RESEARCH ARTICLE

Variability in the validity and reliability of outcome measures identified in a systematic review to assess treatment efficacy of cognitive enhancers for Alzheimer's Dementia

Charlene Soobiah[1,2], Mina Tadrous[2,3], Sandra Knowles[3,4], Erik Blondal[1,2], Huda M. Ashoor[2], Marco Ghassemi[2], Paul A. Khan[2], Joanne Ho[5,6], Andrea C. Tricco[1,2,7], Sharon E. Straus[1,2,8]*

1 Institute for Health Policy, Management & Evaluation, University of Toronto, Toronto, Ontario Canada, **2** Li Ka Shing Knowledge Institute of St. Michael's Hospital, Toronto, Ontario Canada, **3** Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, Ontario Canada, **4** Clinical Pharmacology & Toxicology, Department of Pharmacy, Sunnybrook Health Sciences Centre, Toronto, Ontario Canada, **5** Schlegel Research Institute for Aging, Waterloo, Ontario Canada, **6** Department of Medicine, McMaster University, Hamilton, Ontario Canada, **7** Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario Canada, **8** Division of Geriatric Medicine, Department of Medicine, University of Toronto, Suite RFE 3–805, Toronto, Ontario Canada

* sharon.straus@utoronto.ca

Citation: Soobiah C, Tadrous M, Knowles S, Blondal E, Ashoor HM, Ghassemi M, et al. (2019) Variability in the validity and reliability of outcome measures identified in a systematic review to assess treatment efficacy of cognitive enhancers for Alzheimer's Dementia. PLoS ONE 14(4): e0215225. https://doi.org/10.1371/journal.pone.0215225

Editor: Stephen D. Ginsberg, Nathan S Kline Institute, UNITED STATES

Received: October 29, 2018

Accepted: March 29, 2019

Published: April 18, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The systematic review was funded by Drug Safety and Effectiveness Network/Canadian Institutes for Health Research (CIHR) grant TGE-117529 obtained by SES and ACT. The funders had no role in study design, data collection and

Abstract

Introduction

Selection of optimal outcome measures is a critical step in a systematic review; inclusion of uncommon or non-validated outcome measures can impact the uptake of systematic review findings. Our goals were to identify the validity and reliability of outcome measures used in primary studies to assess cognition, function, behaviour and global status; and, to use these data to select outcomes for a systematic review (SR) on treatment efficacy of cognitive enhancers for Alzheimer's Dementia (AD).

Methods

Articles fulfilling the eligibility criteria of the SR were included in a charting exercise to catalogue outcome measures reported. Outcome measures were then assessed for validity and reliability. Two independent reviewers abstracted data on outcome measures and validity and reliability reported for cognition, function, behaviour and global status.

Results

129 studies were included in the charting exercise; 57 outcome measures were identified for cognition, 21 for function, 13 for behaviour and 10 for global status. A total of 35 (61%) cognition measures, 10 (48%) functional measures, 8 (61%) behavioural measures and four (40%) of global status measures were only used once in the literature. Validity and reliability information was found for 51% of cognition measures, 90% of function and global status measures and 100% of behavioural measures.

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

While a large number of outcome measures were used in primary studies, many of these were used only once. Reporting of validity and reliability varied in AD studies of cognitive enhancers. Core outcome sets should be used when available; when they are not available researchers need to balance frequency of reported outcome measures, their respective validity and reliability, and preferences of knowledge users.

Systematic review registration

CRD#42012001948

Introduction

Outcome measures are tools, instruments or scales used to assess an outcome. For example, the Activities of Daily Living (ADL) [1, 2] is to assess function in older adults. Selection of appropriate outcome measures for inclusion in a systematic review is imperative, to ensure research relevance for knowledge users [3]. Knowledge users are individuals who may use research findings to make a decision and can include patients, clinicians, policymakers, or researchers [4, 5]. Inclusion of non-validated or uncommon measures in a systematic review can make it difficult for knowledge users to interpret and utilize findings to make informed decisions [6, 7]. Multiple measures exist for a particular outcome and their selection for use in systematic reviews can be challenging; a researcher must identify measures that are valid, reliable, and clinically relevant [8].

We previously conducted a systematic review and network meta-analysis on the comparative safety and efficacy of cognitive enhancers for treatment of Alzheimer's Dementia (AD). This review was conducted with geriatricians and policymakers (i.e. knowledge users) who wanted to use the results to inform decision-making on medication use in Canada [9, 10]. During the systematic review process, numerous outcome measures were identified for our pre-specified outcomes (i.e., cognition, function, behaviour, and global status), as such we sought to identify the validity and reliability of outcome measures used in primary studies to assess out pre-specified outcomes; and, to use these data to select outcomes for inclusion in our systematic review.

Methods

As our systematic review methods and results were previously published [9], the focus of this paper is on how outcome measures were identified for inclusion in the review. Our systematic review was registered (CRD#42012001948) [10] and included experimental and observational studies that reported on cognitive enhancers approved for use in Canada (i.e., donepezil, galantamine, rivastigmine and memantine) for patients with AD. Studies had to report on at least one pre-specified outcome (i.e., cognition, function, behaviour, and global status) to be considered eligible for inclusion in the systematic review. Several electronic databases were searched from inception to December 31, 2011 to identify studies. Two independent reviewers assessed each citation and full text article against eligibility criteria. Studies fulfilling eligibility criteria were included in the present study.

Charting exercise

Primary studies fulfilling the eligibility criteria were included in a data charting exercise. For each study, we catalogued the measures that were used for our pre-specified outcomes of interest. This approach is frequently used in scoping reviews to synthesize information and understand knowledge gaps [11] This step was conducted prior to data abstraction in the systematic review -. The literature search for the systematic review was updated prior to publication; however, we used the original literature search for this current study.

A standardized spreadsheet (Excel) was created to capture the measures used to assess each outcome. A calibration exercise was conducted with reviewers (CS, ACT, JH, EB, HA, MG, PAK) on a random subset of studies (n = 10) until adequate agreement was achieved (>80% agreement) and the spreadsheet was modified accordingly. Reviewers independently abstracted outcome measures using the standardized spreadsheet.

Validity and reliability assessment

To obtain validity and reliability information for each measure in the charting exercise a three-step process was employed. First, references of primary studies included in the systematic review were used to locate the measure citation. Second, if the measure was not cited, an electronic literature search was conducted in MEDLINE, Mental Measures Yearbook, Health and Psychosocial Instruments and/or Google Scholar (from inception to January 2015) using the measure name and keywords such as 'validation', 'psychometric' and 'reliability'. Third, if the validity or reliability information for a measure were not located by searching, authors of the included study were contacted to request this information.

A second standardized spreadsheet was created to capture validity and reliability information and a calibration exercise was conducted with reviewers (CS, MT, SK) on a random subset of studies (n = 10) until adequate agreement was achieved (>80% agreement). The spreadsheet was modified accordingly. Next, pairs of reviewers (CS, MT, and SK) abstracted validity and reliability information independently.

Validity and reliability information were categorized using the description reported by the authors in the cited studies. When type of validity or reliability examined was not explicitly stated or was unclear, we used an established framework to categorize the data [11]. This framework suggests that all forms of validity fall under the category of construct validity, which can be further divided into translational (face or content) and criterion validity (concurrent, convergent, predictive, discriminant and predictive; S1 Table) [11]. To estimate reliability of an outcome measure, internal consistency, test-retest and inter-rater reliability were used (S1 Table) [11]. Funding was categorized as industry sponsored (e.g., funding received from the private sector), mixed funding (e.g., funding derived from private and public sectors) and non-industry sponsored (e.g., funding from the public sector).

Analysis

Data from the charting exercise and validity and reliability assessments were analyzed descriptively using frequencies in Excel.

Results

Charting results

In total, 15,556 citations were screened, of which 129 full-text articles were included in the charting exercise. Overall, 101 unique outcome measures were identified including: 57 measures for cognition, 21 for function, 13 for behaviour, and 10 for global status (S2, S3, S4 and S5 Tables).

Frequency of outcome measures

For cognition assessment, the most frequent outcome measures identified from our literature search were the: Mini-Mental State Exam (MMSE)[12] reported in 80 studies; Alzheimer's Disease Assessment Scale–cognitive subscale (ADAS-cog) [13] reported in 61 studies; and Severe Impairment Battery (SIB)[14] reported in 13 studies (S2 Table). Only 7 (12%) measures were used in more than 5 primary studies. A total of 35 (61%) cognition measures were used once.

For function assessment, the most frequently used outcome measures were: Activities of Daily Living (ADL) [2] reported in 19 studies, Alzheimer's disease Cooperative Studies–ADL (ADAS-ADL)[15] and Disability Assessment for Dementia (DAD)[16, 17] reported in 7 studies (S3 Table). Ten (48%) functional measures were used once.

For behavioral assessment, the most frequently reported outcome measures were the: Neuropsychiatric Inventory (NPI) [18] reported in 36 studies; Behavioural Pathology in Alzheimer's Disease Rating Scale (BEHAVE-AD)[19, 20]; and, Cohen-Mansfield agitation Inventory (CMAI)[21] reported in 5 studies (S4 Table). Eight (61%) behavioural measures were used once.

For global status, the most frequently reported outcome measures were the Clinician Interview-Based Impression of Change plus caregiver input (CIBIC-plus)[22] reported in 35 studies, and the Clinical Global Impression of Change (CGIC)[23] reported in 10 studies (S5 Table). Four (40%) global status measures were used once.

Reporting of validity and reliability of outcome measures

Of the 101 outcome measures identified from the 129 primary studies, 74 (73%) outcome measures were supported by citation of references while 27 (27%) outcome measures were not (Fig 1). Of the 74 outcome measures that had citations, 57 (77%) citations reported evidence of validity and/or reliability (i.e., 22 for cognitive measures, 17 for functional measures, 11 for behavioural measures, 7 for global measures) (Fig 1). The citations for the remaining 17 outcome measures (23%) did not contain validity or reliability information or the source was irretrievable. Citations for these 17 outcome measures included 8 textbooks, 8 journal articles and 4 test manuals. We were unable to access the cited test manuals for Digit Span, Digit Symbols Test, Category Fluency Test, and the Wechsler Adult Intelligence Scale, as the material was proprietary. The citation identified for the MENFIS outcome measure, led to a non-English article [24] and translation was not possible.

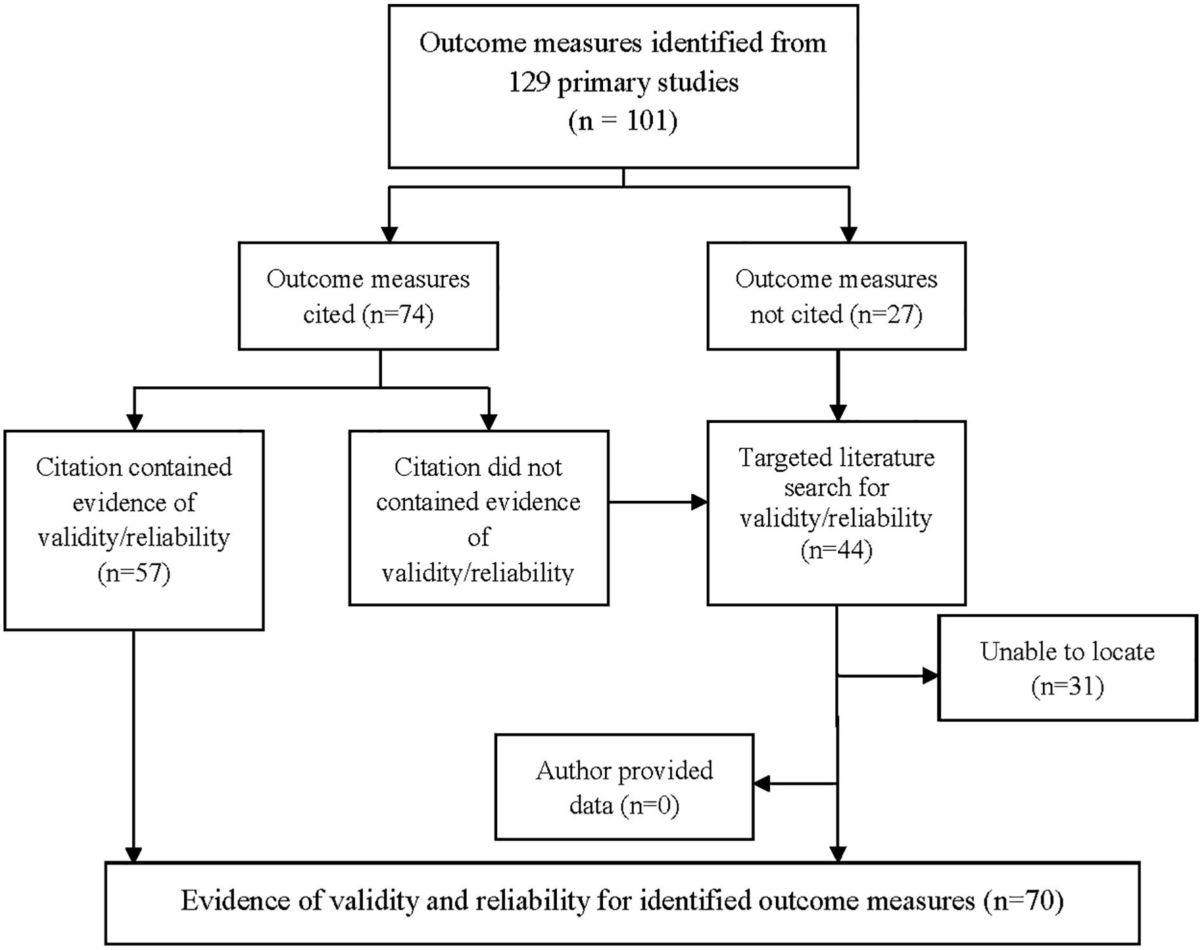
We conducted a targeted literature search for 44 outcome measures that were not supported by appropriate citations in the primary studies; these included 35 cognitive measures, 4 functional measures, 3 global measures and 2 behavioural measures. Validity and reliability could not be located for 31 (31%) outcome measures, the majority (n = 28) of which addressed cognition (Fig 1). Six study authors were contacted; however validity or reliability data were not provided.

Validity and reliability data could not be located for 31 outcome measures that were reported in 20 primary studies. Of these 20 studies, 7 (35%) studies were funded by a mix of industry and non-industry sponsors, 6 (30%) were funded by industry sponsors, and, 2 (10%) were funded by non-industry sponsors. Five (25%) studies did not disclose funding (S6 Table).

Overall, validity and reliability information was located for 51% of cognitive measures (n = 29), 90% of functional measures (n = 10) and 90% global status measures (n = 9). Validity or reliability data were available for all behavioural outcome measures (n = 13).

Cognition outcome measures

Validity and reliability information was identified for 29 of 57 (51%) cognition measures (Table 1). Concurrent and convergent were the most frequently reported forms of validity, while content or face validity were the least reported forms.


Outcome measures identified from 129 primary studies (n = 101)
Outcome measures cited (n=74)
Outcome measures not cited (n=27)
Citation contained evidence of validity/reliability (n=57)
Citation did not contained evidence of validity/reliability
Targeted literature search for validity/reliability (n=44)
Unable to locate (n=31)
Author provided data (n=0)
Evidence of validity and reliability for identified outcome measures (n=70)

Note: Multiple citations were used to obtain validity and reliability data.

Fig 1. Flow diagram of locating validity and reliability.

https://doi.org/10.1371/journal.pone.0215225.g001

The Frontal Assessment Battery (FAB)[25], Global Deterioration Scale (GDS)[26], Groton Maze Learning Test, (GMLT)[27, 28] the Rey-Osterrieth Complex Figure copy (ROCF-copy) [29–31] and the Wechsler Memory Scale (WMS)[32] had evidence of more than two forms of validity (Table 1). Four cognitive outcome measures did not have evidence of validity, namely the Controlled Oral Word Association Test (COWAT)[33], Functional Linguistic Communication Inventory (FLCI)[34, 35], Raven Coloured Matrices (RCM)(36), and Wechsler Logical Memory Test (WLMT) [37, 38].

Test-retest and inter-rater reliability were the most frequently reported forms of reliability for cognition measures. Eleven (38%) of the 29 cognition measures had evidence of two forms of reliability, 10 (34%) had evidence of one form of reliability and eight (27%) had no evidence of reliability (Table 1).

Function outcome measures

Validity and reliability information were identified for 19 of the 21 (90%) function measures. Concurrent and convergent validity were the most frequently reported forms of validity. The Bristol Activities of Daily Living (BADL)[53], and the Caregiver Perceived Burden

Table 1. Validity and reliability of cognitive measures used in treatment efficacy studies of Alzheimer's Dementia (n = 29).

Name of Scale (YR)	Validity						Reliability		
	Face/Content	Construct	Concurrent	Predictive	Convergent	Discriminant	Internal consistency	Test-retest	Inter-rater
ADAS (1984)[13]			✓					✓	✓
ADAS-Cog (1984)[13]			✓					✓	✓
ADAS-OT (1984)[13]							✓		
ASHA-FACS (2008)[39]			✓				✓		✓
BDS (1968)[40]				✓					
CAMCOG (1986)[41]					✓				✓
CDR (1988)[42]					✓		✓		✓
CDR-SB (1988)[43, 44]					✓		✓		✓
CDT (1989)[45]					✓			✓	
CERAD-cog (1989)[46]		✓						✓	✓
COWAT (1996)[33]							✓	✓	
DWRT (1989)[36]				✓				✓	
FAB (2000)[25]			✓			✓	✓		✓
FLCI (1994)[34, 35]								✓	
GDS (1982)[26]				✓	✓				
GMLT (2008)[27, 28]		✓			✓				
MMSE (1975)[12]			✓					✓	✓
RAVLT (1988)[47]			✓						
RCM (1989)[36]								✓	
ROCF-copy (1990)[29–31]		✓		✓			✓		✓
ROCF-recall (1990)[31]					✓			✓	✓
SIB (1997)[14]			✓					✓	
SKT (1992)[48]		✓							
SR (2011)[49]			✓						
TMT (1958)[50, 51]						✓	✓		
TMT-A (1958)[51]						✓			
WLMT (1993) [37, 38]								✓	✓
WMS (1990)[32]		✓			✓				
ZVT (1985)[52]					✓				

Abbreviations: ADAS Alzheimer's Disease Assessment Scale; ADAS-Cog Alzheimer's Disease Assessment Scale- Cognitive subscale; ADAS-OT Alzheimer's Disease Assessment Scale—Orientation Test; ASHA-FACS American Speech-Language Hearing Association- Functional Assessment of Communication Skills for Adults Basic needs and social communication subscales; BDS Blessed Dementia Scale; CAMCOG Cambridge Cognitive Examination; CDR Clinical Dementia Rating Scale; CDR-SB Clinical Dementia Rating Scale–Sum of Boxes; CDT Clock Drawing Test; CERARD-cog Consortium to Establish a Registry for Alzheimer's Disease–cog subscale; COWAT Controlled Oral Word Association Test; DWR Delayed word recall; FAB Frontal Assessment Battery; FLCI Functional Linguistic Communication Inventory; GDS Global Deterioration Scale; GMLT Groton Maze Learning Task; MMSE Mini Mental State Exam; RAVLT Rey Auditory Verbal Learning Test; RCM Raven Colored Matrices; ROCF-recall Rey-Osterrieth complex figure recall; ROCF-copy Rey-Osterrieth complex figure copy; SIB Severe Impairment Battery; SR Story Recall; SKT Syndrome Kurtz test; TMT Trail Making Test; TMT-A Trail Making Test A; WLMT Wechsler Logical Memory Test; YR year ZVT Zahlen–Verbindungs Test.
NOTE: Year reported is based on the earliest published paper reporting on validity or reliability for each outcome measure. Validity data for the following scales/ measures could not be located: Digit Span; Digit Symbols test, Stroop Test; Stockholm Gerontology Research Center test; Verbal Fluency, Clock Recognition, Stockholm Gerontology Research Center test- D-prime value; Word Paradigm–free recall; Cambridge Automated Neuropsychiatric Test Assessment Battery; Cognitive Drug Research Test Battery; Category Fluency Test; Computerized Memory Battery Test; Forced Delayed Recognition; Immediate Visual Memory; Multiple Feature Target Cancellation; Non-Demanding Test of Visual Attention; NYU Stories Test- Delayed Recognition Subscale; Oral Production Test, Reading and Setting a Clock Test; Serial Reaction Time Task; Spatial Span; Test of Constructional Praxis; Temporal Rule Induction; Token Test, Visual Motor Gestalt Test; Wechsler Adult Intelligence Scale; Word Fluency; and Word Learning

https://doi.org/10.1371/journal.pone.0215225.t001

Table 2. Validity and reliability of functional status measures used in treatment efficacy studies for Alzheimer's Dementia (n = 19).

Name of Scale (YR)	Validity						Reliability		
	Face/Content	Construct	Concurrent	Predictive	Convergent	Discriminant	Internal consistency	Test-retest	Inter-rater
AAIQOL (1996)[58]	✓			✓				✓	
ADCS-ADL (1997)[15]	✓		✓					✓	
ADCS-ADL-severe (2005)[59]			✓				✓	✓	
ADFACS (2014)[60]			✓				✓		
ADL (1970)[1]			✓				✓	✓	✓
BADL (1996)[53]	✓	✓	✓					✓	
BI (1997)[61, 62]					✓				✓
CMCS (2000)[55]					✓				
CPBQ (2012)[54]	✓		✓		✓		✓	✓	
DAD (1999)[17, 63]	✓		✓				✓	✓	✓
FAST (1992)[64]			✓						✓
FRS (1989)[65]			✓			✓		✓	✓
GAFS (2006)[66]					✓				✓
GAS (1989)[67]					✓				✓
IADL (1970)[1]			✓		✓		✓	✓	✓
IDDD (1991)[68]			✓		✓		✓		
NOSGER (1991)[57]			✓				✓	✓	✓
PDS (1989)[69]	✓						✓	✓	
ZBI (1980)[70, 71]							✓		✓

Abbreviations: AAIQoL Activity & Affect Indicators of Quality of Life; ADCS-ADL Alzheimer's Disease Cooperative Studies Activities of Daily Living Inventory; ADCS-ADL-severe Alzheimer's Disease Cooperative Studies Activities of Daily Living Severe impairment subscale; ADFACS Alzheimer's Disease Functional Assessment and Change Scale, BI Barthel Index; BADL Bristol Activities of Daily Living; CPBQ Caregiver-Perceived burden Questionnaire; CMCS Caregiver-rated Modified Crichton Scale; DAD Disability Assessment for Dementia; FAST Functional Assessment Screening Tool; FRS Functional Rating Scale; GAST Global Assessment of Functioning Scale; GAS Goal Attainment Scale; IADL Instrumental Activities of Daily Living; IDDD Interview for Deterioration in Daily living activities in Dementia; NOSGER Nurses Observation Scale for Geriatric Patients; ADL Physical Self-Maintenance/Activities of Daily Living; PDS Progressive Deterioration Scale; ZBI Zarit Burden Interview. **Note:** Year reported is based on the earliest published paper reporting on validity or reliability for each outcome measure. Validity data for MENFIS and CBQ could not be located

https://doi.org/10.1371/journal.pone.0215225.t002

Questionnaire (CPBQ)[54] had evidence of three forms of validity. Six of the 19 (32%) measures had evidence of two forms of validity and 11 (58%) had evidence of one form of validity (Table 2).

Test-retest was the most frequently reported measure of reliability. All function measures had evidence of at least one form of reliability with the exception of the caregiver-rated modified Crichton scale (CMCS) which did not report reliability [55]. Activities of Daily Living (ADL) [1, 56], DAD [16, 17], Instrumental Activities of Daily Living (IADL)[1] and the Nurses Observation Scale for Geriatric Patients (NOSGER)[57] each had evidence of all forms of reliability (Table 2).

Behaviour outcome measures

Validity and reliability information were identified for all 13 behavioural outcome measures. Face/content and convergent validity were the most frequently reported forms of validity. The Behavioural Rating Scale for geriatric patients (BRS)[72, 73] and the Geriatric Depression Scale (GDS)[74] had evidence of more than three forms of validity. We were unable to find validity information for the Apathy Scale (AS) [75].

Table 3. Validity and reliability of behavioural status measures used in treatment efficacy studies in Alzheimer's Dementia (n = 13).

Name of Scale	Validity						Reliability		
	Face/Content	Construct	Concurrent	Predictive	Convergent	Discriminant	Internal consistency	Test-retest	Inter-rater
AS (1992)[75]							✓	✓	✓
BEHAVE-AD (1990)[19, 73, 78]		✓							✓
b-NPI (2000)[79]					✓			✓	
b-PRS (1962)[80–84]	✓	✓						✓	✓
BRS (1997)[72, 76, 85]	✓	✓			✓		✓	✓	✓
CA-NPI (2004)[86]					✓			✓	
CGRS (1989)[87]	✓								✓
CMAI (1989)[21, 88]		✓							✓
GDS (1983)[74]	✓			✓	✓		✓	✓	
NPI (1994)[18]	✓		✓				✓	✓	✓
NPI-CDS (2000)[79]			✓					✓	✓
NPI-NH (2001)[77]			✓						
PAS (1995)[89]					✓				✓

Abbreviations: AS Apathy Scale; BEHAVE-AD Behavioural Pathology in Alzheimer's Disease Rating Scale, BRS Behavioural Rating Scale for Geriatric Patients, b-NPI Brief Neuropsychiatric Inventory Scale; b-PRS Brief Psychiatric Rating Scale; CA-NPI Caregiver-Administered-Neuropsychiatric Inventory; CMAI Cohen-Mansfield Agitation Inventory; CGRS Crichton Geriatric Rating Scale; GDS Geriatric Depression Scale; NPI-CDS Neuropsychiatric Inventory–Caregiver distress scale; NPI Neuropsychiatric Inventory; NPI-NH Neuropsychiatric Inventory–Nursing Home version (NPI-NH); PAS Pittsburgh Agitation Scale YR year. **Note:** Year reported is based on the earliest published paper reporting on validity or reliability for each outcome measure.

https://doi.org/10.1371/journal.pone.0215225.t003

Inter-rater and test-retest were the most frequently reported forms of reliability for behavioural outcome measures. The Apathy scale (AS)[75], Behavioural Rating Scale for geriatric patients (BRS)[72, 76] and Neuropsychiatric Inventory (NPI)[18] had evidence of all forms of reliability. Three (23%) measures had two forms of reliability, whereas six (46%) had evidence of one form of reliability. We were unable to find evidence of reliability for the Neuropsychiatric Inventory nursing home version (NPI-NH)[77] (Table 3).

Global status outcome measures

Validity and reliability of global status measures were located for 9 of 10 (90%) global status outcome measures. Construct, convergent and discriminant validity were the most frequently reported forms of validity. Four (44%) of the nine global status measures had evidence of two forms of validity and two (22%) had one form of validity. The caregiver-rated global impression (Caregiver-rated GI)[90] and the CIBIC-plus(22) outcome measures did not have any associated validity evidence (Table 4).

Test-retest reliability was the most frequently reported form of reliability. We were unable to locate evidence of reliability for the Clinical Global Impressions of Change (CGIC) scale [23, 91].

Discussion

Overall, 129 studies were included in the systematic review on treatment efficacy of cognitive enhancers for AD patients and from these articles, we identified 101 measures for our outcomes of interest (57 cognition measures, 21 function measures, 13 behaviour measures and 10 global status measures). We identified validity and reliability data for 51% of cognition measures, 100% of behaviour measures and 90% of function measures and global status measures.

Table 4. Validity and reliability of global status measures used in treatment efficacy studies in Alzheimer's Dementia (n = 9).

Name of Scale	Validity							Reliability		
	Face/ Content	Construct	Criterion	Concurrent	Predictive	Convergent	Discriminant	Internal consistency	Test-retest	Inter-rater
ADCS-CGI (1997)[92]	✓				✓				✓	
CRGI (1994)[90]									✓	
CGIC (1996)[23, 91]						✓	✓			
CGIC-severe (1992)[91, 93]						✓			✓	✓
CGI-I (2006)[91]						✓	✓		✓	
CIBIC- plus (1994)[22, 90]									✓	✓
CSS (1990)[94, 95]		✓		✓				✓		
GBS (1982)[96, 97]		✓		✓						✓
SCB (1991)[98]		✓					✓	✓		✓

Abbreviation: ADCS-CGIC Alzheimer's Disease Cooperative Studies–Clinician Global Impressions of Change; CSS Caregiver Stress Scale; CRGI Caregiver-rated Global Impression; CGIC-severe Clinical Global Impression of Change—severe subscale; CGIC Clinical Global Impression of Change; CGI-I Clinical Global Impression of Improvement; CIBIC-plus Clinician Interview-Based Impression of Change plus caregiver input; GBS Gottfries-Bråne-Steen Scale, SCB Screen for Caregiver Burden.
Note Year reported is based on the earliest published paper reporting on validity or reliability for each outcome measure. Patient Global Assessment (PGA) scale could not be located

https://doi.org/10.1371/journal.pone.0215225.t004

Studies in which validity or reliability were not supported by a citation were supported by funding from industry or a mix of industry and non-industry sources.

Our study findings are consistent with previous studies that examined psychometric properties of AD measures. Demers and colleagues reviewed the psychometric properties of outcome measures used in AD trials and reported their validity and reliability [99–102]. They reported on three measures for global status, eight measures for function and eight measures for behaviour. Across outcome measures, they found variable evidence of reliability and validity. Of note, they identified that function and behaviour measures lacked evidence of validity and reliability [99–102]. The work by Dermers and colleagues was based on a report in 2000 and identified 26 trials that reported use of cognitive enhancers for AD patients. Our study included a systematic search for randomized controlled trials as well as observational data and did not include any limitations on years of publication. As such, we identified more outcome measures for function, behaviour, and global status and also examined cognition, which was not reported in their previous work.

Locating validity and reliability information for cognition was a challenge. A majority (n = 24) of the valid and reliable scales to measure cognition were published before 2000, suggesting that cognition measures used in AD trials have not changed substantially over time. We observed a high number of cases where a cognition measure was only used in one study, which makes it difficult to make comparisons or draw conclusions across studies. In 2001, the MMSE [12] became a proprietary measure, which means a licensed version had to be purchased before use [103]. Many researchers and clinicians are compelled to use open source tools to assess cognition. The Montreal Cognitive Assessment (MoCA) measure has evidence of validity and but was not used in any of the included studies and it has not been validated for use in as many settings as the MMSE.

Our study has some limitations. First, the charting exercise was conducted on the original systematic review search (December 2011) rather than updated literature search. We conducted the charting exercise to gain insight on which measures were used in studies to assess outcomes; as such, we did not update the charting exercise as outcome measures were already selected. Second, we followed a three-step process for identifying validity and reliability information rather than completing a systematic search for each outcome measure. Given limited resources, we felt our three-step approach was feasible and yielded information needed to select outcome measures for our systematic review. Third, we did not assess the methodological quality of the validity and reliability information and merely categorized whether there was reported evidence of validity or reliability based on how authors reported the information. Lastly, we did not search grey literature sources to obtain validity and reliability evidence, given limited resources.

Conclusions

Our paper highlights the variability in the reporting of outcome measures used in AD studies. We identified multiple outcome measures reported in the primary studies; many of these were used only once in the primary studies that were included in our systematic review. The large number of measures used in studies makes it difficult to synthesize the evidence. Cataloguing and assessing validity and reliability of each outcome measure in studies can be resource intensive; using core outcome sets (i.e., agreed upon outcomes and measures used in a particular discipline) in systematic reviews is recommended as this may streamline outcome selection [6]. In lieu of core outcome sets, researchers need to balance frequency of reported outcome measures, their respective validity and reliability, and preferences of knowledge users.

Supporting information

S1 Table. Validity and Reliability Definitions.
(PDF)

S2 Table. Frequency of Cognitive Outcome Measures (n = 57).
(PDF)

S3 Table. Frequency of Functional Outcome Measures (n = 21).
(PDF)

S4 Table. Frequency of Behavioural Outcome Measures (n = 13).
(PDF)

S5 Table. Frequency of Global Status Outcome Measures (n = 10).
(PDF)

S6 Table. Missing References and Sources of Funding.
(PDF)

Acknowledgments

We would like to thank Alissa Epworth for conducting the targeted literature search and Jessica Comilang, Katarzyna Pyka, and Alistair Scott for their help in locating citations and full-text articles.

Author Contributions

Conceptualization: Charlene Soobiah, Sharon E. Straus.

Formal analysis: Charlene Soobiah.

Funding acquisition: Andrea C. Tricco, Sharon E. Straus.

Investigation: Charlene Soobiah, Mina Tadrous, Sandra Knowles, Erik Blondal, Huda M. Ashoor, Marco Ghassemi, Paul A. Khan, Joanne Ho, Andrea C. Tricco.

Methodology: Charlene Soobiah, Sharon E. Straus.

Supervision: Sharon E. Straus.

Writing – original draft: Charlene Soobiah.

Writing – review & editing: Mina Tadrous, Sandra Knowles, Erik Blondal, Huda M. Ashoor, Marco Ghassemi, Paul A. Khan, Joanne Ho, Andrea C. Tricco, Sharon E. Straus.

References

1. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Nursing Research. 1970; 19(3):278.
2. Lawton MP. Scales to measure competence in everyday activities. Psychopharmacology bulletin. 1988; 24(4):609–14. PMID: 3074322
3. Straus S, Tetroe J., Graham ID.,Straus S., Tetroe J., Graham ID.,. Knowledge Translation in Health Care: Moving from Evidence to Practice2014.
4. Segal J. Choosing the important outcomes for a systematic review of a medical test. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ); 2012. Contract No.: 12-EHC075-EF.
5. Wallace J, Byrne C, Clarke M. Making evidence more wanted: a systematic review of facilitators to enhance the uptake of evidence from systematic reviews and meta-analyses. Int J Evid Based Healthc. 2012; 10(4):338–46. https://doi.org/10.1111/j.1744-1609.2012.00288.x PMID: 23173658
6. Clarke M, Williamson PR. Core outcome sets and systematic reviews. Syst Rev. 2016; 5:11. https://doi.org/10.1186/s13643-016-0188-6 PMID: 26792080
7. Wallace J, Nwosu B, Clarke M. Barriers to the uptake of evidence from systematic reviews and meta-analyses: a systematic review of decision makers' perceptions. BMJ Open. 2012; 2(5).
8. Mullen E. Choosing outcome measures in systematic reviews: Critical challenges. Research on Social Work Practice. 2006; 16(1):84–90.
9. Tricco AC, Ashoor HM, Soobiah C, Rios P, Veroniki AA, Hamid JS, et al. Comparative Effectiveness and Safety of Cognitive Enhancers for Treating Alzheimer's Disease: Systematic Review and Network Metaanalysis. J Am Geriatr Soc. 2018.
10. Tricco AC, Vandervaart S, Soobiah C, Lillie E, Perrier L, Chen MH, et al. Efficacy of cognitive enhancers for Alzheimer's disease: protocol for a systematic review and network meta-analysis. Syst Rev. 2012; 1:31. https://doi.org/10.1186/2046-4053-1-31 PMID: 22742585
11. Trochim W. The Research Methods Knowledge Base (2nd ed). Cincinnati, OH: Atomic Dog Publishing; 2001.
12. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research. 1975; 12(3):189–98. PMID: 1202204
13. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. The American journal of psychiatry. 1984.
14. Schmitt FA, Ashford W, Ernesto C, Saxton J, Schneider LS, Clark CM, et al. The severe impairment battery: concurrent validity and the assessment of longitudinal change in Alzheimer's disease. The Alzheimer's disease cooperative study. Alzheimer disease and associated disorders. 1996; 11:S51–6.
15. 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. Alzheimer Disease & Associated Disorders. 1997; 11:33–9.
16. Gauthier S, Gélinas I, Gauthier L. Functional disability in Alzheimer's disease. International Psychogeriatrics. 1997; 9(S1):163–5.
17. Gélinas I, Gauthier L, McIntyre M, Gauthier S. Development of a functional measure for persons with Alzheimer's disease: the disability assessment for dementia. American Journal of Occupational Therapy. 1999; 53(5):471–81. PMID: 10500855
18. 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 (12):2308–. PMID: 7991117
19. Reisberg B, Auer SR, Monteiro IM. Behavioral pathology in Alzheimer's disease (BEHAVE-AD) rating scale. International Psychogeriatrics. 1997; 8(S3):301–8.

20. Reisberg B, Borenstein J, Salob SP, Ferris SH. Behavioral symptoms in Alzheimer's disease: phenomenology and treatment. The Journal of clinical psychiatry. 1987.

21. Cohen-Mansfield J. Assessment of disruptive behavior/agitation in the elderly: function, methods, and difficulties. Journal of geriatric psychiatry and neurology. 1995.

22. Boothby H, Mann A, Barker A. Factors determining interrater agreement with rating global change in dementia: The cibic-plus. International journal of geriatric psychiatry. 1995; 10(12):1037–45.

23. Kørner A, Lauritzen L, Bech P. A psychometric evaluation of dementia rating scales. European psychiatry. 1996; 11(4):185–91. https://doi.org/10.1016/0924-9338(96)88389-1 PMID: 19698448

24. Homma A. NR, Ishii T., Hasegawa K. Development of a new rating scale for dementia in the elderly: Mental function impairment scale (MENFIS). Japanese Journal of Geriatric Psychiatry. 1991; 2 (10):1217–22.

25. Dubois B, Slachevsky A, Litvan I, Pillon B. The FAB A frontal assessment battery at bedside. Neurology. 2000; 55(11):1621–6. PMID: 11113214

26. Reisberg B, Ferris SH, de Leon MJ, Crook T. The Global Deterioration Scale for assessment of primary degenerative dementia. The American journal of psychiatry. 1982.

27. Pietrzak RH, Maruff P, Mayes LC, Roman SA, Sosa JA, Snyder PJ. An examination of the construct validity and factor structure of the Groton Maze Learning Test, a new measure of spatial working memory, learning efficiency, and error monitoring. Archives of Clinical Neuropsychology. 2008; 23(4):433–45. https://doi.org/10.1016/j.acn.2008.03.002 PMID: 18448309

28. Pietrzak RH, Maruff P, Snyder PJ. Convergent validity and effect of instruction modification on the groton maze learning test: A new measure of spatial working memory and error monitoring. International Journal of Neuroscience. 2009; 119(8):1137–49. PMID: 19922344

29. Loring DW, Martin RC, Meador KJ, Lee GP. Psychometric construction of the Rey-Osterrieth complex figure: Methodological considerations and interrater reliability. Archives of Clinical Neuropsychology. 1990; 5(1):1–14. PMID: 14589539

30. Deckersbach T, Savage CR, Henin A, Mataix-Cols D, Otto MW, Wilhelm S, et al. Reliability and validity of a scoring system for measuring organizational approach in the Complex Figure Test. Journal of Clinical and Experimental Neuropsychology. 2000; 22(5):640–8. https://doi.org/10.1076/1380-3395(200010)22:5;1-9;FT640 PMID: 11094399

31. Berry DT, Allen RS, Schmitt FA. Rey-Osterrieth Complex Figure: Psychometric characteristics in a geriatric sample. The Clinical Neuropsychologist. 1991; 5(2):143–53.

32. Altepeter TS, Adams RL, Buchanan WL, Buck P. Luria Memory Words Test and Wechsler Memory Scale: Comparison of utility in discriminating neurologically impaired from controls. Journal of clinical psychology. 1990; 46(2):190–3. PMID: 2324303

33. Ruff R, Light R, Parker S, Levin H. Benton controlled oral word association test: Reliability and updated norms. Archives of Clinical Neuropsychology. 1996; 11(4):329–38. PMID: 14588937

34. Bayles KA. TC. The Functional Linguistic Communication Inventory. Tucson, AZ: Canyonlands Publishing; 1994.

35. McGilton KS, Rochon E, Sidani S, Shaw A, Ben-David BM, Saragosa M, et al. Can We Help Care Providers Communicate More Effectively With Persons Having Dementia Living in Long-Term Care Homes? Am J Alzheimers Dis Other Demen. 2017; 32(1):41–50. https://doi.org/10.1177/1533317516680899 PMID: 27899433

36. Knopman DS, Ryberg S. A verbal memory test with high predictive accuracy for dementia of the Alzheimer type. Archives of neurology. 1989; 46(2):141–5. PMID: 2916953

37. Woloszyn DB, Murphy SG, Wetzel L, Fisher W. Interrater agreement on the Wechsler Memory Scale-Revised in a mixed clinical population. The Clinical Neuropsychologist. 1993; 7(4):467–71.

38. Lo AH, Humphreys M, Byrne GJ, Pachana NA. Test–retest reliability and practice effects of the Wechsler Memory Scale-III. Journal of neuropsychology. 2012; 6(2):212–31. https://doi.org/10.1111/j.1748-6653.2011.02023.x PMID: 22257421

39. de Carvalho IAM, Mansur LL. Validation of ASHA FACS–Functional Assessment of Communication Skills for Alzheimer Disease Population. Alzheimer Disease & Associated Disorders. 2008; 22(4):375–81.

40. Blessed G, Tomlinson BE, Roth M. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. The British Journal of Psychiatry. 1968.

41. Roth M, Tym E, Mountjoy C, Huppert FA, Hendrie H, Verma S, et al. CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. The British journal of psychiatry. 1986; 149(6):698–709.

42. Morris JC. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. International psychogeriatrics. 1997; 9(S1):173–6.

43. Burke WJ, Miller JP, Rubin EH, Morris JC, Coben LA, Duchek J, et al. Reliability of the Washington University clinical dementia rating. Archives of neurology. 1988; 45(1):31–2. PMID: 3337672

44. Cedarbaum JM, Jaros M, Hernandez C, Coley N, Andrieu S, Grundman M, et al. Rationale for use of the Clinical Dementia Rating Sum of Boxes as a primary outcome measure for Alzheimer's disease clinical trials. Alzheimer's & Dementia. 2013; 9(1):S45–S55.

45. Sunderland T, Hill JL, Mellow AM, Lawlor BA, Gundersheimer J, Newhouse PA, et al. Clock drawing in Alzheimer's disease. Journal of the American Geriatrics Society. 1989; 37(8):725–9. PMID: 2754157

46. Morris J, Heyman A, Mohs R, Hughes J, Van Belle G, Fillenbaum G, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assesment of Alzheimer's disease. Neurology. 1989; 39(9):1159–. PMID: 2771064

47. Macartney-Filgate MS, Vriezen ER. Intercorrelation of clinical tests of verbal memory. Archives of Clinical Neuropsychology. 1988; 3(2):121–6. PMID: 14591264

48. Overall JE, Schaltenbrand R. The SKT neuropsychological test battery. Topics in geriatrics. 1992; 5 (4):220–7.

49. Baek MJ, Kim HJ, Ryu HJ, Lee SH, Han SH, Na HR, et al. The usefulness of the story recall test in patients with mild cognitive impairment and Alzheimer's disease. Aging, Neuropsychology, and Cognition. 2011; 18(2):214–29.

50. Reitan R. The Relation of the Trail Making Test to Organic Brain Damage1. Group. 1955; 3(1):1.

51. Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. Perceptual and motor skills. 1958; 8(3):271–6.

52. Oswald WD, Fleischmann UM. Psychometrics in aging and dementia: advances in geropsychological assessments. Archives of gerontology and geriatrics. 1985; 4(4):299–309. PMID: 3833084

53. BUCKS RS, Ashworth D, Wilcock G, Siegfried K. Assessment of activities of daily living in dementia: development of the Bristol Activities of Daily Living Scale. Age and ageing. 1996; 25(2):113–20. PMID: 8670538

54. Erder MH, Wilcox TK, Chen W-H, O'Quinn S, Setyawan J, Saxton J. A new measure of caregiver burden in Alzheimer's disease: The caregiver-perceived burden questionnaire. American Journal of Alzheimer's Disease & Other Dementias®. 2012; 27(7):474–82.

55. Homma A, Takeda M, Imai Y, Udaka F, Hasegawa K, Kameyama M, et al. Clinical efficacy and safety of donepezil on cognitive and global function in patients with Alzheimer's disease. A 24-week, multicenter, double-blind, placebo-controlled study in Japan. E2020 Study Group. Dement Geriatr Cogn Disord. 2000; 11(6):299–313. https://doi.org/10.1159/000017259 PMID: 11044775

56. 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(12):914–9.

57. Spiegel R, Brunner C, Ermini-Fünfschilling D, Monsch A, Notter M, Puxty J, et al. A New Behavioral Assessment Scale for Geriatric Out-and In-Patients: the NOSGER (Nurses' Observation Scale for Geriatric Patients). Journal of the American Geriatrics Society. 1991; 39(4):339–47. PMID: 2010583

58. Albert S, Castillo-Castaneda C, Sano M, Jacobs D, Marder K, Bell K, et al. Quality of life in patients with Alzheimer's disease as reported by patient proxies. Journal of the American Geriatrics Society. 1996; 44(11):1342–7. PMID: 8909350

59. Galasko D, Schmitt F, Thomas R, Jin S, Bennett D, Ferris S. Detailed assessment of activities of daily living in moderate to severe Alzheimer's disease. Journal of the International Neuropsychological Society. 2005; 11(4):446–53. PMID: 16209425

60. Manero R, Casals-Coll M, Sánchez-Benavides G, Rodríguez-de Los Reyes O, Aguilar M, Badenes D, et al. Diagnostic Validity of the Alzheimer's Disease Functional Assessment and Change Scale in Mild Cognitive Impairment and Mild to Moderate Alzheimer's Disease. Dementia and geriatric cognitive disorders. 2014; 37(5–6):366–75. https://doi.org/10.1159/000350800 PMID: 24556708

61. Richards SH, Peters TJ, Coast J, Gunnell DJ, Darlow M-A, Pounsford J. Inter-rater reliability of the Barthel ADL Index: how does a researcher compare to a nurse? Clinical rehabilitation. 2000; 14 (1):72–8. https://doi.org/10.1191/026921500667059345 PMID: 10688347

62. Fricke J, Unsworth CA. Inter-rater reliability of the original and modified Barthel Index, and a comparison with the Functional Independence Measure. Australian Occupational Therapy Journal. 1997; 44 (1):22–9.

63. Feldman H, Sauter A, Donald A, Gelinas I, Gauthier S, Torfs Ka, et al. The disability assessment for dementia scale: a 12-month study of functional ability in mild to moderate severity Alzheimer disease. Alzheimer Disease & Associated Disorders. 2001; 15(2):89–95.

64. Sclan SG, Reisberg B. Functional assessment staging (FAST) in Alzheimer's disease: reliability, validity, and ordinality. International psychogeriatrics. 1992; 4(3):55–69.

65. Loewenstein DA, Amigo E, Duara R, Guterman A, Hurwitz D, Berkowitz N, et al. A new scale for the assessment of functional status in Alzheimer's disease and related disorders. Journal of Gerontology. 1989; 44(4):P114–P21. PMID: 2738312

66. Hilsenroth MJ, Ackerman SJ, Blagys MD, Baumann BD, Baity MR, Smith SR, et al. Reliability and validity of DSM-IV axis V. American Journal of Psychiatry. 2000; 157(11):1858–63. https://doi.org/10.1176/appi.ajp.157.11.1858 PMID: 11058486

67. Stolee P, Stadnyk K, Myers AM, Rockwood K. An individualized approach to outcome measurement in geriatric rehabilitation. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences. 1999; 54(12):M641–M7.

68. Teunisse S, Derix MM, Van Crevel H. Assessing the severity of dementia: patient and caregiver. Archives of Neurology. 1991; 48(3):274–7. PMID: 2001184

69. DeJong R, Osterlund O, Roy G. Measurement of quality-of-life changes in patients with Alzheimer's disease. Clinical therapeutics. 1989; 11(4):545–54. PMID: 2776169

70. Zarit SH, Reever KE, Bach-Peterson J. Relatives of the impaired elderly: correlates of feelings of burden. The gerontologist. 1980; 20(6):649–55. PMID: 7203086

71. Zarit SH, Anthony CR, Boutselis M. Interventions with care givers of dementia patients: comparison of two approaches. Psychology and aging. 1987; 2(3):225. PMID: 3268213

72. Mack JL, Patterson MB, Tariot PN. Behavior Rating Scale for Dementia: development of test scales and presentation of data for 555 individuals with Alzheimer's disease. Journal of geriatric psychiatry and neurology. 1999; 12(4):211–23. https://doi.org/10.1177/089198879901200408 PMID: 10616870

73. Patterson MB, Schnell AH, Martin RJ, Mendez MF, Smyth KA, Whitehouse PJ. Assessment of behavioral and affective symptoms in Alzheimer's disease. Journal of geriatric psychiatry and neurology. 1990; 3(1):21–30. PMID: 2346584

74. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. Journal of psychiatric research. 1983; 17(1):37–49.

75. Starkstein SE, Mayberg HS, Preziosi TJ, Andrezejewski P, Leiguarda R, Robinson RG. Reliability, validity, and clinical correlates of apathy in Parkinson's disease. The Journal of neuropsychiatry and clinical neurosciences. 1992; 4(2):134–9. https://doi.org/10.1176/jnp.4.2.134 PMID: 1627973

76. Patterson MB, Mack JL, Mackell JA, Thomas R, Tariot P, Weiner M, et al. A Longitudinal Study of Behavioral Pathology Across Five Levels of Dementia Severity in Alzheimer's Disease: The CERAD Behavior Rating Scale for Dementia. Alzheimer Disease & Associated Disorders. 1997; 11:40–4.

77. Wood S, Cummings JL, Hsu M-A, Barclay T, Wheatley MV, Yarema KT, et al. The use of the neuropsychiatric inventory in nursing home residents: characterization and measurement. The American Journal of Geriatric Psychiatry. 2001; 8(1):75–83.

78. Sclan SG, Saillon A, Franssen E, Hugonot-Diener L, Saillon A, Reisberg B. THE BEHAVIOR PATHOLOGY IN ALZHEIMER'S DISEASE RATING SCALE (BEHAVE-AD): RELIABILITY AND ANALYSIS OF SYMPTOM CATEGORY SCORES. International Journal of Geriatric Psychiatry. 1996; 11(9):819–30.

79. Kaufer DI, Cummings JL, Ketchel P, Smith V, MacMillan A, Shelley T, et al. Validation of the NPI-Q, a brief clinical form of the Neuropsychiatric Inventory. The Journal of neuropsychiatry and clinical neurosciences. 2000; 12(2):233–9. https://doi.org/10.1176/jnp.12.2.233 PMID: 11001602

80. Overall JE, Gorham DR. The brief psychiatric rating scale. Psychological reports. 1962; 10(3):799–812.

81. Crippa J, Sanches R, Hallak J, Loureiro S, Zuardi A. A structured interview guide increases Brief Psychiatric Rating Scale reliability in raters with low clinical experience. Acta Psychiatrica Scandinavica. 2001; 103(6):465–70. PMID: 11401662

82. Flemenbaum A, Zimmermann RL. Inter-and intra-rater reliability of the Brief Psychiatric Rating Scale. Psychological Reports. 1973; 33(3):783–92.

83. Gabbard GO, Coyne L, Kennedy LL, Beasley C, Deering CD, Schroder P, et al. Interrater reliability in the use of the Brief Psychiatric Rating Scale. Bulletin of the Menninger Clinic. 1987; 51(6):519. PMID: 3427257

84. Ownby RL, Koss E, Smyth KA, Whitehouse PJ. The factor structure of the Brief Psychiatric Rating Scale in Alzheimer's disease. Journal of geriatric psychiatry and neurology. 1994; 7(4):245–50. https://doi.org/10.1177/089198879400700410 PMID: 7826495

85. Weiner MF, Koss E, Patterson M, Jin S, Teri L, Thomas R, et al. A comparison of the Cohen-Mansfield agitation inventory with the cerad behavioral rating scale for dementia in community-dwelling persons with Alzheimers disease. Journal of psychiatric research. 1998; 32(6):347–51. PMID: 9844950

86. Kang SJ, Choi SH, Lee BH, Jeong Y, Hahm DS, Han IW, et al. Caregiver-administered neuropsychiatric inventory (CGA-NPI). Journal of geriatric psychiatry and neurology. 2004; 17(1):32–5. https://doi.org/10.1177/089198873258818 PMID: 15018695

87. COLE MG. Inter-rater reliability of the Crichton geriatric behavioural rating scale. Age and ageing. 1989; 18(1):57–60. PMID: 2631691

88. Cohen-Mansfield J, Marx MS, Rosenthal AS. A description of agitation in a nursing home. Journal of gerontology. 1989; 44(3):M77–M84. PMID: 2715584

89. Rosen J, Burgio L, Kollar M, Cain M, Allison M, Fogleman M, et al. The Pittsburgh Agitation Scale: a user-friendly instrument for rating agitation in dementia patients. The American Journal of Geriatric Psychiatry. 1995; 2(1):52–9.

90. 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(12):2315–. PMID: 7991118

91. Bandelow B, Baldwin DS, Dolberg OT, Andersen HF, Stein DJ. What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? The Journal of clinical psychiatry. 2006; 67(9):1428–34. PMID: 17017830

92. Schneider LS, Olin JT, Doody RS, Clark CM, Morris JC, Reisberg B, et al. Validity and reliability of the Alzheimer's Disease Cooperative Study-Clinical Global Impression of Change. Alzheimer Disease & Associated Disorders. 1997; 11:22–32.

93. Dahlke F, Lohaus A, Gutzman H. Reliability and clinical concepts underlying global judgments in dementia: implications for clinical research. Psychopharmacology Bulletin. 1992.

94. Feldman H, Gauthier S, Hecker J, Vellas B, Emir B, Mastey V, et al. Efficacy of donepezil on maintenance of activities of daily living in patients with moderate to severe Alzheimer's disease and the effect on caregiver burden. Journal of the American Geriatrics Society. 2003; 51(6):737–44. PMID: 12757558

95. Pearlin LI, Mullan JT, Semple SJ, Skaff MM. Caregiving and the stress process: An overview of concepts and their measures. The gerontologist. 1990; 30(5):583–94. PMID: 2276631

96. Bråne G, Gottfries C, Winblad B. The Gottfries-Bråne-Steen scale: validity, reliability and application in anti-dementia drug trials. Dementia and geriatric cognitive disorders. 2001; 12(1):1–14. https://doi.org/10.1159/000051230 PMID: 11125236

97. Gottfries C-G, Bråne G, Gullberg B, Steen G. A new rating scale for dementia syndromes. Archives of gerontology and geriatrics. 1982; 1(4):311–21. PMID: 7186327

98. Vitaliano PP, Russo J, Young HM, Becker J, Maiuro RD. The screen for caregiver burden. The Gerontologist. 1991; 31(1):76–83. PMID: 2007478

99. Demers L, Oremus M, Perrault A, Wolfson C. Review of outcome measurement instruments in Alzheimer's disease drug trials: introduction. Journal of geriatric psychiatry and neurology. 2000; 13(4):161–9. https://doi.org/10.1177/089198870001300401 PMID: 11128056

100. Perrault A, Oremus M, Demers L, Vida S, Wolfson C. Review of outcome measurement instruments in Alzheimer's disease drug trials: psychometric properties of behavior and mood scales. Journal of geriatric psychiatry and neurology. 2000; 13(4):181–96. https://doi.org/10.1177/089198870001300403 PMID: 11128058

101. Demers L, Oremus M, Perrault A, Champoux N, Wolfson C. Review of outcome measurement instruments in Alzheimer's disease drug trials: psychometric properties of functional and quality of life scales. Journal of geriatric psychiatry and neurology. 2000; 13(4):170–80. https://doi.org/10.1177/089198870001300402 PMID: 11128057

102. Oremus M, Perrault A, Demers L, Wolfson C. Review of outcome measurement instruments in Alzheimer's disease drug trials: psychometric properties of global scales. Journal of geriatric psychiatry and neurology. 2000; 13(4):197–205. https://doi.org/10.1177/089198870001300404 PMID: 11128059

103. Newman JC, Feldman R. Copyright and open access at the bedside. N Engl J Med. 2011; 365 (26):2447–9. https://doi.org/10.1056/NEJMp1110652 PMID: 22204721