF. Serial neuropsychological assessment with the National Institute of Mental Health (NIMH) AIDS abbreviated neuropsychological battery. Arch Clin Neuropsychol. 2001;16(1):9–18. [PubMed: 14590189]

56. Falleti MG, Maruff P, Collie A, Darby DG. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. J Clin Exp Neuropsychol. 2006;28(7):1095–1112. [PubMed: 16840238]

57. Kaptchuk TJ, Miller FG. Placebo effects in medicine. N Engl J Med. 2015;373(1):8–9. [PubMed: 26132938]

58. Vos SJB, Verhey F, Frölich L, et al. Prevalence and prognosis of Alzheimer's disease at the mild cognitive impairment stage. Brain. 2015;138(5):1327–1338. [PubMed: 25693589]

59. Kim YJ, Cho SK, Kim HJ, et al. Data-driven prognostic features of cognitive trajectories in patients with amnestic mild cognitive impairments. Alzheimers Res Ther. 2019;11(1):10. [PubMed: 30670089]

60. Jutten RJ, Sikkes SAM, Van Der Flier WM, Scheltens P, Visser PJ, Tijms BM. Finding treatment effects in Alzheimer trials in the face of disease progression heterogeneity. Neurology. 2021;96(22):e2673-e2684.

61. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge University Press; 2011.

62. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–745. [PubMed: 20494804]

63. Kristensen LQ, Muren MA, Petersen AK, Van Tulder MW, Gregersen Oestergaard L. Measurement properties of performance-based instruments to assess mental function during activity and participation in traumatic brain injury: a systematic review. Scand J Occup Ther. 2020;27(3):168–183. [PubMed: 31725339]

64. Mokkink LB, De Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1171–1179. [PubMed: 29260445]

65. Soobiah C, Tadrous M, Knowles S, et al. Variability in the validity and reliability of outcome measures identified in a systematic review to assess treatment efficacy of cognitive enhancers for Alzheimer's Dementia. PLoS One. 2019;14(4):e0215225.

66. Cappelleri JC, Zou KH, Symonds T, et al. Patient-Reported Outcomes: Measurement, Implementation and Interpretation. Crc Press; 2013.

67. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159–1170. [PubMed: 29550964]

68. Papp KV, Buckley R, Mormino E, et al. . Clinical meaningfulness of subtle cognitive decline on longitudinal testing in preclinical AD. Alzheimers Dement. 2020;16:552–560. [PubMed: 31759879]

69. Ganz AB, Beker N, Hulsman M, et al. Neuropathology and cognitive performance in self-reported cognitively healthy centenarians. Acta Neuropathol Commun. 2018;6(1):1–13. [PubMed: 29298724]

70. Beker N, Sikkes SAM, Hulsman M, et al. Longitudinal maintenance of cognitive health in centenarians in the 100-plus study. JAMA Netw Open. 2020;3(2):e200094-e200094.

71. Brickman AM, Cabo R, Manly JJ. Ethical issues in cross-cultural neuropsychology. Appl Neuropsychol. 2006;13(2):91–100. [PubMed: 17009882]

72. Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and White elders. J Int Neuropsychol Soc. 2002;8(3):341. [PubMed: 11939693]

73. Puente SG, Van Eijck M, Jochems W. Empirical validation of characteristics of design-based learning in higher education. Int J Eng Educ. 2013;29(2):491.

74. Agranovich A, Puente A. Do Russian and American normal adults perform similarly on neuropsychological tests? Preliminary findings on the relationship between culture and test performance. Arch Clin Neuropsychol. 2007;22(3):273–282. [PubMed: 17331697]

75. Boone K, Victor T, Wen J, Razani J, Ponton M. The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. Arch Clin Neuropsychol. 2007;22(3):355–365. [PubMed: 17320344]

76. Franzen S, Vanâ Denâ Berg E, Kalkisim Y, et al. Assessment of visual association memory in low-educated, non-western immigrants with the modified visual association test. Dement Geriatr Cogn Disord. 2019;47(4–6):345–354. [PubMed: 31319408]

77. Manly JJ. Advantages and disadvantages of separate norms for African Americans. Clin Neuropsychol. 2005;19(2):270–275. [PubMed: 16019708]

78. Szlyk JP, Myers L, Zhang Y, Wetzel L, Shapiro R. Development and assessment of a neuropsychological battery to aid in predicting driving performance. J Rehabil Res Dev. 2002;39(4):483–496. [PubMed: 17638145]

79. Cahn-Weiner DA, Farias ST, Julian L, et al. Cognitive and neuroimaging predictors of instrumental activities of daily living. J Int Neuropsychol Soc. 2007;13(5):747–757. [PubMed: 17521485]

80. Gold DA. An examination of instrumental activities of daily living assessment in older adults and mild cognitive impairment. J Clin Exp Neuropsychol. 2012;34(1):11–34. [PubMed: 22053873]

81. Royall DR, Lauterbach EC, Kaufer D, Malloy P, Coburn KL, Black KJ. The cognitive correlates of functional status: a review from the Committee on Research of the American Neuropsychiatric Association. J Neuropsychiatry Clin Neurosci. 2007;19(3):249–265. [PubMed: 17827410]

82. Marshall GA, Sikkes SAM, Amariglio RE, et al. P4–245: the association between instrumental activities of daily living and cortical amyloid in cognitively normal older adults screening for the a4 study. Alzheimers Dement. 2019;15(7S_Part_26):P1372-P1372.

83. Liu-Seifert H, Siemers E, Price K, et al. Cognitive impairment precedes and predicts functional impairment in mild Alzheimer’s disease. J Alzheimers Dis. 2015;47(1):205–214. [PubMed: 26402769]

84. Baker JE, Lim YY, Jaeger J, et al. Episodic memory and learning dysfunction over an 18-month period in preclinical and prodromal Alzheimer’s disease. J Alzheimers Dis. 2018;65(3):977–988. [PubMed: 30103330]

85. Marshall GA, Aghjayan SL, Dekhtyar M, et al. Measuring instrumental activities of daily living in non-demented elderly: a comparison of the new performance-based Harvard Automated Phone Task with other functional assessments. Alzheimers Res Ther. 2019;11(1):1–12. [PubMed: 30611304]

86. Sikkes SA, de Lange-de Klerk ESM, Pijnenburg YAL, et al. A new informant-based questionnaire for instrumental activities of daily living in dementia. Alzheimers Dement. 2012;8(6):536–543. [PubMed: 23102123]

87. Jutten RJ, Peeters CFW, Leijdesdorff SMJ, et al. Detecting functional decline from normal aging to dementia: development and validation of a short version of the Amsterdam IADL Questionnaire. Alzheimers Dement (Amst). 2017;8:26–35. [PubMed: 28462387]

88. Koo BMi Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. Innov Aging. 2019;3(1):igy038.

89. Edgar CJ, Vradenburg G, Hassenstab J. The 2018 revised FDA guidance for early Alzheimer's disease: establishing the meaningfulness of treatment effects. J Prev Alzheimers Dis. 2019;6(4):223–227. [PubMed: 31686092]

90. Cook CE. Clinimetrics corner: the minimal clinically important change score (MCID): a necessary pretense. J Man Manip Ther. 2008;16(4):82E–83E. [PubMed: 19119392]

91. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res. 2021;30(10):2729–2755. [PubMed: 34247326]

92. Sperling R, Mormino E, Johnson K. The evolution of preclinical Alzheimer's disease: implications for prevention trials. Neuron. 2014;84(3):608–622. [PubMed: 25442939]

93. Rentz DM, Parra Rodriguez MA, Amariglio R, Stern Y, Sperling R, Ferris S. Promising developments in neuropsychological approaches for the detection of preclinical Alzheimer's disease: a selective review. Alzheimers Res Ther. 2013;5(6):58. [PubMed: 24257331]

94. Jutten RJ, Sikkes SAM, Amariglio RE, et al. Identifying sensitive measures of cognitive decline at different clinical stages of Alzheimer's disease. J Int Neuropsychol Soc. 2020:1–13.

95. Rentz DM, Papp KV. Commentary on composite cognitive and functional measures for early stage Alzheimer's disease trials. Alzheimers Dement (Amst). 2020;12(1):e12012.

96. Jonaitis EM, Koscik RL, Clark LR, et al. Measuring longitudinal cognition: individual tests versus composites. Alzheimers Dement (Amst). 2019;11:74–84. [PubMed: 31673596]

97. Schneider LS, Goldberg TE. Composite cognitive and functional measures for early stage Alzheimer's disease trials. Alzheimers Dement (Amst). 2020;12(1):e12017.

98. Dong A, Toledo JB, Honnorat N, et al. Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: links to cognition, progression and biomarkers. Brain. 2017;140(3):735–747. [PubMed: 28003242]

99. Dowling NM, Hermann B, La Rue A, Sager MA. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. Neuropsychology. 2010;24(6):742. [PubMed: 21038965]

100. Langbaum J, Hendrix S, Ayutyanont N, et al. Establishing composite cognitive endpoints for use in preclinical Alzheimer's disease trials. J Prev Alzheimers Dis. 2015;2(1):2. [PubMed: 26273569]

101. Donohue MC, Sun CK, Raman R, et al. . Cross-validation of optimized composites for preclinical Alzheimer's disease. Alzheimers Dement (N Y). 2017;3(1):123–129. [PubMed: 28758145]

102. Dubbelman MA, Jutten RJ, Tomaszewski Farias SE, et al. Decline in cognitively complex everyday activities accelerates along the Alzheimer's disease continuum. Alzheimers Res Ther. 2020;12(1):1–11.

103. Cloutier S, Chertkow H, Kergoat MJ, Gélinas I, Gauthier S, Sylvie S. Trajectories of decline on instrumental activities of daily living prior to dementia in persons with mild cognitive impairment. Int J Geriatr Psychiatry. 2021;36(2):314–323. [PubMed: 32892375]

104. Webster L, Groskreutz D, Grinbergs-Saull A, et al. Development of a core outcome set for disease modification trials in mild to moderate dementia: a systematic review, patient and public consultation and consensus recommendations. Health Technol Assess. 2017;21(26):1–192.

105. Mormino EC, Papp KV, Rentz DM, et al. Early and late change on the preclinical Alzheimer's cognitive composite in clinically normal older individuals with elevated amyloid β. Alzheimers Dement. 2017;13(9):1004–1012. [PubMed: 28253478]

106. Papp KV, Rentz DM, Orlovsky I, Sperling RA, Mormino EC. Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: the PACC5. Alzheimers Dement (N Y). 2017;3(4):668–677. [PubMed: 29264389]

107. Jutten RJ, Harrison J, Lee Meeuw Kjoe PR, et al. A novel cognitive-functional composite measure to detect changes in early Alzheimer's disease: test–retest reliability and feasibility. Alzheimers Dement (Amst). 2018;10:153–160. [PubMed: 29780863]

108. Jutten RJ, Dicks E, Vermaat L, et al. Impairment in complex activities of daily living is related to neurodegeneration in Alzheimer's disease–specific regions. Neurobiol Aging. 2019;75:109–116. [PubMed: 30557769]

RESEARCH IN CONTEXT

Systematic review:

The authors reviewed PubMed for literature on clinical outcome assessments (COAs) to assess treatment efficacy in Alzheimer's disease (AD) clinical trials. Compared to the rapidly evolved measures of biomarkers in predementia stages of AD, relatively less progress has been made in the evolution of COAs for those stages.

Interpretation:

Based on lessons learned from past AD trials, we propose a multi-step framework to advance the selection and implementation of more effective COAs in clinical trials of early AD. This framework can be applied when selecting existing tests or items for a composite measure to include as a COA as well as when developing entirely novel measurement instruments for cognition to implement in future clinical trials.

Future directions:

This manuscript provides concrete recommendations that could contribute to the much-needed consensus and use of more appropriate COAs to assess the efficacy of putative treatments for AD.

Highlights

- We discuss lessons learned on capturing cognitive changes in predementia stages of AD.
- We propose a framework for the design and evaluation of performance based cognitive tests for use in early AD trials.
- We provide recommendations to facilitate the implementation of more effective cognitive outcome measures in AD trials.

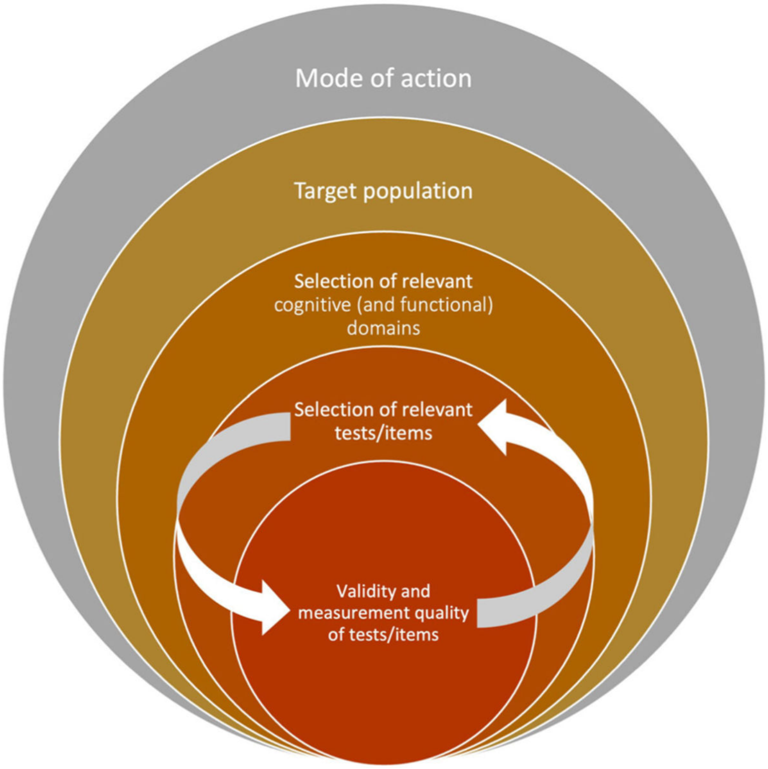
Mode of action
Target population
Selection of relevant cognitive (and functional) domains
Selection of relevant tests/items
Validity and measurement quality of tests/items

FIGURE 1.

Recommendation for selecting clinical outcome assessments (COAs) for Alzheimer's disease trials based on the hypothesized mode of action, target population (e.g., clinical disease stage) and measurement characteristics of available tests

TABLE 1

FDA 2018 Draft guidance recommendations on clinical outcome assessments (COAs) by clinical stage of Alzheimer's disease.[4]

Stage	Description	Diagnostic label	Recommendations on COAs as described in FDA 2018 draft guidance
1	Patients with characteristic pathophysiologic changes of AD but no evidence of clinical impact.	Preclinical AD	- A clinically meaningful benefit in cognition cannot be measured in these patients because there is no clinical impairment to assess (assuming that the duration of a trial is not sufficient to observe and assess the development of clinical impairment during the conduct of the trial). - An effect on the characteristic pathophysiologic changes of AD, as demonstrated by an effect on various biomarkers, analyzed asa primary efficacy measure, may, in principle, serve as the basis for an accelerated approval - As with the use of neuropsychological tests, a pattern of treatment effects seen across multiple individual biomarker measures would increase the persuasiveness of the putative effect.
2	Patients with characteristic pathophysiologic changes of AD and subtle detectable abnormalities on sensitive neuropsychological measures, but no functional impairment.	Preclinical AD	- A possible approach is to conduct a study of sufficient duration to allow the evaluation of the measures discussed below for Stage 3 patients - Alternatively, FDA will consider strongly justified arguments that a persuasive effect on sensitive measures of neuropsychological performance may provide adequate support for a marketing approval. - A pattern of putatively beneficial effects demonstrated across multiple individual tests would increase the persuasiveness of the finding; conversely, a finding on a single test unsupported by consistent findings on other tests would be less persuasive
3	Patients with characteristic pathophysiologic changes of AD, subtle or more apparent detectable abnormalities on sensitive neuropsychological measures, and mild but detectable functional impairment.	Prodromal AD/MCI due to AD	- Ideally, the outcome measure used in this stage of disease will provide an assessment of meaningful cognitive function. - An integrated scale that adequately and meaningfully assesses both daily function and cognitive effects is acceptable as a single primary efficacy outcome measure. - The independent assessment of daily function and cognitive effects is also an acceptable approach. In this setting, effect on a sensitive measure of neuropsychological performance of uncertain independent clinical meaning (e.g., a word-list recall test) should not allow for an overall finding of efficacy in the absence of meaningful functional benefit.

Abbreviations: AD, Alzheimer's disease; COAs, clinical outcome assessments; FDA, Food and Drug Administration.

TABLE 2

Overview of challenges and lessons learned from Alzheimer's disease (AD) trials with regard to clinical outcome assessments (COAs), as well as a recommendation framework for the evaluation of performance based cognitive tests based on the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) methodology

Challenge	**Whatwe have learned**	**Related validity aspects**	**Considerations**	**Recommendations (for examples of applications, see Section 3.3)**
What to measure	Existing tests are less focused on the cognitive processes that are relevant in early (pre)clinical stages of AD.	Content validity (including face validity) Interpretability (including ecological validity)	All items refer to relevant aspects of the construct to be measured AND together comprehensively reflect the construct to be measured. All test aspects (items, Instructions, response options) are understood by the target population as Intended.	a. Test should have a clear theoretical framework with respect to the construct of Interest and target population. b. The target population and clinical experts should be involved In development of test content and material. c. Test should be shown to be suitable (i.e., comprehensible, relevant, comprehensive, culturally appropriate) for the target population.
How to measure	Tests are insufficiently sensitive to disease progression in early stages of AD. It remains unclear whether (changes in) test scores are clinically meaningful.	Construct validity (including structural validity) Interpretability (Including clinical meaningfulness)	The scoring algorithm (single score, weighted/ unweighted composite score) is an adequate reflection of the dimensionality (one or multiple domains) being measured. Test captures the construct of interest and Is associated with related constructs and not associated with unrelated constructs. Test is sensitive to disease progression and stable In the absence of change.	a. A factor analysis should be performed to demonstrate adequate model fit and score use (single score or domain scores). b. Test score should be validated against relevant clinical (e.g., everyday functioning, quality of life) and/or biological measures (e.g., neurodegeneration, amyloid or tau biomarkers) and relationships are in line with hypotheses. c. An Independent validation study on the ability to capture cognitive change in the target population should be performed. d. It should be examined what (change In) scores patients and caregivers perceive as clinically meaningful.
Who to measure	Tests are insufficiently suitable across groups (e.g., different cultures). Tests are differentially sensitive to change across different stages of AD.	Cross-cultural validity (Including measurement Invariance) Longitudinal validity (Including responsiveness)	Test performance is not Influenced by cross-cultural differences OR a cross-cultural adaptation has been made. Test is equally sensitive to change across disease stages OR is known to be sensitive to change In the specific target population.	a. A representative and diverse target population should be Involved in test development and validation. b. Test should have no measurement Invariance across groups and show no Important differential item functioning when comparing groups based on demographic characteristics (such as age, sex, education) and cultural aspects. c. The selection of tests/items should be adapted to disease stage of Interest. d. Test should show no or limited range restrictions in scoring in the target population.