ELSEVIER

Alzheimer's & Dementia 5 (2009) 324–339

Alzheimer's & Dementia

Perspectives

Scales as outcome measures for Alzheimer's disease

Ronald Black[a], Barry Greenberg[b], J. Michael Ryan[a], Holly Posner[c], Jeffrey Seeburger[d], Joan Amatniek[e], Malca Resnick[f], Richard Mohs[g], David S. Miller[h], Daniel Saumier[i,j], Maria C. Carrillo[k,*], Yaakov Stern[l]

[a]*Wyeth Research, Collegeville, PA, USA*
[b]*Toronto Western Research Institute, Toronto, Ontario, Canada*
[c]*Eisai Medical Research, Inc., New York, NY, USA*
[d]*Merck Research Laboratories, North Wales, PA, USA*
[e]*Ortho-McNeil Neurologics, Inc., Raritan, NJ, USA*
[f]*Forest Laboratories, Inc., New York, NY, USA*
[g]*Eli Lilly & Co., Indianapolis, IN, USA*
[h]*United BioSource Corporation, Wayne, PA, USA*
[i]*Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada*
[j]*BELLUS Health, Inc., Laval, Quebec, Canada*
[k]*Alzheimer's Association, Chicago, IL, USA*
[l]*Department of Clinical Neuropsychology, Taub Institute, Columbia University, New York, NY, USA*

Abstract The assessment of patient outcomes in clinical trials of new therapeutics for Alzheimer's disease (AD) continues to evolve. In addition to assessing drugs for symptomatic relief, an increasing number of trials are focusing on potential disease-modifying agents. Moreover, participants with AD are being studied earlier in their course of disease. As a result, the limitations of current outcome measures have become more apparent, as has the need for better instruments. In recognition of the need to review and possibly revise current assessment measures, the Alzheimer's Association, in cooperation with industry leaders and academic investigators, convened a Research Roundtable meeting devoted to scales as outcome measures for AD clinical trials. The meeting included a discussion of methodological issues in the use of scales in AD clinical trials, including cross-cultural issues. Specific topics related to the use of cognitive, functional, global, and neuropsychiatric scales were also presented. Speakers also addressed academic and industry initiatives for pooling data from untreated and placebo-treated patients in clinical trials. A number of regulatory topics were also discussed with agency representatives. Panel discussions highlighted areas of controversy, in an effort to gain consensus on various topics.

Keywords: Neuropsychological assessment; Outcome measure; Cognitive testing; Functional scale; Global scale

1. Introduction

Before 1984, consensus measures did not exist for diagnosing or assessing the progression of Alzheimer's disease (AD). Clinical trials were heterogeneous, inclusion criteria were vague and various, and outcomes were idiosyncratic. For example, diagnostic inclusion criteria comprised vague entities such as organic brain syndrome, senile cognitive decline, or organic psychosyndrome. Outcomes included miscellaneous clinician rating scales and various neuropsychological subscales. In the early 1980s, there were attempts to arrive at a consensus on criteria and measures. The National Institute of Neurological and Communicative Disorders and Stroke—Alzheimer's Disease and Related Disorders Criteria for Alzheimer's Disease (also known as the McKhann criteria) were proposed and immediately applied as inclusion criteria in dementia trials [1]. Early clinical trial work with physostigmine in healthy participants and in participants with AD led to the development of the Alzheimer's Disease

*Corresponding author. Tel.: 312-335-5722; Fax: 866-741-3716.
E-mail address: Maria.Carrillo@alz.org

1552-5260/09/$ – see front matter
doi:10.1016/j.jalz.2009.05.667

Assessment Scale (ADAS) as a cognitive-assessment instrument and outcome, specifically for clinical trials [2,3]. Other neuropsychological assessments were also developed for trials at this time.

Early experiences in AD clinical trials led a United States Food and Drug Administration (FDA) advisory panel in 1989 to recommend that AD clinical trials be at least 3 months, preferably 6 months, in duration and use a standard cognitive-assessment instrument and a clinician's global assessment as primary outcomes. As a result, the vast majority of AD registration clinical trials have been 6 months long, and used the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog) [3] as the primary cognitive outcome. Over time, 6 months were considered insufficient, and trials were lengthened to 12 months. Most of these trials targeted participants with mild to moderate levels of AD severity.

Recently, with the increasing interest in disease modification and the enrollment of participants at milder stages of AD, AD clinical trials have been lengthened to 18 months. Most trials still rely on the Mini-Mental State Examination (MMSE) [4] and Clinical Dementia Rating (CDR) [5] for staging, the ADAS-Cog for cognitive outcomes, the Alzheimer's Disease Cooperative Study-Activity of Daily Living (ADCS-ADL) [6] or the Disability Assessment for Dementia (DAD) [7] for activities of daily living, the CDR or Clinician's Global Impression of Change (CGIC) [8] for global clinical measures, and the Neuropsychiatric Inventory (NPI) [9] for assessing behavior. However, the move toward 18-month trials presents some significant technical issues, such as subject retention and how to handle the increased number of dropouts [10].

Current outcome measures may vary greatly in the linearity of decline over these longer trials, and in their relative sensitivity to change across different strata of disease severity. This raises questions about the scales used in trials to measure progression, and about what should be considered a meaningful difference in the ADAS-Cog and other measures. There is also a perception that control groups are not deteriorating as rapidly as they did in the 1980s and early 1990s on the scales now used in trials, although no clear evidence exists for this.

One issue with current measures, including the ADAS-Cog, involves the broad distribution of cognitive test scores at baseline, even within a narrowly defined group such as mild AD. Moreover, within-subject change is highly variable, with considerable overlap in scores and standard deviations between placebo and treated groups. Overall, the clinical decline in patients in placebo groups may be relatively small compared with the variability in patients, even in 18-month trials, so it may be difficult to detect a drug treatment effect if one exists. Irizarry et al., examining individual scores over 6, 12, or 18 months, reported considerable participant variation, wherein some deteriorated and some actually showed improvement while on a placebo [11]. This variability seems to increase over time, suggesting that for longer trials, the data may produce greater deviations. In addition to between-subject differences, ADAS-Cog variability is compounded by site-to-site and country-to-country differences.

2. Methodological issues in clinical trials

Cognitive scales are essential for AD clinical trials because decline in cognition is the defining symptom. For this reason, cognitive tests are generally given in phase 2 trials, and are a regulatory requirement in phase 3.

2.1. Measurement properties of cognitive tests

A good cognitive test or test battery for AD trials should sample all major cognitive functions affected by AD, should be sensitive over a range of impairment levels, reliable, and valid, should have minimal floor and ceiling effects, should be sensitive to longitudinal changes with minimal practice (learning) effects, and should provide a composite measure of overall performance. The test must also work in the real world, and cannot overtax study participants. Information on practice effects is important, and the availability of equivalent forms for repeated measurements is necessary [12].

Perhaps what is most required in any cognitive test is content validity, or the extent to which the test actually measures what it is intended to measure. Validity cannot be achieved without good interrater reliability. Sensitivity over a range of cognitive-ability levels is becoming increasingly important as trials in cognitively normal people and people with mild dementia become more common. There is a need to improve psychometric properties in this regard. Test bias is often not adequately addressed, and can be problematic, especially when transferring tests to other languages or cultures. Test bias, because of changing psychometric raters during trials, is also a potential issue. It becomes increasingly difficult to maintain rater consistency as trials become increasingly longer.

The ADAS-Cog has been the gold standard for cognitive assessment in clinical trials, but has some limitations. It does not adequately measure certain domains, including delayed memory, attention, and executive function. The Alzheimer's Disease Cooperative Study (ADCS) has improved upon this by adding new tests that address those components [13]. In addition, floor effects make it less useful for longitudinal studies with severe AD. Instead, the Severe Impairment Battery or Modified Ordinal Scales of Psychological Development are often used [14,15]. At the other end of the spectrum, more sensitive cognitive tests are needed for participants with mild cognitive impairment (MCI). Primary prevention trials may require a different set of tests to detect very small changes in memory, typically the first domain affected, at the normal end of the spectrum. Measurement scales that can be conducted at home or over the phone would also be advantageous.

An important factor in the implementation of cognitive scales is that the numbers generated by the scale in question are measurements of the central dependant variable (in the context of assessing an AD patient, the variable would represent functional or physical states of the brain) on which clinical decisions are based. Poorly chosen rating scales can

jeopardize the success of a clinical trial. The core requirements for any scale are that: 1) the numbers generated are linked to the measurements, and 2) the items on the instrument are linked to the variable it intends to measure [16,17].

2.2. Analysis and interpretation of cognitive-scale data

Numbers do not directly translate to measurements unless a theory or definition links the two. Old theories posit that observed scores can be equated to true scores plus some error, but this method has been difficult to test. Two new psychometric theories, the Rasch measurement (RM) theory and item response theory (IRT), attempt to correlate numbers and measurements in a more robust fashion [18–22].

The RM and IRT articulate theories about the relationship between numbers generated by rating scales and measurements. These are complex theories and require sophisticated software for analysis, but they deliver the type of measurements clinical trials want and need in rating-scale data.

These two theories can provide information about the scales and how they relate to the persons being measured. Analyses show how items line up along a continuum, where measurements can be improved, when response categories are working or not, and when the scale and sample are mistargeted [23]. They indicate whether a measurement is stable over time and across different samples. The analyses allow ''item banking'' or the addition of items to a single scale to improve sensitivity.

The RM and IRT can impart significant information about participants. For example, the relationship between raw scores and linear measurements is almost always S-shaped, i.e., a change of 10 points in the middle of the ADAS-Cog involves about 0.6 units in the middle of the scale, but 21 units at the end. With RM and IRT, the curve is apparent and can easily be followed. The RM and IRT also allow direct comparisons of person measures. They can equate across samples, identify whether people from different samples can be directly compared, and identify outliers.

2.3. Statistical analysis of trial data

Another important component that has evolved over the years involves the methods for analyzing longitudinal data. The selegiline/alpha-tocopherol trial published in 1997 [24], for example, was analyzed based on survival endpoints, or a random-effects model. Since then, analysis of covariance (ANCOVA) [25] has been more frequently used, and the generalized estimating equation (GEE) [26] and random-effects modeling approaches were recently used to analyze psychometric data. Analysis of covariance is frequently used by imputing missing data according to the last-observation-carried-forward approach, whereas GEE and random-effects models assume that progression can reasonably be modeled based on linear changes over time, which may or may not be true.

Many factors can be weighed when choosing methods of analysis, but perhaps the three most common considerations are distribution of scale, the extent of missing data, and design. The recent ADCS trial of nonsteroidal anti-inflammatory agents (NSAIDs), which used ANCOVA, demonstrated that ADAS-Cog has a good distribution of scale [27]. The later homocysteine (HC) trial, using GEE, also demonstrated good Gaussian distribution [28].

Dealing with missing data, however, has been an issue in most neurodegenerative trials. In both the NSAID and HC trials, there was a 20% dropout, but the HC trial continued for 18 months, whereas the NSAID trial was 1 year in duration. In terms of trial design, models such as GEE have the advantage of providing a formal approach to deal with missing data. Traditionally, AD trials have defined their primary analyses in terms of baseline and final data points, while ignoring intermediate data. Alternatively, modeling approaches such as GEE use all of the data and deal with dropouts in a rational way, without requiring an imputation of missing values. The GEE also handles correlation structure, i.e., changes in scales and variance through time, and the within-subject relationship between test scores over time. One drawback of GEE, however, is its requirement of a ''missing completely at random'' assumption. This assumption requires that missing data and dropouts be independent of all data, observed and unobserved. This is a step down from ''missing at random'' (MAR), which allows missing data to be dependent on observed data. The MAR assumption allows for a differential dropout between treatments. Random-effects models require a MAR assumption, which is an advantage, although random-effects models also require that a correlation structure be specified a priori, which can be challenging.

Some additional considerations are becoming increasingly important. Floor and ceiling effects are becoming an issue as more severely and less severely affected populations are being followed [29]. This was apparent in individual ''spaghetti'' plots in the selegiline trial, which showed that in the severe group, the data reached a plateau at the bottom end of the scale for some participants. Nonlinearity is also becoming a concern as trials are conducted over longer and longer follow-up periods. The placebo group might follow a linear regression, for example, but the treatment arm might be biphasic. In fact, both the NSAID and the HC trials suggested nonlinearity, although linearity still seems the best fit [30]. Linear splines may be one strategy for dealing with nonlinearity. In linear splines, data are modeled to a series of linear lines, each jointed at the various assessment visits. Otherwise, using an ANCOVA method that focuses on final-visit endpoints may be the best option.

2.4. Assessment settings in clinical trials

For large-scale prevention trials, there are obvious advantages to conducting cognitive testing in the home rather than in clinical settings, especially in rural communities. In-home assessment may result in a more representative population sample, better retention, and lower cost, but there are challenges, including the assessment itself, data capture and

remote data collection, and the general need for and implementation of new technologies. The most important question concerns whether in-home assessment will be embraced by study participants.

The Home-Based Assessment Trial is an ongoing study aimed at testing the feasibility of in-home testing [31,32]. It involves three different data-collection methods: mail-in with live telephone backup; automated telephone, using interactive voice recognition software; and computer-based. The protocols focus on several domains, including cognitive, functional, global, behavioral, quality of life, and pharmacoeconomic. The trial will evaluate the instruments and measure medication adherence, in this case using a vitamin capsule. This 4-year trial includes individuals aged 75 years and older who are not demented, who do not have any neurologic or neurodegenerative disorder, and who are not on prescription cognition-enhancing drugs. The participants are living independently, which for the purpose of this study means that they have control over their mail, telephone, and computer. Its design also ensures that 1 in 5 participants are from a diverse population or minority group. Trial sites are required to perform assessments and in-home visits, to ensure the best recruitment and retention. The study was designed to enroll 600 participants, but a small pilot study was conducted first to assess feasibility and to develop standard operating procedures. The pilot trial also measured how efficient the full trial might be, i.e., how often participants needed telephone help or a site visit to deal with their computer kiosk.

Recruitment for the pilot showed some surprising reasons for nonparticipation, including the size of the computer, fear of having strangers in the home, or unwillingness to switch from their current vitamin regimen. Of 60 individuals, only 48 were randomized, and 9 dropped out. Seven of those dropouts were in the computer arm, suggesting that the technological aspect might have been more intimidating than first imagined.

The computer kiosk was most intensive in terms of training, but training seemed to be retained. Cognitive assessments demonstrated good test-retest reliability between baseline and 1-month assessments. According to the overall impression, participants were not as enthusiastic about technology in their home, and training took longer than anticipated. At the time of preparation of this paper, the full trial had screened 289 subjects and randomized 245. More than half were aged over 80 years, and their level of education was fairly high.

2.5. Cultural issues in cognitive testing

Cognitive testing does not always translate across cultures or geographic boundaries. One question of importance to ''low-and-middle-income countries'' concerns whether clinical tests from developed nations translate across cultural and socioeconomic boundaries. In assessing that question, there are both cultural and methodological boundaries to consider. Different cultures may regard memory loss in different ways, and not all may think of it as a disease. For that reason, dementia may not be treated equally or reported in all countries. Likewise, not all methodologies will be equally applicable throughout the globe. Predictions suggest that by 2040, 71% of people with dementia will be in developing countries, so it is imperative that we learn more about the disease in those populations [33].

Even if dementia were to be reported equally across countries, cross-cultural issues remain in terms of measurement. Appropriate instruments are not available in many local languages, or measures may not be culturally validated, such that measurement characteristics (e.g., reliability, sensitivity, and specificity) may change across populations. Often, local norms are nonexistent, as are norms for people with little or no formal education. Because of all these uncertainties, comparing studies in different countries is fraught with problems. Scores may not necessarily be comparable. An MMSE score of 23 may represent very different levels of impairment in different populations.

Terms such as ''culture'' are used in different ways. Cultural differences are often invoked with regard to studies carried out within a given country's ''ethnic minorities,'' but a minority in one country could be the majority in another. In cross-cultural studies, equivalence of assessment is the goal. The measure should tap into the same cognitive domain in all populations.

Language and linguistic structure may heavily influence testing ability as well. A reading measure that was devised in English, and that relies on nonphonetic spelling, may not translate well into Spanish, which is spelled phonetically, and even those with little formal education can read as well as those with high school degrees. Other languages have different grammatical structures that make ''translating'' tests difficult. In English, for example, the MMSE requires only a verb and a subject for sentence construction. This requirement would not make sense in a language such as Japanese, which requires noun modifiers for a full sentence.

An Indo-American study that tested participants in Pittsburgh and Ballabgarh, India, exemplified some cross-cultural problems [34]. A large proportion (75%) of the population-based sample in India was illiterate, and could not read a written word list. The list had to be read aloud to them, thus changing a test condition. The test group had little difficulty with category fluency. However, they could not perform initial letter or phoneme fluency tasks. These tasks require words to be thought of as having an initial sound, which is a meaningless concept for illiterate individuals. Standard naming tests were problematic because the volunteers were not accustomed to two-dimensional graphic representations. They could, however, name objects without difficulty. Visuospatial tasks involving drawing or copying were virtually impossible for older adults who had never previously held a pencil. An alternative visuospatial task, such as arranging sticks in a particular pattern, is feasible, as demonstrated in a Nigerian study [35].

Merely translating a test into another language will not deliver a valid measurement, nor should trial investigators

be asked to translate all or parts of a scale, as is often the case. Test modification and development in different cultural settings are necessary, and for cross-cultural studies, equivalence and harmonization are essential for making meaningful comparisons. Ideally, a cross-national study should be designed from its inception to be reliable and valid at all sites at which it will be conducted, rather than letting different sites use different tests or different versions (other than translations) of the same tests. Tests should also be calibrated according to appropriate norms for each population.

2.6. Panel discussion points

- How different are RM and IRT? In IRT, one seeks to model the dataset, and if data do not fit the model, the model is changed to fit the data. In RM, the idea is to stay with the model and seek to understand why data do not fit the model. This is because the RM has very specific mathematical properties.
- Fixed instrument versus one that is inherently flexible: for example, if ADAS-Cog did not fit the RM model, it would be possible and permissible to adjust the task until it did. This raises the possibility of adding items that potentially increase sensitivity across a very broad spectrum, e.g., from individuals with AD to college students, and then focusing on a subset in a trial.
- No change versus inability to measure change in a particular domain: in trials where scores are not changing, for example, is the wrong scale being used? Is the domain really changing and not detected by the measure, or is there really no significant change?

3. Cognitive scales

Cognitive decline is a nonlinear continuum from normal through MCI to dementia. Where one sits on this continuum may therefore determine the sensitivity of a cognitive measure. In the ADCS MCI Trial, for example, placebo groups did not change much over 36 months according to both MMSE and ADAS-Cog/13 item version [36]. Those who were *APOE* ε4-positive did show greater change. Similarly, in the Alzheimer's Neuroimaging Initiative (ADNI), MCI participants showed little change in ADAS-Cog over 12 months, with a fair amount of variability from participant to participant. Very little annual change was evident on the CDR scale and in ADNI MCI participants (about 0.7). In ADNI, the Auditory Verbal Learning Test also indicated little movement [37], but with substantial variability among participants, which translated into very little differences among groups, whereas individual changes were apparent. Data from the National Alzheimer's Coordinating Center Uniform Data Set for amnestic MCI also showed that CDR and MMSE findings did not change much over 1 year [38]. However, the community-based Mayo Clinic Olmsted County Study of Aging showed that raw scores of measurements in several domains, observed over 12–15 months, resulted in a deterioration of memory in amnestic MCI participants [39]. In language, category fluency showed some loss, as did the Trail Making Test-B of executive function and the block design test of visuospatial memory. Normalized data, however, resulted in little change in four domains (memory, language, attention, and visuospatial memory) over 12–15 months.

These four different datasets suggest that in mildly impaired participants, there is very little movement in these scores, at least over the relative short term. Individual indices, and enriching populations with specific genotypes or imaging data, may offer a better picture. Computerized testing may also allow for mild changes in cognition to be captured, by measuring chronometric components of cognition.

The ADAS-Cog was originally an 11-item test [3]. There were several subsequent additions that may or may not be used [13,40]. The Neuropsychological Test Battery (NTB) is an attempt to address areas of function that are believed to be ignored or poorly measured by the ADAS-Cog [41]. The six commonly used tests (three of memory, and three of executive function) of the NTB have actually been used for decades, so there may be an opportunity to capitalize on more recent advances.

Current scales of cognition (e.g., ADAS-Cog and NTB) in clinical trials tend to use composite scores [41]. They avoid the statistical difficulties of dealing with multiple individual test scores. However, regulatory requirements do not always insist on composite scores. Measuring key areas of cognition separately may be more informative, and should be considered.

Factor-structure analysis can reveal some interesting characteristics of these composite tests, and may provide principled statistical support for the grouping of test measures into separate cognitive domains, such as episodic memory, working memory, and attention [42]. Combining measures in this way represents a helpful methodology for assessing drug effects in different cognitive domains, while precluding the necessity to analyze all outcome metrics from the selected tests. But there is one important caveat: a recent analysis suggested that a factor structure apparent at baseline may not hold up over time in participant groups. This is a major issue to be considered if factor analysis is to be used as a way of grouping outcome measures. One possibility is that with more extensive training before the test is used in a trial situation, the factor structure might be more stable. Averaging two or three measurements administered during a baseline period might induce more stability.

Ceiling and floor effects are also a concern with ADAS-Cog in clinical trials of MCI and mild AD. In MCI, for example, a majority of participants with amnestic MCI had scores of zero in 9 out of 11 ADAS-Cog tests [43]. This affected the assessment of practice effects. Participants continually demonstrating ceiling effects can make it more difficult to detect practice effects. Understanding practice effects is crucial for trials, because practice effects can obscure the true rate of decline in placebo groups. The NTB, on the other hand, is

more sensitive to participants with mild disease, and reduces the variability in neuropsychological test scores by using a prebaseline exposure to tests to induce stability.

The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) is another cognitive test that is relatively simple to administer, has minimal practice effects, and seems sensitive in MCI and mild-AD participants [44]. It measures immediate and delayed memory, attention, language, and visuospatial/constructional domains. It can be administered in 20 to 25 minutes, and was translated into and validated in multiple languages [45]. The RBANS has not been used in large, published AD or MCI clinical trials, but was shown to detect interventions in traumatic brain injury and schizophrenia trials [46,47]. This scale could be useful in AD clinical trials, but has not been compared directly with ADAS-Cog.

Some properties of RBANS that add to its suitability as an outcome measure, above and beyond tests such as ADAS-Cog and NTB, include:

- Population-based norming, with understandable scaling (mean, 100; SD, 15);
- Routine use in clinical diagnostic work, creating a link between clinical practice and clinical trials; and
- A global score composed of distinct neurocognitive domain scores (memory, attention, language, and visuospatial), allowing for post hoc exploration of effects.

Another approach to be considered for large trials entails computerized testing, which has many advantages, including easier data capture and validation, better standardization, enhanced reliability, automated scoring, and greater precision [48]. Computerized testing makes it possible to capture unique data that cannot be collected by other means, and that may be more sensitive to cognitive decline (e.g., reaction time, adaptive responses, speech files, and repetition effects). Computers can also assess practice effects, and may even allow them to be used in analyses. Computer-based testing, however, involves some obstacles. These include cost, technical maintenance, the need for specialized training, barriers to acceptance, and generalizability. Some computerized tests are proprietary and may contain limits on the extent to which specific tests may be modified for the specific needs of a trial.

Several computerized tests were described in the literature, and may be suitable for measuring cognition (Table 1) [49–59]. These tests vary considerably in terms of their domains, types of interface used, age groups, populations (e.g., AD or MCI), and number of volunteers on whom the tests were validated. In regard to how these may be developed in the future, efforts are underway to compare computerized tests with ''legacy'' tests, to use more automated assessments in-home, to take advantage of the unique aspects of computer-based testing to identify special properties of drugs, and to develop new testing paradigms.

In developing new or improved scales, the AD community may be able to capitalize on the experience of those working on diseases that affect cognition. In schizophrenia, cognitive function is a better predictor of functional performance than ''positive'' symptoms (e.g., hallucinations or paranoia), and this may be relevant to people with MCI who, although cognitively impaired, seem to have normal function [60]. Measuring performance in MCI participants may bring out subtleties that have not been appreciated. In fact, using the University of California San Diego (UCSD) Performance-Based Skills Assessment [61], a relationship was revealed between cognitive scores and function in MCI participants.

Table 1
Commonly used computerized cognitive tests

ANAM [49]
CANS-MCI [50]
CANTAB [51]
CNS Vital Signs [52]
CNTB [53]
COGDRAS-D [54]
CogState [55]
Cognitive and Stability Index (CSI) [56]
MCI Screen (MCIS) [57]
MicroCog [58]
Mindstream [59]

4. Functional scales

Loss of function is a key component of the diagnostic criteria for AD, and is often one of the primary endpoints in clinical trials, usually in regard to activities of daily living. Several functional outcome scales are currently being used in AD trials, including the DAD [7], ADCS-ADL [6], AD Functional Assessment Change Scale, Interview for Deterioration in Daily Living Activities in Dementia, and Progressive Deterioration Scale. Most of these are heavily influenced by informant reports. The most commonly used are the DAD and the ADCS-ADL, which address similar constructs, although their scoring may involve slightly different emphases.

Several clinical trials of compounds for AD used functional scales, and the most successful treatment for mild to moderate AD showed a stabilization of functional scales over 12 months [62]. In moderate to severe AD participants, an accelerated decline may occur in placebo groups, but floor effects have not been observed. Thus, the scales are valid, even at fairly low MMSE scores [62].

Current scales are known to exhibit different rates of change in AD. Moreover, they may not be ideal for very early disease stages. In fact, the suggestion was made to remove the functional decline requirement from the diagnostic criteria for prodromal AD and replace it with a biomarker requirement, such as cerebrospinal fluid measurements or brain imaging [63]. In terms of early diagnosis, the DAD scale cannot predict progression to AD, but the social and occupational domains of the Functional Rating Scale may have some predictive value [64]. Overall, however, existing scales may overemphasize instrumental and basic activities, and inadequately test social function.

The ADCS-ADL was a product of the ADCS instrument study [6]. Originally 45 ADLs were investigated and pretested to see if they were rigorous enough to withstand the criteria demanded of a routine test [65]. The ADCS-ADL eventually included a series of questions that covered 23 activities with graded responses, so that gradual transitions in performance could be detected. Total scores range from 0 to 78, and they correlate with dementia severity according to the MMSE. Score decline was measured over a 12-month follow-up period, and an ''inverted U'' effect was evident. This was also the case for ADAS-Cog, in that change is more gradual in the mildest and severest cases. The ADCS-ADL was implemented in galantamine and homocysteine studies, and seems to track functional decline sufficiently for use in clinical trials. It depends on information from an informant, however, and thus has an inherent degree of variability and subjectivity.

The ADCS-ADL was modified to suit MCI cohorts by selecting items that are most sensitive to change in the mildest AD patients. The modified scale, with 18 items and a score ranging from 0 to 54 [36], was used in the ADCS MCI trial. Initially, over 90% of participants (n = 769) had top scores on 10 of 18 items, whereas the other eight items seemed to be more sensitive to some sort of baseline impairment. Over the 36 months of the trial, 222 conversions to AD occurred. Those who converted showed further decline in those eight items that indicated impairment at baseline. Converters also showed decline on the remaining 10 items, whereas nonconverters showed no overall decline.

The concept of MCI originally held that mildly impaired people would exhibit cognitive decline, but undergo no change in daily function. However, minimally impaired independent activities of daily living (IADLs) were suggested for inclusion among the consensus criteria for amnestic MCI [66], notwithstanding that specific IADL performance in this group is not well-characterized. It is possible that some threshold of cognitive decline must be reached before declines in daily function become evident. However, the sensitivity of the instrument used to measure functional impairment will play a role in whether such problems are detected or not. Many of the available functional scales were designed to be relevant to dementia, but may not be applicable to MCI. For instance, individuals with MCI may still be able to shop independently, but have problems in remembering shopping items, finding the car in the parking lot, or efficiently planning a sequence of stops in a shopping routine. A number of studies using instruments sensitive to subtle problems in daily function did, in fact, show that MCI is often associated with mild functional impairments [67].

Importantly, recent evidence suggests that in people with MCI, mild problems in functional abilities at baseline are associated with more rapid disease progression [68] and a greater risk of conversion to dementia [69,70]. Moreover, particularly in individuals with low levels of education or from an ethnic minority background, measures of functional impairment may actually be a better predictor of subsequent disease progression than baseline measures of cognitive function. This may be the case because functional measures tend to be less affected by background or demographic factors than neuropsychological tests, which can be strongly influenced by factors such as education.

Given that functional changes in mildly affected individuals with MCI yield valuable information, which aspects of daily function should be measured in these people? In this group, IADLs seem much more affected than ADLs. In one recent study [71] of 18 ADCS-MCI-ADLs, 14 (particularly those with a strong memory component) differentiated individuals with MCI from control participants. Other approaches may involve applying cognitive models to better define real-world situations of cognition, breaking down component parts that make up IADLs, targeting processes that underlie aspects of activity performance, and indentifying early difficulties that precipitate more global dependencies in IADLs. Other approaches may involve the completion of daily journals or activity logs by informants or participants, or the use of computerized technology. But can a more sensitive test be developed to better detect subtle functional changes in mildly affected individuals?

The Everyday Cognition Scale (ECog) measures everyday manifestations of cognitive impairments in six different domains, including memory, language, visuospatial skills, planning, organization, and divided attention. It has been used in clinical trials. It differentiates clinical groups well (i.e., normal, MCI, and dementia), with a good range in each group and no appreciable ceiling or floor effects, particularly in MCI. The effect size between normal and MCI is about 0.5 to 1.0 standard deviations. Between MCI and dementia, the effect size doubles [67]. Different subtypes of MCI also show different patterns of performance on ECog [72]. To cite another advantage of ECog, it appears to be largely independent of educational level and ethnic status.

Other functional scales that may be useful in MCI populations include the Informant Questionnaire on Cognitive Decline in the Elderly [73], the Functional Capacities of Daily Living [74], the Functional Assessment Questionnaire [75], and the AD8 Dementia Screening Interview [76].

The relationship between cognition and functional impairment is crucial. Not only do diagnostic criteria rely on functional impairment, but if the cognitive components that underlie functional impairment could be identified, it may be possible to target interventions directly to those components, and to better predict who is at future risk for functional decline. To address this aspect, it is worthwhile to study cognitive and neuroimaging correlates of IADLs.

Some cross-sectional and longitudinal studies examined these relationships. Many cross-sectional studies indicate that executive function more accurately predicts IADLs than any other area of cognition [77]. Executive function was also much more predictive than demographic variables such as age, education, or health status [78]. Many studies also showed that memory is a significant predictor of IADLs, although there was significant variability in those studies

[79,80]. In terms of neuroimaging correlates, cross-sectional studies suggest that in some disorders, white-matter disease (subcortical hyperintensities) accounts for significant variance in ADLs [81–83].

Longitudinal studies support cross-sectional data on executive function, with many studies showing that executive function predicts decline in ADLs over several years in vascular dementia [84]. A recent study of several parameters, including executive function, memory, and neuroimaging correlates (hippocampal volume, white-matter hyperintensities, and cortical gray-matter volume), found that although memory and executive function were associated with baseline IADL scores, only executive function was independently associated with rate of change in IADLs [82,85]. In regard to neuroimaging correlates, both hippocampal and cortical gray-matter volumes were associated with baseline IADL, but only hippocampal volume was associated with IADL change. When psychometric and imaging analyses were combined, only executive function accounted for a significant portion of variance in future daily function.

4.1. Panel discussion points

- Informant versus participant reporting: which is better? Consensus indicates that when informant and participant scores diverge, the participant is more likely to convert to AD.
- Social function versus ADL: capacity for social interaction, confidence in social settings, and other criteria should also be considered.
- How relevant are these measures to drug development? Are subtle differences in different scoring systems relevant, given that acetylcholinesterase (AChE) inhibitors, for example, have such small effect sizes? Is it likely that we will want drugs with even less of an effect size, or one that treats only one domain? Are subtle functional measures likely to be useful from the perspective of drug development, or can industry ignore them?
- Better diagnosis: functional measures may lead to better and earlier diagnoses, which could be important for earlier treatment and earlier labeling.

5. Neuropsychiatric scales and global ratings in clinical trials

Although AD is seen primarily as a cognitive disorder, it can also cause neuropsychiatric symptoms that may, in fact, be the most treatable. In people with AD, the lifetime risk of developing some type of psychopathology is 100%. Psychiatric symptoms run the gamut and include psychosis, depression, agitation, aggression, and anxiety. Neuropsychiatric symptoms have a major impact on the person with AD and on caregiver quality of life. As well as leading to earlier institutionalization, they indicate more rapid cognitive decline [86]. ''Mild behavioral impairment'' may herald the conversion from MCI to AD.

Although there are consensus criteria for some of these syndromes in AD, such as psychosis and depression, these constructs have not been validated. Measuring psychiatric symptoms in people with AD depends on multiple factors, e.g., the target population (normal controls, MCI, or AD participants), their psychiatric status at baseline, and the objective of the study (to diagnose, characterize, or assess change). In intervention studies, it is important to have a behavioral hypothesis in mind, e.g., is psychopathology to be relieved, prevented, or delayed? Although regulatory trials tend to focus on statistically significant between-group differences for approval, there is also a need to better inform clinical practice and assess clinical significance in terms of incorporating global scales as adjunct measures, including responders' analyses, and considering a broader range of alternative trial designs (e.g., survival to switch/discontinuation), and including better measures of caregiver's impressions.

For AD, focused as well as more general scales are in use. The Cornell Scale for Depression in Dementia and the Cohen-Mansfield Agitation Inventory are examples of the former. General scales include the Brief Psychiatric Rating Scale [87], the Revised Memory and Behavior Problems Checklist [88], the Multidimensional Observation Scale for Elderly Subjects [89], the Behavior Rating Scale for Dementia (BRSD) of the Consortium to Establish a Registry for Alzheimer's Disease [90,91], the Psychogeriatric Dependency Rating Scale [92], and the NPI [9]. The most comprehensive are the Behavior Rating Scale for Dementia and the NPI. The simplest is the Psychogeriatric Dependency Rating Scale. The NPI is most widely used at present, despite its major limitation of not accounting for rater judgment.

The AD trials in which neuropsychiatric symptoms were the main focus include studies of antipsychotic drugs (e.g., carbamazepine [93], risperidone [94], and quetiapine [95]), drugs with putative neuroprotective effects (e.g., valproate) [96], and cholinergics (e.g., galantamine) [97]. Although several scales were used to demonstrate efficacy in clinical trials, the selection of a neuropsychiatric scale depends on what one wants to measure. Is the issue a characterization of behavior at a given point or points in time, a need to establish the presence or delay of a particular sign or symptom, or the need to assess change after an intervention? Critical methodological issues will influence choice, including target population, behavioral hypothesis (e.g., symptom reduction or secondary prevention), feasibility considerations, whether subjective or caregiver distress is to be measured, and sources of information (Table 2). Reduction in scale scores after an intervention may reach statistical significance, but still leave open the question of clinical significance. Thus some investigators may rely on global clinical impressions to address clinical meaningfulness. The literature provides many examples of different uses and approaches. For instance, effectiveness outcomes were used in the Clinical Antipsychotic Trial of Intervention Effectiveness study for Alzheimer's Disease (CATIE-AD) trial as another means of addressing methodological issues [98].

Table 2
Methodological issues to consider in measurement of behavior

What is the target population?
Normal MCI, dementia, specific dementia diagnosis (severity will matter)?
With/without psychopathology at baseline?
Setting?
What is the objective of the study?
Characterization of behavior
Establish presence/absence of *any* behavioral domain
Establish presence/absence of *specific* behavioral domains
Assess change after intervention
For interventional study: what is the behavioral hypothesis?
Relief of psychopathology once present
Secondary prevention/delayed onset of psychopathology
Primary prevention
Does the scale address domains of interest?
What are its psychometric properties and extent of data in populations of interest?
Does it rate frequency?
Does it rate severity (although this is subjective)?
Does it rate degree of disruptiveness/distress (subjective)?
What is the relevant time window?
What are the sources of information: observation, interview, informant?
Are informant qualifications specified?
What are the rater qualifications?
Are there adequate training materials?
What are the feasibility issues?
How reliable is the scale?
How sensitive is the scale?
Are there validity data?
What is the study duration?
What is the frequency of visits?
Are behavioral data available by telephone, internet, survey?
Is it clear how to analyze?
For example, NPI: total, item-by-item, reduction in symptom present, remission of symptom present, emerging symptom.

What about prevention studies? A secondary outcome in a galantamine trial suggested that, according to the NPI, the emergence of behavioral symptoms might be delayed in participants taking drug versus placebo, suggesting that secondary prevention for psychopathology is possible [99]. The issue becomes whether one can reliably measure emerging or incident psychopathology. Currently, the ADCS Valproate Neuroprotection Trial is asking just that question, using as an endpoint a threshold in a modified version of the NPI that must be reached and maintained over a period of weeks, in association with a clinician's assessment that the NPI change is clinically significant during that time.

Scales must be chosen to suit the domain to be measured. One may want to measure the presence or absence of *any* behavioral domain, or the presence or absence of a *specific* domain. One example of neuropsychiatric measurement in a community setting is the Cache County Study in Utah of people aged 65 years and older [100]. Based on the NPI scale, over a 5-year period, the prevalence of neuropsychiatric symptoms in this population increased, such that 90% of the population had at least one symptom, and the symptoms were worsening. Using a cutoff of 10 on the NPI as a baseline, for example, 70% of people had an NPI greater than 10 after ≥ 3 years.

Why have drugs for neuropsychiatric symptoms in AD not fared so well? Perhaps the wrong symptoms are being targeted. Better measures may help address this problem, and measures that can function in both broad and narrow spectra would be ideal. One possibility a clinician-rated NPI measure (the NPI-C) [101]. This is being developed by a 14-site international collaborative. It preserves and expands some key domains, and also adds new ones, to provide depth.

5.1. Global assessments

Global assessments have long been considered the ultimate test of a drug's antidementia effects, and though that may be true, not all clinicians agree with that sentiment. Nevertheless, global assessments play a crucial role in the diagnosis and staging of dementia, in assessing disease progression, and as outcome measures in clinical trials. Most global assessments require a skilled clinician, which can be both an advantage and a disadvantage (the tests can be time-consuming). Most are also based on interview with the person being diagnosed and/or a knowledgeable informant, and the critical point is an ability to judge change in function over time. The scales also have the benefit of assessing multiple domains, i.e., not just cognition and function, but behavior as well, and they were developed to be independent of other data, including neuropsychological measurements.

The main advantage of global scales involves their assessment of intra-individual change, i.e., how a particular individual has changed relative to previous abilities. Because the scales use individual participants as their own controls, they are not influenced by comorbidities, experiences, level of education, or cultural differences. They also avoid floor-and-ceiling effects. However, global scales pose some disadvantages. The interview can be time-consuming, informants can be unreliable or even unavailable, and the scales require judgment.

Global scales can be divided into those that are standardized and semistructured, and those that are individualized. The latter include the Clinician Interview-Based Impression of Change (CIBIC) and the CIBIC-Plus [102,103], which includes an informant interview. The CIBIC is the based on the underlying idea that if a clinician can detect a change, then it must be clinically meaningful. The individualized outcome measures have good face validity. With the CIBIC, there is not much reliability between physicians, and the test has proven hard to standardize. The ADCS-CGIC has more structure than the CIBIC [8]. It also has good test-retest reliability and predictive validity.

The Clinical Dementia Rating (CDR) is the most structured, and has good interrater reliability [5]. It is sensitive to even small degrees of clinically meaningful change, with changes paralleling psychometric scores. The CDR can be turned into a slightly more quantitative measure by using the ''sum of boxes'' (SB) approach, which basically turns the scale into a 0-to-18 scale by rating each of six domains as 0, 0.5, 1, 2, or 3, with higher scores indicating more

impairment [104]. The CDR-SB is very sensitive and is predicted to be equal in power to many neuropsychological measures, but at smaller sample sizes [105]. For example, the CDR-SB showed a significant deleterious effect in the ADCS Estrogen Trial, whereas the MMSE, ADCS-CGIC, and ADAS-Cog all showed trends [106]. The CDR, which was translated into several languages, is being applied in multicenter trials such as ADNI and the ADCS, and was also incorporated into the National Alzheimer's Coordinating Center Uniform Data Set package. This database will follow people longitudinally, and the CDR is part of the analysis.

Goal attainment scaling (GAS) is another individualized global measure [107,108]. A goal could be the adjustment of some behavior, such as repetitive questioning. In some cases, reducing repetitive questioning by 20% might be seen as a big improvement, but if the caregiver or person affected by AD thinks it is not good enough, then the individualized measure takes that into account. A GAS score is based on a mathematically calculated increase or decrease from a baseline score, set at 50. Scores higher than 50 indicate improvement in one or more goals, and a score of less than 50 indicates deterioration. Goal attainment scaling can pick up longitudinal changes that are not captured by the ADAS-Cog or CIBIC-Plus.

Goal attainment is particularly important to patients and caregivers, and very often a group of goals is connected, i.e., when one behavior improves so does another, or when an individual reports one behavior, another is usually also reported. These connections can be graphed, and may provide valuable information about an individual's behaviors and what is important to them. A recent web-based interview-assessment study showed that repetitive questioning, language difficulties (such as linguistic expression and word-finding difficulties), social interactions/withdrawal, misplacing objects, and telephone use were among the most connected behaviors [109]. For example, those symptoms that are most highly connected often improve or deteriorate in concert.

Recent trial data show that GAS scores can detect improvements in people taking galantamine [110]. For repetitive questioning, 70% of participants showed improvement over 16 weeks, versus only 27% of participants on placebo, and many more of the controls showed deterioration (30% versus 10%).

There may be ways to improve on current global scales. They could be modified to include assessments of people's quality of life and ''behavioral competence,'' and the measurement properties of current scales could be better defined. Furthermore, several scales in current use make standardization difficult. Adopting a single instrument would aid in comparisons of antidementia drugs.

Many global studies rely on patient-reported outcomes (PROs), which, from a therapeutic perspective, are a direct measure of treatment benefit. This is one reason for the growing interest in PROs recently. However, in AD as in other dementias, the participant may be a poor informant, and as a result, the PRO may not be accurate. The FDA has released a draft guidance document for industry on the use of PROs to support labeling claims [111]. Label claims should be supported by ''substantial evidence'' based on adequate and well-controlled investigations. Those investigations should have appropriate methods for assessments of outcome that are well-defined and reliable. The most important concept in the draft guidance states that measurements should begin with the goal in mind. After a treatment-benefit claim is identified, an appropriate instrument can be developed to justify that claim.

The conceptual framework of an instrument is crucial. This conceptual framework is not a complicated entity, but a simple description of how each item relates to and contributes to the score and to other items in the instrument. The concept can be improved and validated over time, until the instrument is capable of measuring the intended concept. Crucial to this are content validity (i.e., the items and domains measure the intended concepts as outlined in the concept framework), reliability, construct validity, and an ability to detect change. Validity is not absolute, because it often depends on the population under study, disease severity, and clinical design setting.

To help with the development of instruments, one can reference the FDA Target Product Profile (TPP), a guidance document for both industry and review staff [112]. The TPP, or development plan summary, is designed to smooth the process of developing instruments by fostering communication between the regulatory agency and product sponsor early in the development process. A well-developed TPP can facilitate communication by providing the proper context for discussion in terms of labeling goals.

The FDA and industry are also working on a consortium approach to outcome-measures development. The PROLabels database (http://www.mapi-prolabels.org/) was developed to provide easy access to the PROs included in the approved labeling of products in Europe and the United States [113]. The PROLabels database is a unique online tool for collecting information on medical products that have received a PRO labeling claim from the FDA and/or European Medicines Agency (EMEA). It was codeveloped by the Mapi Research Trust (Lyon, France) and Mapi Values (Boston, MA) in 2006. The PROLabels updated database of 2008 provides a clearer picture of the use of PROs to assess patients' treatment benefits. In addition, it facilitates comparisons between United States and European regulatory agencies.

5.2. *Panel discussion points*

- Goal attainment scaling from a regulatory perspective: because GAS is individualized, will it satisfy regulatory agencies? There is no regulatory experience with those scales, so their appropriateness needs to be addressed. In theory, GAS appears to be a good idea.
- Value of informants in global assessment: the CIBIC was designed to avoid the need for an informant, but that is hard to achieve in clinical practice. That is why

the CIBIC-Plus was introduced. Informants have a tendency to constrain the clinician's assessment of the person affected by AD.

- Behavioral symptoms: to what extent are these symptoms different in AD than in non-AD settings, and how can one decide whether a specific claim for AD-related behavioral symptoms is appropriate? This issue may be resolved on a case-by-case basis, and may rely on how much is known about the pathology that leads to the symptom. Dementia in Parkinson's disease (PD), for example, may appear similar to AD, but if a drug can treat it by attacking the underlying pathology of PD, then there may be justification for a claim of PD-specific dementia. According to the general regulatory perspective, to grant a claim for a specific symptom that occurs in a larger clinical context, there must be something specific about that symptom.
- Are behavioral symptoms the core features of a disease? That question is debatable, because not every person affected by AD manifests behavioral symptoms. Behavioral symptoms are not used to define or diagnose the disease.

6. Cooperative analysis of data

Academia, small and large industry, and government agencies such as the National Institutes of Health have held their traditional niches in AD drug development. Whereas academia mostly focuses on pathophysiology and genetics, industry mostly focuses on the development of lead compounds, biomarker studies, and proof-of-concept and efficacy trials. Those traditional divisions have become blurred, however, with an accompanying opportunity for greater collaboration. The biomarker initiative in the late 1990s, for example, led to the development of positron emission tomography and single photon emission computed tomography (SPECT) ligands, and more recently to ADNI, a collaboration between industry and academia [114]. Industry and the National Institute on Aging have primarily funded ADNI, with an initial focus on cerebrospinal fluid biomarkers and subsequently Pittsburgh compound B (PIB) imaging, with additional funding from the Alzheimer's Association and General Electric [115–117]. A major question at present concerns whether there will be a consensus on cognitive and functional scales in time for a renewal of ADNI [118].

In the ADNI Study, the ADAS-Cog, MMSE, and CDR-SB all showed baseline differences among normal control, MCI, and AD groups, with clear trends toward decline that differed across groups. These trends are also evident in composite scores. The data suggest that to detect a modest difference in rate of decline (25%) in AD, a clinical trial of 400 to 800 participants would be needed in each arm of the study. In an MCI trial, additional participants would be required. Memory tests fare little better in this regard. This finding suggests that better and faster ways are needed to detect AD progression or reduction in progression, and this is a key motivation for ADNI. The real challenge will be to turn these results into practical measures of change that are highly informative, and that add to the standard measures.

Another recently initiated major collaboration was a private foundation-funded meeting to address the problem of placebo data in MCI and AD trials. There has been some concern in the AD research community that cognitive measures in clinical trials are inadequate, that decline among placebo subjects is not being captured, and that participants may somehow be healthier, more diverse, or even milder than before. With the advent of clinical trials for participants with MCI, there is clearly a need to improve these measures.

7. Regulatory issues

7.1. Validation of new scales

An ability to demonstrate the validity of a scale for use in a primary outcome measure in a registration trial is crucial. In the United States, the FDA examines new scale proposals a little differently than in the past, and would probably investigate any new scale more closely than before. Correlation with traditional scales may also be an important requisite. There is an increasing emphasis on a comprehensive validation process for psychometric scales, primarily for cognitive measures.

As for noncognitive measures, the FDA considers functional and global scales to serve the same purpose. Global scales perhaps offer a cruder assessment, and so the FDA probably relies more on face validity for those, but the revelation that the CDR-SB can actually be more sensitive than cognitive measures raises the question of whether it would suffice as a primary outcome. The FDA would have to scrutinize that possibility carefully.

On cognitive scales, and in particular effects on specific cognitive measures, it is hard to predict what kind of claim could be made if a drug affected only executive function, for example. The FDA has not been faced with that scenario in practice, but the possibility is worthy of consideration. In terms of the bottom line, the FDA would require more systematic validation of a scale than they did in the past. Obtaining that validation would likely be a long process.

The European perspective is similar in that if a scale is in use, the track record probably speaks for itself, but if it is newly developed, then proper validation is required. An important issue involves whether the scale is developed for the purpose of measuring an effect of a specific drug, or to measure a change in some clinically relevant domain. The scale should be developed to measure those domains first, and then used to measure the effect of a drug. In this regard, a scale must have face validity, construct validity, and reliability.

7.2. Acceptable primary cognitive endpoints

The NTB has found more widespread use in clinical trials. The FDA has informed at least one sponsor that the NTB is

acceptable. Formal validation has not been achieved, but the test has a certain amount of face validity.

In the new European draft guidelines, the NTB is recognized, and is more or less ready to be accepted as a primary outcome, although there is some reticence, mostly because it is viewed as a scale made to fit a certain type of drug. There are also at least four different versions of the NTB, which make acceptance more complicated.

7.3. Acceptability of scale variations

Situations may arise where different versions of the same scale are used to obtain data. From the FDA's perspective, this is already the case to a certain degree, e.g., in international studies where scales have to be translated: by definition, scales in different languages are different versions. Furthermore, different forms or different specifics of scales are often used to avoid practice effects. There does not appear to be an a priori reason to prohibit this, if the versions are valid and can be combined.

The European Union perspective is different on this issue. They will treat different versions of scales as different scales, e.g., the four different versions of the NTB.

7.4. Coadministering scales

If the FDA deems the NTB acceptable, then it would not require an additional scale to be run in parallel, but the FDA certainly recognizes the value in having both.

In Europe, for the foreseeable future, the EMEA will likely require, or at least look favorably on, sponsors who run the NTB in parallel with the ADAS-Cog, so that some clinical relevance can be attributed to the NTB.

7.5. Nonprimary endpoints and labeling

The FDA has a fairly standard position on nonprimary endpoints and labeling. For a primary outcome, the measurement must be of a different domain than the primary endpoint. For example, the primary outcome could be cognition, and the secondary outcome could be global. A prospective labeling plan acceptable to the FDA is needed, as is an adequate statistical plan to deal with multiple comparisons and other statistical issues.

In the European Union, only the primary outcome measure is described in the statistical process control (SPC) plan. Moreover, in Europe, advertisements can be geared directly only to the payers and not to consumers, and cost effectiveness is of concern, rather than the SPC.

7.6. Midtrial design or analysis changes

Because the field is rapidly evolving and dementia trials can be lengthy, sponsors may consider midtrial modifications in design or analysis. From the FDA's perspective, this seems unnecessarily complex and unlikely to be approved, although it remains within the realm of possibility. Midtrial modifications would depend entirely on case-by-case specifics.

In Europe, such modifications may depend on the extent to which circumstances have changed. Adaptive designs are not well-considered, but are also not completely impossible. However, a change that affects the outcome would not be well-favored. Analytical changes made before a trial is unblinded, for example, would be tolerable, and in general it is best if any potential change is built into the plan in advance.

7.7. Survival analysis and dichotomized endpoints

For the FDA, survival endpoints are typically well-accepted, and outcomes timed to MCI or AD incidence are not a problem. Dichotomized endpoints are not as easy to deal with, and it may not be clear what they even mean. They may be acceptable, or problematic. There is no a priori objection, but acceptability depends on the details.

In Europe, survival analysis is well-accepted and, in fact, is seen as having good face validity. However, some weight must be given to the difference in time that survival analysis reveals. For example, a drug that delays conversion by 2 months, assuming that conversion could be exactly defined, is not likely to be approved. Dichotomized endpoints are also possible, though clinical relevance is the most fundamental issue.

7.8. Consensus in identification of suitable outcome measures for early-stage trials

Clearly, outcome measures for early-stage, disease-modification trials will be needed. Identification of suitable measures will undoubtedly require cooperation among the pharmaceutical industry, academia, and regulatory agencies. Some changes (e.g., the NTB) can be performed at a local level, with the agreement of the FDA and the sponsor, but other changes, such as developing a rationale for studying disease modification, will need the cooperation of all stakeholders. In Europe, newly available funding could be used for just such a process.

7.9. Panel discussion points

- Different effects in primary and secondary outcomes: how would they affect approval? How would the FDA view no drug effect in the primary outcome measure (i.e., ADAS-Cog) and a small drug effect on a secondary outcome measure (i.e., NTB)? The FDA has never placed an effect-size requirement on the ADAS-Cog, because it was assumed that the effects would be picked up by global scales, but the problem may be that the global scale is really just another ADAS-Cog. There is no way of knowing that, but if it were true, it would be a problem.
- Different versions: how different is different? There are different versions of the ADAS-Cog, for example, and different language versions. In this regard, the devil is in the details. One could call two different scales by

the same name, but that would not make them different versions of the same scale. Moreover, if some items on a scale can be discarded, then there must be redundant tests, but if the tests are not redundant and they are measuring different things, then are they not different scales?

- What is the way to approach the challenge of collaborating on devising scales? The ADNI is moving into MCI and milder aspects of the MCI spectrum, and will need to modify instruments or pick up the NTB. Engaging people with computer battery expertise may be productive. Either way, there must be a strategy to selecting a scale so that it is measures what it is intended to measure. Whether the best measure turns out to be computer-based or pen-and-paper is then secondary.
- Clinical relevance: As the field moves toward making a diagnosis in the MCI range of cognitive impairment, a drug effect may improve memory to a small degree but have a larger positive effect on a biomarker. If a drug does not significantly benefit patient function or provide clinical benefit, is it clinically relevant? Does such an outcome have any meaning? This is a one of the current challenges, and the literature is limited in its ability to clarify this point.

8. Conclusions

This review has covered a wide series of important issues regarding outcome measures in AD clinical trials. As the length of clinical trials increases and individuals with milder deficits are included in trials, it becomes clearer that our standard cognitive measures may not be sufficiently sensitive for measuring current status or changes over time. Several possible solutions were discussed, e.g., adding more items to the ADAS-Cog to increase its sensitivity to subtle deficit, or evaluations of more focused domains, or adapting global summary scales. The need for more sensitivity has to be balanced against the FDA's desire for efficacy, as determined by a clinically meaningful change, as opposed to simply statistically significant improvement on a scale. In addition, it is worth considering which problems with current measures will disappear when a drug that has a major impact on AD pathology is developed, and which issues need to be addressed to detect the efficacy of a new therapeutic intervention with a strong effect size. Similar considerations apply to measures of function, psychopathology, and global clinical-efficacy ratings. Progress in basic research also presents new challenges. For example, careful consideration must be given to how biomarkers might complement assessment scales in clinical trials. The Research Roundtable discussion provided an excellent summary of promising trends and ideas for improving outcome measures in clinical trials. It pointed out areas of need and provided direction for coordinated efforts to improve the scales used in trials, and highlighted some notable progress in this direction. The Research Roundtable's findings could provide a focal point for new consortia that attempt to share and use data that have already been collected to address these important issues.

Acknowledgments

The Scales as Outcome Measures for Alzheimer's Disease Subcommittee of the Alzheimer's Association Research Roundtable thanks all the speakers who participated in this meeting, including Lon Schneider, University of Southern California; Steven Ferris, New York University; Jeremy Hobart, Peninsula Medical School; Ronald Thomas, University of California at San Diego; Mary Sano, Mount Sinai School of Medicine; Mary Ganguli, University of Pittsburgh; Ronald Petersen, Mayo Clinic; John Harrison, i3 Research; Christopher Randolph, Loyola University Medical Center; Jeffrey Kaye, Oregon Health and Science University; Terry Goldberg, Litwin-Zucker Alzheimer's Research Center; Howard Feldman, University of British Columbia and Vancouver Coastal Health; Douglas Galasko, University of California at San Diego; Sarah Farias, University of California; Deborah Cahn-Weiner, University of California, San Francisco; Pierre Tariot, Banner Alzheimer Institute; Constantine Lyketsos, Johns Hopkins Bayview Medical Center; John Morris, Washington University School of Medicine; Kenneth Rockwood, Dalhousie University; Laurie Burke, Food and Drug Administration; William Potter, Merck & Co.; Laurel Beckett, University of California, Davis; Marilyn Albert, Johns Hopkins University; Russell Katz, Food and Drug Administration; and Cristina Sampaio, Faculdade de Medicina de Lisboa. We also thank Tom Fagan, science writer, and DeLois Powe of the Alzheimer's Association for their assistance in preparing the manuscript.

References

[1] Mckhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939–44.

[2] Davis KL, Mohs RC, David BM, Horvath TB, Greenwald BS, Rosen WG, et al. Oral physostigmine in Alzheimer's disease. Psychopharmacol Bull 1983;19:451–3.

[3] Mohs RC, Cohen L. Alzheimer's Disease Assessment Scale (ADAS). Psychopharmacol Bull. 1998;24:627–8.

[4] Folstein MF, Folstein SE, McHugh PR. Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician. 1975. J Psychiatr Res 1975;12:189–98.

[5] Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry 1982; 140:566–72.

[6] Galasko D, Bennett D, Sano M, Ernesto C, Thomas R, Grundman M, et al. An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. The Alzheimer's Disease Cooperative Study. Alzheimer Dis Assoc Disord 1997;11(Suppl. 2):S33–9.

[7] Gelinas I, Gautheir L, McIntyre M, Gauthier S. Development of a functional measure for persons with Alzheimer's Disease: the Disability Assessment for Dementia. Am J Occup Ther 1998; 53:471–81.

[8] Schneider LS, Olin JT, Doody RS, Clark CM, Mooris JC, Reisberg B, et al. Validity and reliability of the Alzheimer's Disease Cooperative Study—Clinical Global Impression of Change (ACDS-CGIC). Alzheimer Dis Assoc Disord 1997;11(Suppl. 2):S22–32.

[9] Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. Neurology 1994; 44:2308–14.
[10] Schneider LS. Prevention therapeutics of dementia. Alzheimers Dement 2008;4(Suppl. 1):S122–30.
[11] Irizarry MC, Webb DJ, Bains C, Barrett SJ, Lai RY, Laroche JP, et al. Predictors of placebo group decline in the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) in 24 week clinical trials of Alzheimer's disease. J Alzheimers Dis 2008;14:301–11.
[12] American Psychological Association, National Council on Measurement in Education, American Educational Research Association. Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association; 1999.
[13] Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. Alzheimer Dis Assoc Disord 1997;11(Suppl.):S21–31.
[14] Saxton J, McGonigle-Gibson KL, Swihart AA, Boller F. The Severe Impairment Battery (SIB) Manual. Pittsburgh, PA: Alzheimer's Disease Research Center; 1993.
[15] Auer SR, Sclan SC, Yaffee RA, Reisberg B. The neglected half of Alzhneimer disease: cognitive and functional concomitants of severe dementia. J Am Geriatr Soc 1994;42:1266–72.
[16] Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Raing sales as outcome measuere for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007;6:1094–105.
[17] Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley; 1968.
[18] Hambelton RK, Swaminathan H. Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff; 1985.
[19] Write BD. Solving measurement problems with the Rasch model. J Educ Meas 1977;14:97–116.
[20] Write BD, Stone MH. Best Test Design: Rasch Measurement. Chicago: MESA; 1979.
[21] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Education Research; 1960.
[22] Andrich D. Rasch Models for Measurement. Beverley Hills, CA: Sage; 1988.
[23] Andrich DA. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.
[24] Sano M, Ernesto C, Thomas RG, Klauber MR, Schafer K, Grundman M, et al. A controlled trial of selegiline, alpha-tocopherol, or both as treatement for Alzheimer's disease. The Alzheimer's Disease Cooperative Study. N Engl J Med 1997;336:1216–22.
[25] Mulnard RA, Cotman CE, Kawas C, van Dyck CH, Sano M, Doody R, et al. Estrogen replacement therapy for treatment of mild to moderate Alzheimer disease: a randomized controlled trial. Alzheimer's Disease Cooperative Study. JAMA 2000;283:1007–15.
[26] Petersen RC, Thomas RD, Grundman M, Bennett D, Doody R, Ferris S, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. N Engl J Med 2005;352:2379–88.
[27] Aisen PS, Schafer KA, Grundman M, Pfeiffer E, Sano M, Davis KL, et al. Effects of rofecoxib or naproxen vs placebo on Alzheimer disease progression: a randomized controlled trial. JAMA 2003; 289:2819–26.
[28] Aisen PS, Schneider LS, Sano M, Diaz-Arrastia R, van Dyck CH, Weiner MF, et al. High-dose B vitamin supplementation and cognitive decline in Alzheimer disease: a randomized controlled trial. Alzheimer Disease Cooperative Study. JAMA 2008;300:1774–83.
[29] Thomas RB, Berg JD, Sano M, Thal L. Analysis of longitudinal data in an Alzheimer's disease clinical trial. Stat Med 2000;19:1433–40.
[30] Dawson JD, Lagakos SW. Size and power of two-sample tests of repeated measures data. Biometrics 1993;49:1022–32.
[31] Sano M, Zhu CW, Whitehouse PJ, Edland S, Jin S, Ernstrom K, et al. ADCS Prevention Instrument Project: pharamcoeconomics: assessing health-related resource use among healthy elderly. Alzheimer Dis Assoc Disord 2006;20(Suppl. 3):S191–202.
[32] Ferris SH, Aisen PS, Cummings J, Galasko D, Salmon DP, Schneider L, et al. ADCS Prevention Instrument Project: overview and initial results. Alzheimer Dis Assoc Disord 2006;20(Suppl. 3): S109–23.
[33] Ferri CP, Prince M, Brayne C, Brodaty H, Fratigliono L, Ganguli M, et al. Global prevalence of dementia: a Delphi Consensus Study. Lancet 2005;366:2112–7.
[34] Pandav R, Fillenbaum G, Ratcliff G, Dodge H, Ganguli M. Sensitivity and specificity of cognitive and functional screening instruments for Dementia: the Indo-US Dementia Epidemiology Study. J Am Geriatr Soc 2002;50:554–61.
[35] Baiyewu O, Unverzagt FW, Lane KA, Gureje O, Ogunniyi A, Musick B, et al. The Stick Design Test: a new measure of visuoconstructional ability. J Int Neuropsychol Soc 2005;11:598–605.
[36] Grundman M, Petersen RC, Ferris S, Thomas RG, Aisen PS, Bennet DA, et al. Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials. Arch Neurol 2004;61:59–66.
[37] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement 2005;1:55–66.
[38] Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, et al. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Dis Assoc Disord 2007; 21:249–58.
[39] Geda YE, Roberts RO, Knopman DS, Petersen RC, Christianson TJ, Pankratz VS, et al. Prevalence of neuropsychiatric symptoms in mild cognitive impairment and normal cognitive aging: population-based study. Arch Gen Psychiatry 2008;65:1193–8.
[40] Mohs RC, Cohen L. Alzheimer's Disease Assessment Scale (ADAS). Psychopharmacol Bull 1998;24:627–8.
[41] Harrison J, Minassian SL, Jenkins L, Black RS, Koller M, Grundman MPH. The NTB: a Neuropsychological Test Battery for use in Alzheimer's disease clinical trials. Arch Neurol 2007; 64:1323–9.
[42] Lewis MS, Maruff PT, Silbert BS. Examination of the use of cognitive domains in postoperative cognitive dysfunction after coronary artery bypass graft surgery. Ann Thorac Surg 2005;80:910–6.
[43] Winblad B, Gauthier S, Scinto L, Feldman H, Wilcock GK, Truyen L, et al. Safety and efficacy of galantamine in subjects with mild cognitive impairment. Neurology 2008;70:2024–35.
[44] Randolph C, Tierney MC, Mohr E, Chase TN. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary clinical validity. J Clin Exp Neuropsychol 1998; 20:310–9.
[45] Yamashima T, Yoshida M, Kumahashi K, Matsui M, Koshino Y, Higashima M, et al. The Japanese version of RBANS (Repeatable Battery for the Assessment of Neuropsychological Status) [in Japanese]. No To Shinkei 2002;54:463–71.
[46] McKay C, Wertheimer JC, Fichtenberg NL, Casey JE. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): clinical utility in a traumatic brain injury sample. Clin Neuropsychol 2008;22:228–41.
[47] Delle Chiaie R, Salviati M, Fiorentini S, Biondi M. Add-on mirtazapine enhances effects on cognition in schizophrenic patients under stabilized treatment with clozapine. Exp Clin Psychopharamacol 2007;15:563–8.
[48] Wild K, Howieson D, Webbe F, Seelye A, Kaye J. Status of computerized cognitive testing in aging: a systematic review. Alzheimers Dement 2008;4:428–37.
[49] Jones WP, Loe SA, Krach KS, Rager RY, Jones HM. Automated Neuropsychological Assessment Metrics (ANAM) and Woodcock-Johnson III Test of Cognitive Ability: a concurrent validity study. Clin Neuropsychol 2008;22:305–20.

[50] Tornatore JB, Hill E, Laboff J, McGann ME. Self-administered screening for mild cognitive impairment: initial validation of a computerized test battery. J Neuropsychiatry Clin Neurosci 2005; 17:98–105.
[51] Robbins TW, James M, Owen AM, Sahakian BJ, Mcinnes L, Rabbitt P. Cambridge Neuropsychological Test Automated Battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers. Dementia 1994;5:266–81.
[52] Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. Arch Clin Neuropsychol 2006;21:623–43.
[53] Veroff AE, Cutler NR, Sramek JJ, Prior PL, Mickelson W, Hartman JK. A new assessment tool for neuropsychopharmacologic research: the computerized Neuropsychological Test Battery. J Geriatr Psychiatry Neurol 1991;4:211–7.
[54] Simpson PM, Surmon DJ, Wesnes KA, Wilcock GK. The cognitive drug research computerized assessment system for demented patients: a validation study. Int J Geriatr Psychiatry 2004;6:95–102.
[55] Maruff P, Thomas E, Cysique L, Brew B, Collie A, Snyder P, et al. Validity of the CogState Brief Battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. Arch Clin Neuropsychol 2009 March 25;[Epublication ahead of print].
[56] Erlanger DM, Feldman DJ, Kaplan D, Theodoracopulos A. Development and validation of the cognitive stability index, a web-based protocol for monitoring change in cognitive function. Arch Clin Neuropsychol 2000;15:693–4.
[57] Trenkle DL, Shankle WR, Azen SP. Detecting cognitive impairment in primary care: performance assessment of three screening instruments. J Alzheimers Dis 2007;11:323–35.
[58] Johnson JA, Rust JO. Correlational analysis of MicroCog: assessment of cognitive functioning with the Wechsler Adult Intelligence Scale-III for a clinical sample of veterans. Psychol Rep 2003;93:1261–6.
[59] Dwolatzky T, Whitehead V, Doniger GM, Simon ES, Schweiger A. Validity of a novel computerized cognitive battery for mild cognitive impairment. BMC Geriatr 2003;3:4.
[60] Bowie CR, Reichenberg A, Patterson TL, Heaton RK, Harvey PD. Determinants of real-world functional performance in schizophrenia subjects: correlations with cognition, functional capacity, and symptoms. Am J Psychiatry 2006;163:418–25.
[61] Patterson TL, Goldman S, McKibbin CL, Hughs T, Jeste DV. UCSD performance based skills assessment: development of a new measure of everyday functioning for severely mentally ill adults. Schizophr Bull 2001;27:235–45.
[62] Feldman H, Gauthier S, Hecker J, Vellas B, Subbiah P, Whalen E. Donepezil MSAD Study Investigators Group. Neurology 2001; 57:613–20.
[63] DuBois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 2007; 6:734–46.
[64] Hsiung GR, Alipour S, Jacova C, Grand J, Gauthier S, Black S, et al. Transition from cognitively impaired not demented to Alzheimer's disease: an analysis of changes in functional abilities in a dementia clinic cohort. Dement Geriatr Cogn Disord 2008;25:483–90.
[65] Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA 1963;185:914–9.
[66] Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, et al. Mild cognitive impairment—beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. J Intern Med 2004;256:240–6.
[67] Farias S, Mungas D, Reed B, Harvey D, Cahn-Weiner D, DeCarli C. MCI is associated with deficits in everyday functioning. Alzheimer Dis Assoc Disord 2006;20:217–23.
[68] Purser J, Fillenbaum G, Wallace R. Memory complaint is not necessary for diagnosis of mild cognitive impairment and does not predict 10-year trajectories of functional disability, word recall, or short portable mental status questionnaire limitations. J Am Geriatr Soc 2006;54:335–8.
[69] Peres K, Chrysostome V, Fabrigoule C, Orgogozo J, Dartigues J, Bargerger-Gateau P. Restriction in complex activities of daily living in MCI: impact on outcome. Neurology 2006;67:461–6.
[70] Daly E, Zaitchik D, Copeland M, Schmahmann J, Gunther J, Albert M. Predicting conversion to Alzheimer disease using standardized clinical information. Arch Neurol 2000;57:675–80.
[71] Perneczky R, Pohl C, Sorg C, Hartmann J, Tosic N, Grimmer T, et al. Impairment of activities of daily living requiring memory or complex reasoning as part of the MCI syndrome. Int J Geriatr Psychiatry 2006; 21:158–62.
[72] Farias S, Mungas D, Reed B, Cahn-Weiner D, Jagust W, Baynes K, et al. The measurement of everyday cognition (ECog): scale development and psychometric properties. Neuropsychology 2008;22:531–44.
[73] Jorm A, Korten E. Assessment of cognitive decline in the elderly by informant interview. Br J Psychiatry 1988;152:209–13.
[74] Glosser G, Gallo J, Duda N. Visual perceptual functions predict instrumental activities of daily living in patients with dementia. Neuropsychiatry Neuropsychol Behav Neurol 2002;15:198–206.
[75] Pfeffer R, Kurosaki C, Harrah J, Chance S, Filos S. Measurement of functional activities of daily living of older adults in the community. J Gerontol 1982;37:323–9.
[76] Galvin J, Roe C, Xiong C, Morris J. Validity and reliability of the AD8 informant interview in dementia. Neurology 2006;67:1942–8.
[77] Grigsby J, Kaye K, Baxter J, Shetterly SM, Hamman RF. Executive cognitive abilities and functional status among community-dwelling older person in the San Luis Valley Health and Aging Study. J Am Geriatr Soc 1998;46:590–6.
[78] Cahn-Weiner DA, Malloy PF, Boyle PA, Marran M, Salloway S. Prediction of functional status from neuropsychological tests in community-dwelling elderly individuals. Clin Neuropsychol 2000; 14:187–95.
[79] Farias ST, Harrell E, Neumann C, Houtz A. The relationship between neuropsychological performance and daily functioning in individuals with Alzheimer's disease: ecological validity of neuropsychological tests. Arch Clin Neuropsychol 2003;18:655–72.
[80] Farias ST, Mungas D, Jagust W. Degree of discrepancy between self and other-reported everyday functioning by cognitive status: dementia, mild cognitive impairment, and healthy elders. Int J Geriatr Psychiatry 2005;20:827–34.
[81] Cahn DA, Malloy PF, Salloway S, Rogg J, Gillard E, Kohn R, et al. Subcortical hyperintensities on MRI and activities of daily living in geriatric depression. J Neuropsychiatry Clin Neurosci 1996; 8:404–11.
[82] Cahn-Weiner DA, Farias ST, Julian L, Harvey DJ, Kramer JH, Reed BR, et al. Cognitive and neuroimaging predictors of instrumental activities of daily living. J Int Neuropsychol Soc 2007;13:747–57.
[83] Farias ST, Mungas D, Reed B, Haan MN, Jagust WJ. Everyday functioning in relation to cognitive functioning and neuroimaging in community-dwelling Hispanic and non-Hispanic older adults. J Int Neuropsychol Soc 2004;10:342–54.
[84] Boyle PA, Paul RH, Moser DJ, Cohen RA. Executive impairments predict functional declines in vascular dementia. Clin Neuropsychol 2004;18:75–82.
[85] Tomaszewski Farisa S, Cahn-Weiner DA, Harvey DJ, Reed BR, Mungas D, Kramer JH, et al. longitudinal changes in memory and executive functioning are associated with longitudinal change in instrumental activities of daily living in older Adults. Clin Neuropsychol 2008;23:1–16.
[86] Profenno L. Dementia. 3rd ed. London: Hodder Arnold; 2005.
[87] Overall JE, Gorham DR. Brief Psychiatric Rating Scale. Psychol Rep 1962;10:799–812.
[88] Teri L, Truax P, Logsdon R, Uomoto J, Zarit S, Vitaliano PP. Assessment of behavioral problems in dementia: the Revised Memory and Behavior Problems Checklist. Psychol Aging 1992;7:622–31.

[89] Helmes E, Csapo KG, Short JA. Standardization and validation of the Multidimensional Observation Scale for Elderly Subjects (MOSES). J Gerontol 1987;42:395–405.
[90] Morris JC, Heyman A, Mohs RC, Hughes JP, van Velle G, Fillenbaum G, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): part 1—clinical and neuropsychological assessment of Alzheimer's disease. Neurology 1989;39:1159–65.
[91] Mirra S, Heyman A, McKeel D, Sumi SM, Crain BJ, Brownlee LM, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): part II—standardization of the neuropathologic assessment of Alzheimer's disease. Neurology 1991;41:479–86.
[92] Wilkinson IM, Graham-White J. Psychogeriatric dependency rating scales (PGDRS): a method of assessment for use by nurses. Br J Psychiatry 1980;137:558–65.
[93] Olin J, Fox LS, Pawluczyk S, Taggart N, Schneider LS. A pilot randomized trial of carbamazepine for behavioral symptoms in treatment-resistant outpatients with Alzheimer's disease. Am J Geriatr Psychiatry 2001;9:400–5.
[94] De Deyn PP, Rabheru K, Rasmussen A, Bocksberger JP, Dautzenberg LJ, Eriksson S, et al. A randomized trial of risperidone, placebo, and haloperidol for behavioral symptoms of dementia. Neurology 1999;53:946–50.
[95] Paleacu D, Barak Y, Mirecky I, Mazeh D. Quetiapine treatment for behavioral and psychological symptoms of dementia in Alzheimer's disease patients: a 6-week, double-blind, placebo-controlled study. Int J Geriatr Psychiatry 2008;23:393–400.
[96] Lott AD, McElroy SL, Keys MA. Valproate in the treatment of behavioral agitation in elderly patients with dementia. J Neuropsychiatry Clin Neurosci 1995;7:314–9.
[97] Tariot PN, Solomon PR, Morris JC, Kershaw P, Lilienfeld S, Ding C. A 5-month, randomized, placebo-controlled trial of galantamine in AD. Neurology 2000;54:2269–76.
[98] Schnieder LS, Tariot PN, Dagerman KS, Davis SM, Hsaio JK, Ismail MS, et al. Effectiveness of atypical antipsychotic drugs in patients with Alzheimer's disease. N Engl J Med 2006;355: 1525–38.
[99] Profenno LA, Jamimovich L, Holt CJ, Porsteinsson A, Tariot PN. A randomized, double-blind, placebo-controlled pilot trial of safety and tolerability of two doses of divalproex sodium in outpatients with probable Alzheimer's disease. Curr Alzheimer Res 2005; 2:553–8.
[100] Tschanz JT, Treiber K, Norton MC, Welsh-Bohmer KA, Toone L, Zandi PP, et al. A population study of Alzheimer's disease: findings from the Cache County Study of Memory, Health and Aging. Care Manag J 2005;6:107–14.
[101] Lyketsos CG. Neuropsychiatric symptoms (behavioral and psychological symptoms of dementia) and the development of dementia treatments. Int Psychogeriatr 2007;19:409–20.
[102] Knopman DS, Knapp MJ, Gracon SI, Davis CS. The clinician interview-based impression (CIBI): a clinician's global change rating scale in Alzheimer's disease. Neurology 1994;44:2315–21.
[103] Rockwood K, Joffres C. Improving clinical descriptions to understand the effects of dementia treatment: consensus recommendations. Halifax Consensus Conference on Understanding the Effects of Dementia Treatment. Int J Geriatr Psychiatry 2002;17:1006–11.
[104] Morris JC, McKeel DW Jr, Fulling K, Torack RM, Berg L. Validation of clinical diagnostic criteria for Alzheimer's disease. Ann Neurol 1998;24:17–22.
[105] Berg L, Miller JP, Baty J, Rubin EH, Morris JC, Figiel G. Mild senile dementia of the Alzheimer type. 4. Evaluation of intervention. Ann Neurol 1992;31:242–9.
[106] Mulnard RA, Cotman CW, Kawas C, van Dyck CH, Sano M, Doody R, et al. Estrogen replacement therapy for treatment of mild to moderate Alzheimer disease: a randomized controlled trial. Alzheimer's Disease Cooperative Study. JAMA 2000;283:1007–15.
[107] Rockwood K. Capacity, population aging and professionalism. Can Med Assoc J 2006;174:1689.
[108] Kiresuk T, Smith A, Cardillo J, eds. Goal Attainment Scaling: Application, Theory, and Measurement. Hillsdale, NJ: Lawrence Erlbaum; 1994.
[109] Rockwood K, Li Y, Fay S, Mitnitski A. Patient/care partner goals of dementia treatment: comparing clinical trial and web-based accounts. Alzheimers Dement 2008;4(Suppl. 2):T174.
[110] Rockwood K, Fay S, Jarrett P, Asp E. Effect of galantamine on verbal repetition in AD: a secondary analysis of the VISTA Trial. Neurology 2007;68:1116–21.
[111] Guidance for industry patient reported outcome measures: use in medical product development to support labeling claims. February 2, 2006. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071975.pdf.
[112] Guidance for industry and review staff: target product profile—a strategic development process tool. March 30, 2007. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm080593.pdf.
[113] Caron M, Emery MP, Marquis P, Piault E, Scott J. Recent trends in the inclusion of patient-reported outcome data in approved drugs labeling by the FDA and EMEA. Patient Reported Outcomes Online. Available at: http://www.pro-newsletter.com/index.php?option=com_content&task=view&id=224&Itemid=77.
[114] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dementia 2005;1:55–66.
[115] Shaw LM. PENN biomarker core of the Alzheimer's Disease Neuroimaging Initiative. Neurosignals 2008;16:19–23.
[116] Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, et al. Cerebrospinal fluid biomarker signature in Alzheimer's Disease Neuroimaging Initiative subjects. Ann Neurol 2009;65:403–13.
[117] Landau SM, Madison C, Wu D, Cheung C, Foster N, Reiman E, et al. Pinpointing change in Alzheimer's disease: longitudinal FDG-PET analysis from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement 2008;4(Suppl 2):T291–2.
[118] Carrillo MC, Sanders CA, Katz RG. Maximizing the Alzheimer's Disease Neuroimaging Initiative (ADNI) II. Alzheimers Dement 2009;5:271–5.