Composite neurocognitive endpoints in Alzheimer’s disease clinical trials: A commentary

Christopher Randolph

1Department of Neurology, Loyola University Medical Center, Maywood, Illinois
2MedAvante-Prophase

Correspondence
Christopher Randolph, Department of Neurology, Loyola University Medical Center, Maywood, IL.
E-mail: crandol@lumc.edu

Schneider and Goldberg raise several good points in their review of the use of “composite” scales as endpoints in early-stage Alzheimer’s disease (AD) trials (and by extension, early-stage disease in other neurodegenerative dementing disorders). Some of their comments warrant additional explanation, however, and others beg some counterargument.

History: With the advent of the acetylcholinesterase inhibitors, the U.S. Food and Drug Administration (FDA) issued guidance on AD trial endpoints, requiring the dual endpoints of a neurocognitive measure and a “global” measure. The rationale for this was that the clinical meaningfulness of a statistically significant difference on a neurocognitive scale such as the Alzheimer’s Disease Assessment Scale—Cognitive Subscale (ADAS-Cog) was unknown, and that a separate measure of global clinical change would be needed to confirm efficacy. Clinical Global Impression (CGI) scales were initially more commonly used, followed by the substitution of Activities of Daily Living (ADL) scales such as the Alzheimer’s Disease Cooperative Study (ADCS-ADL) for longer trials for which it was assumed change from baseline would be more difficult to measure with CGI scales (although this was never demonstrated empirically). This approach continues to be implemented in trials of mild-moderate dementia due to AD. With the advent of trials in mild cognitive impairment (MCI) and very mild dementia, it became apparent that the ADAS-Cog was unsuitable as a neurocognitive endpoint due to ceiling/practice effects. Before a more suitable neurocognitive battery was established as a standard for this population, the FDA issued draft guidance in 2013 as Schneider and Goldberg note. This guidance indicated that a single composite measure of cognition and function could be acceptable as a sole primary endpoint for this study population, and they provided the CDR as an example. Unsurprisingly, virtually every pivotal trial in this study population since that publication has included the CDR sum of boxes as a primary endpoint. When the first secondary prevention trials were being planned, longitudinal cohort data for at-risk individuals were fairly consistent in demonstrating that neurocognitive decline was measurable several years before the onset of clinically detectable symptoms or functional impairments. As a result, prevention trials moved to various neurocognitive composites as endpoints, and this was recognized as appropriate in the more recent FDA guidance.

Regulatory issues with composites: Apart from the need for any measure to demonstrate suitable psychometric and clinimetric characteristics (and ultimately predict clinically meaningful outcomes), concerns have been raised by the FDA when it comes to composites that mix measurement methodologies. Therefore, while they are accepting of the CDR, which involves ordinal ratings of cognitive and functional impairment based solely upon interview data, they have been critical of scales that mix interview-based data collection with direct examination data, such as the Progressive Supranuclear Palsy Rating Scale (PSPRS). This raises concerns for scales like the Alzheimer’s Disease Composite Scores (ADCOMS), reviewed by Schneider and Goldberg. From a psychometric perspective, this is a legitimate concern, as such scales mix objective psychometric (or neurologic) data with subjective subject (and/or informant) self-report data. It is conceivable with such scales that movement of one of the components could shift the total scale in the absence of movement of the other.

Composite design pitfalls: The majority of the composites implemented in secondary prevention trials were derived from longitudinal cohort data from the at-risk population of interest—either due to...
a genetic risk or evidence of current cerebral amyloid load. The rationale for this is essentially that there would be no other way to conduct a power analysis to support an interventional trial. The pitfalls of this approach, however, are numerous. They include only being able to utilize the tests that were included in the original longitudinal cohort study, which may be less than optimal for the intended purpose; typically do not have alternate forms; may be culturally bound; and usually are administered only on an annual basis; not on 3- or 6-month intervals as is typically the case in clinical trials. This poses a risk for increased practice effects that may flatten the trajectory of the placebo group and obscure a treatment signal. Study teams have attempted to mitigate this problem by creating alternate forms in which possible or substituting subtests from scales that already have equivalent alternate forms, such as the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). It should also be noted that there is an assumption with a strategy of extracting the measures that appear to be most sensitive from a larger battery that these measures will behave in the same manner longitudinally when administered separately. This ignores the possible effects of proactive and retroactive interference upon performance. That is, larger neurocognitive batteries that contain subtests that overlap neurocognitive domains tend to drive down performance on constituent subtests. Substantially reducing the number of subtests in the battery may eliminate sources of interference and subsequently boost performance on the remaining subtests.

Clinical meaningfulness and normative data: Treatment effects on any composite measure must ultimately translate into clinically meaningful effects to achieve/maintain regulatory approval. Demonstrating that a treatment slows the trajectory of neurocognitive decline in a preclinical sample is only the first step to establishing clinically meaningful change. In addition, with the exception of the RBANS, none of the composites reviewed by Schneider and Goldberg have any normative data, and interpretation of change in performance on composites without normative data is extremely difficult. With existing normative data, it is much easier to understand change scores. For example, if a preclinical sample of subjects at baseline has an RBANS total scale index score of 100, it is clear that they are at the normal mean at baseline (50th percentile). If, after 24 months that sample’s mean score is 85, as a group they have fallen from the 50th percentile of their age norm to the 16th percentile. This is change of a magnitude that is easily understood even by a lay audience. The availability of normative data to aid in the interpretation of performance is a parameter that should be added to the table created by Schneider and Goldberg.

Neurocognitive domain selection and measurement considerations: Schneider and Goldberg raise questions about the assumption that a composite measure will be more effective in tracking disease progression than a single domain measure, how certain measures are conceptualized from a domain standpoint, whether orientation is an optimal measure of anterograde memory, and whether the use of differential subtest weighting is psychometrically appropriate. These are all valid concerns. Differential weighting of subtest scores is a psychometrically dubious strategy, has the potential for introducing instability in measurement that may suggest change in global neurocognitive status that is not truly reflective of a subject’s overall performance, and may even produce counterintuitive findings. Although Schneider and Goldberg suggest the use of a factor analytic approach to establish domain construct validity, this is actually not an optimal approach, and may be highly misleading. A more suitable approach is the exploration of convergent and divergent validity, in conjunction with clinical validation via group studies. The questions raised by them regarding the use of orientation measures in many composites are valid. Orientation is obviously dependent upon the integrity of anterograde memory and Schneider and Goldberg point out that this measure is largely useless in cognitively normal subjects and even in many subjects with MCI due to ceiling effects. The reason that orientation measures appear to emerge as useful in tracking progression in symptomatic AD is likely due to the limited measurement of anterograde memory with other measures in these longitudinal cohort studies. In such situations, orientation may be able to capture variance in anterograde memory status that is not otherwise measured. Finally, Schneider and Goldberg question whether a global or composite measure is intrinsically more suitable to measuring neurocognitive decline in AD than a single domain measure. In considering this question, it is important to note that individuals with AD vary quite a bit with respect to the degree of impairment in different neurocognitive domains in the earlier stages of the disease. Impairments of language and visuospatial function (for example) may produce as much, or more, distress and disability as impairment of anterograde memory. By focusing on a single domain, clinically relevant neurocognitive decline may be missed. In addition, it is highly unlikely that a single domain measurement would be as sensitive to disease progression as multidomain measurement, provided that all domain measurements met appropriate psychometric and clinimetric standards (not always a given).

Summary: Schneider and Goldberg pose a number of provocative questions about the conceptual, psychometric, and practical considerations in the creation of composite neurocognitive endpoints in early AD trials. I share a number of their concerns, particularly with the practice of relying upon sampling measures only from preexisting longitudinal cohort studies, the pitfalls of which are reviewed above. The European Prevention of Alzheimer’s Disease (EPAD) Scientific Advisory Group for Clinical and Cognitive Outcomes took a more systematic, literature-based approach to endpoint selection. To summarize the properties of a suitable battery that were explored, the characteristics are as follows:

1. Should be sensitive to all (or most) of the neurocognitive deficits associated with disease stage. In this case, early symptomatic AD.
2. Should have full range of measurement (absence of floor or ceiling effects) for target population. In this case, normal older adults through MCI due to AD into the early mild dementia due to AD stage.
3. Should have well-established psychometric (eg, reliability) and clinical validity.
4. Should have coherence with biomarker changes.
5. Should be cross-culturally applicable.
6. Should have alternate forms with evidence of minimal practice effects.

The only existing neurocognitive test battery that met these criteria was the RBANS, which had the additional appeal of having well over 30 validated translations readily available. For the EPAD longitudinal cohort study, the RBANS was also supplemented with some experimental exploratory neurocognitive measures.

Finally, as a field, we had the opportunity to compare the performance of different iterations of the composite batteries and the RBANS in multiple secondary prevention trials—unfortunately, some of these trials have been halted due to toxicity or futility. This may limit a head-to-head comparison of the sensitivity of the various batteries to disease progression, but enough work continues to enable the production of benchmarking data for some of the batteries in ongoing studies, and for comparison of the performance of these batteries with other clinical endpoints and biomarker data.

DISCLOSURES
Dr. Randolph is the author of the RBANS and receives royalties from the copyright holder, Pearson, Inc.

REFERENCES
1. Leber P. Guidelines for the Clinical Evaluation of Antidementia Drugs. First Draft. New Canaan, CT: US Food and Drug Administration; 1990.
2. FDA. United States Food and Drug Administration. Guidance for industry Alzheimer’s disease: Developing drugs for the treatment of early stage disease (FDA-2013-D-0077) DRAFT (US Food and Drug Administration, Center for Drug Evaluation and Research), 2013.
3. FDA. Early Alzheimer’s disease: Developing drugs for treatment guidance for industry. Available at: https://www.fda.gov/...fda-guidance.../alzheimers-disease-developing-drugs-treatment. Published February 2018.
4. Donohue MC, Sperling RA, Salmon DP, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol. 2014;71(8):961-970. https://doi.org/10.1001/jamaneurol.2014.803
5. Delis DC, Jacobson M, Bondi MW, Hamilton JM, Salmon DP. The myth of testing construct validity with factor analysis or correlations with normal or mixed clinical populations: lessons from memory assessments. J Int Neuropsych Soc. 2003;9:936-946.
6. Ritchie K, Ropacki M, Albala B, et al. Recommended cognitive outcomes in preclinical Alzheimer’s disease: consensus statement from the European Prevention of Alzheimer’s Dementia project. Alzheimers Dement. 2017;13(2):186-195. https://doi.org/10.1016/j.jalz.2016.07.154

How to cite this article: Randolph C. Composite neurocognitive endpoints in Alzheimer’s disease clinical trials: A commentary. Alzheimer’s Dement. 2020;12:e12010. https://doi.org/10.1002/dad2.12010