



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

A systematic review of the validity and reliability of apathy scales in neurodegenerative conditions

Citation for published version:

Radakovic, R, Harley, C, Abrahams, S & Starr, JM 2014, 'A systematic review of the validity and reliability of apathy scales in neurodegenerative conditions', *International Psychogeriatrics*, vol. 27, no. 6, pp. 903-923.
<https://doi.org/10.1017/S1041610214002221>

Digital Object Identifier (DOI):

[10.1017/S1041610214002221](https://doi.org/10.1017/S1041610214002221)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

International Psychogeriatrics

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



**A systematic review of the validity and reliability of apathy scales in
neurodegenerative conditions**

Ratko Radakovic^{* a, b, c, d}, Catharine Harley^e, Sharon Abrahams^{a, b, d} and John M.
Starr^c

^a *Psychology- School of Philosophy, Psychology & Language Sciences, University of
Edinburgh, UK*

^b *Anne Rowling Regenerative Neurology Clinic, University of Edinburgh, UK*

^c *Alzheimer Scotland Dementia Research Centre, University of Edinburgh, UK*

^d *Euan MacDonald Centre for MND Research, University of Edinburgh, UK*

^e *Western General Hospital, University of Edinburgh, UK*

* Correspondence should be addressed to: Ratko Radakovic, University of Edinburgh,
Department of Psychology, 7 George Square, Edinburgh, UK, EH8 9JZ Tel.
+441316509867

Email addresses: r.radakovic@sms.ed.ac.uk, radakovic.ratko@gmail.com

Abstract

Background: There are several scales used to detect apathy in disease populations. Since apathy is a prevalent symptom in many neurodegenerative diseases, this is an especially important context in which to identify and compare scales.

Aims: To provide an overview of apathy scales validated in generic and specific neurodegenerative disease populations, compare validation studies' methodological quality and the psychometric properties of the validated apathy scales.

Methods: A systematic review of literature was conducted of articles published between 1980 and 2013. The final articles selected for review were rated on methodological quality and the psychometric properties of the scales used were interpreted.

Results: Sixteen articles validating apathy scales were included in the review, five in a generic neurodegenerative sample and eleven in specific neurodegenerative samples. The methodological quality of specific studies varied from Poor to Excellent. The highest quality, which had psychometrically favourable scales, were the Dementia Apathy Interview and Rating and the Apathy Evaluation Scale- Clinical version in Alzheimer's Disease and the Lille Apathy Rating Scale in Parkinson's disease. Generic neurodegenerative disease validation studies were of average methodological quality and yielded inconsistent psychometric properties.

Conclusions: Several instruments can be recommended for use in some specific neurodegenerative diseases. Other instruments should either be validated or developed to assess apathy in more generic populations.

Key words: apathy, rating scales, psychogeriatrics, dementia, motor disorders

Introduction

Historically, apathy is reported to have undergone a transformation in its definition. Originally, rooted in *apatheia*, meaning without (*a-*) passion (*-pathos*), it was regarded as a virtue in Greek Stoic philosophy, allowing individuals to be more objective. However, the modern meaning of apathy has come to be regarded as undesirable and, as a part of the psychiatric field, can be a syndrome or symptom relating to disease. Clinically, as a behavior, apathy is most often characterised as a deficit in motivation, which is displayed through reduced interest, emotions and goal directed behaviors (Marin, 1996). Apathy has been found to have a variable relationship with various symptoms and experiences, such as depression and boredom. As a symptom, apathy has a variable association with depression dependent on the neurodegenerative disease (Levy *et al.*, 1998). However, depression is expressed as negative and, in the case of bipolar depression, extremely positive affect (Levy *et al.*, 1998), whereas, apathy is observed as emotional neutrality, where neither positive nor negative emotions are observed. Boredom, on the other hand, is a widespread human experience that is linked, through motivational elements, to apathy. However, research has shown them to be statistically distinct when measured by scales (Goldberg *et al.*, 2011). This is due to apathy scales being designed to measure apathy as a symptom rather than a domain of normal motivational behavior.

Apathy has been frequently observed as a symptom of neurodegenerative diseases (for review see van Reekum, Stuss and Ostrander, 2005; Chase, 2011) and is assessed by a variety of methods. Both clinically and in research, these methods often range from administration of measurement scales to observational ratings of behaviors by an experienced or trained individual. While specific methods of assessing apathy have guidelines to ensure that it is administered in a standardised

fashion, the problem is that there are many different methods to measure apathy. This creates uncertainty of which method would be most appropriate to measure apathy in a given pathological population. The utility of these methods are often based on previous studies, which have validated the measure in a specific pathological population. However, these methods and the studies validating them are seldom compared within literature.

Since apathy is a common symptom in neurodegenerative disease, this is an important area in which to produce a comprehensive review of how apathy is measured both generally and for specific neurodegenerative diseases. It would also be beneficial to compare the validity and reliability of assessment methods, so that the highest quality, well-validated and reliable method can be chosen for a specific patient group.

Apathy in neurodegenerative disease:

Apathy is common in all types of dementia (Clarke *et al.*, 2008) and most prominent in both Alzheimer's disease (AD) and vascular dementia (VaD; Jonsson *et al.*, 2010). In both small and large vessel VaD, apathy is a consistent symptom with negative prognostic and treatment implications (Staekenborg *et al.*, 2010). In AD, apathy was found to occur in 61% to 92% of patients (e.g. Landes *et al.*, 2005) with an almost equally high prevalence in frontotemporal dementia (FTD; e.g. Mendez *et al.*, 2008). However, recent research has found differing apathy profiles which can distinguish between AD and FTD (Quaranta *et al.*, 2012). This distinction between AD and FTD was also observed in functional impairments of activities of daily living, which could be associated with apathy level (Mioshi *et al.*, 2009).

Apathy is also a prominent symptom in amyotrophic lateral sclerosis (ALS; Grossman *et al.*, 2007; Girardi *et al.*, 2011), the most common form of motor neurone disease (Goldstein and Abrahams, 2013). A large cohort study found that apathy was the most common behavioral symptom in ALS, occurring in 31% of patients (Witgert *et al.*, 2010). Recent neuroimaging research in ALS has shown evidence of neuroanatomical correlates relating to apathy of abnormalities in the anterior cingulate cortex (Woolley *et al.*, 2011). In a recent review of social cognition in ALS apathy was observed as the most prevalent, overlapping behavioral change in both ALS and FTD (Abrahams, 2011).

In other motor disorders such as Parkinson's disease (PD), over a third of patients have been found to exhibit apathy (Pedersen *et al.*, 2009). Apathy may be predictive of cognitive decline and even dementia in PD (Dujardin *et al.*, 2009) but may also have marked variability in its relation to the clinical presentation of PD (Dujardin *et al.*, 2007).

Apathy is also a persistent and progressive symptom across all stages of Huntington's disease (HD; Thompson *et al.*, 2012) with a variable prevalence reported between 34% and 76% associated with stage of disease (van Duijn *et al.*, 2007). A study comparing HD gene mutation carriers versus non-carriers found that the latter had 0% prevalence compared to carriers who had 32% prevalence (van Duijn *et al.*, 2010). Functional decline was also found to be greater in HD patients with apathy, independent of motor or cognitive symptoms (Hamilton *et al.*, 2003).

Due to this variable prevalence of apathy in these neurodegenerative diseases, this review focuses on apathy as a symptom of these specific diseases.

Emerging concepts:

Apathy is conceptualised by the measurement scales associated with it. It is originally defined by a substructure of cognition, behavior and emotion (Marin *et al.*, 1991) but is measured as singular concept, with specified clinical cut-offs dependent on the scale.

A previous review by Clarke *et al.* (2011) looked at all the different measurement methods to assess apathy in research and clinical practice and found that most used rating scales or scale-based interviews. These scales were self-assessments, informant-based or clinical semi-structured interviews for patients or informants. The studies reviewed included a range of pathological populations with variable aetiologies (Clarke *et al.*, 2011). However, Clarke *et al.*'s review did not directly examine apathy as a symptom of specific neurodegenerative diseases and was not a systematic review of literature. The present review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher *et al.*, 2009) criteria to methodologically identify the quality of existent apathy scales and define which scales have been validated in neurodegenerative diseases. As apathy occurs as a symptom in a majority of neurodegenerative diseases (for review see van Reekum, Stuss and Ostrander, 2005; Chase, 2011) it is important to examine methods of apathy assessment that can identify apathy as a symptom in any neurodegenerative disease. Furthermore, as previously mentioned, patients with specific neurodegenerative diseases, such as HD and AD, exhibit apathy in variable prevalence. So it is important

to classify scales that were specifically designed, and therefore only validated, in specific neurodegenerative disease so that they may be used as such in the future. This will allow for identification of apathy scales that are neurodegenerative disease generic and specific. This will result in recommendations for suitable apathy scale choices to be made in future research and clinical practice in a disease group where apathy is a prevalent symptom.

Quality, Validity and Reliability:

The methodology and the measures used within a research study are important determinants of article quality. Individual articles employ different patient groups, inclusion and exclusion criteria, index tests and reference or gold standards of diagnosis, all of which are described with different levels of detail. Due to this variation, we considered it important to use a standardized tool for quality assessment of each study.

In addition, we considered the psychometric properties reported by each study of the specific apathy instrument, which are commonly presented in terms of validity and reliability. Validity is best described as accuracy of an instrument, and can be defined through content, concurrent, convergent, congruent and discriminant or divergent. Content validity relates to the design of a measure and whether the instrument is measuring what it is aiming to measure. Concurrent, convergent or congruent validity are often grouped together as measuring construct validity and measure whether the instrument correlates with other well-validated methods of detection. This method of detection is known as a gold or reference standard. In terms of apathy, a gold standard would be the diagnosis of apathy by one or two trained

psychiatrists via observation, a structured or semi structured clinical interview.

Discriminant or divergent validity is also an important psychometric property and is where two components, or constructs, that are theoretically unrelated are actually shown to dissociate. With symptomatic apathy, this behavioral component is often depression (Levy *et al.*, 1998), which is due the emotional nature of patients with depression and, conversely, the emotionally neutral nature of apathetic patients. This apathetic component also differentiates it from exhilarating happiness observed in elation, the uncomfortable uneasiness of anxiety and excessive agitation associated with irritability. These behavioral components can also be used to determine discriminant validity.

Reliability is best defined as the consistency of the measure. This relates to the replicability of findings independent of how accurate they are. This can be determined through between-item correlations of the specific measure, which is known as internal consistency reliability and usually measured by Cronbach's alpha (Cronbach, 1951). Test-retest reliability is another form where the relationship between responses to the measure over several different time periods is recorded and correlated. Another form of reliability is inter-rater based, which examines whether two, or more informants or interviewers responses or scores agree.

Both validity and reliability are important determinants of an assessment scale's utility and a good apathy measure should embody both of these characteristics measured in the same sample. Therefore, the focus of this review will be on studies that examine both the validity and reliability of scales. However, there are no gold

standards of interpretation regarding validity and reliability psychometric statistics, therefore previously used guidelines were adopted to categorise them.

Study aims:

Based on the above points and issues, the aims of this review are:

1. To determine apathy assessment scales that have been validated in neurodegenerative populations
2. To determine neurodegenerative generic and neurodegenerative specific apathy assessment scales
3. To quantify the methodological quality of the apathy assessment validation studies
4. To compare reliability and validity statistics of each apathy assessment scale as to establish suitable neurodegenerative generic and specific apathy assessment methods

Methods

This systematic review was conducted following the PRISMA guidelines (Moher *et al.*, 2009). The completed checklist can be found in the Supplementary materials (see figure S1).

Eligibility criteria

Only articles, abstracts and letters limited to English language publications between 1980 and 2013 containing human participants, aged 18 or older were screened.

Information Sources

The main search was conducted on 31 December 2013. The electronic databases used were EMBASE, PSYCHINFO, MEDLINE and PUBMED.

Search criteria

Through the use of Boolean operators, the database searches were restricted to an apathy term ('Apathy' or 'Motivation' or 'Amotivation'), neurodegenerative disease term ('Dementia' or 'Alzheimer Disease' or 'Alzheimer's Disease' or 'Multiinfarct Dementia' or 'Vascular Dementia' or 'Frontotemporal Dementia' or 'Semantic Dementia' or 'Parkinson Disease' or 'Parkinson's Disease' or 'Motor Neuron Disease' or 'Amyotrophic Lateral Sclerosis' or 'Huntingtons Disease' or 'Huntingtons Chorea'), a scale term ('Rating Scale' or 'Rating Scales' or 'Summed Rating Scale' or 'Psychological Rating Scale' or 'Psychiatric Status Rating Scale' or 'Questionnaire' or 'Questionnaires' or 'Measurement') and a validity or reliability term ('Reliability' or 'Validity' or 'Test Reliability' or 'Interrater Reliability' or 'Statistical Reliability' or 'Test Validity' or 'Statistical Validity').

Study selection

After applying the above mentioned search criteria and eligibility criteria, duplicate articles were removed. Study selection was conducted in two main stages following the electronic search. In the first stage, two authors (R.R. and C.H.) went through the titles and abstracts independently, selecting which abstracts to include in the next stage, exclude from the next stage and which were ambiguous or disagreed upon by the aforementioned two authors. Inclusion was based on the following: study

involved a neurodegenerative disease population, used scalar or quantifiable methods of assessing apathy levels, and recorded both validity and reliability statistics. The ambiguous or disagreed upon abstracts were then forwarded to the third author (J.S.) who made suggestions, of inclusion, exclusion or examination of whole articles.

In the second and final stage, the selected abstracts were then examined as full articles by authors R.R. and C.H., independently. These were classed as include in the review, exclude from the review and ambiguous or disagreed upon by authors R.R. and/or C.H. Articles that were included in the review had to meet the same criteria as in the first stage but authors were additionally asked to record the neurodegenerative disease assessed, apathy assessment methods used and validity and reliability statistics produced. In all cases, the authors had to provide reasons for exclusion or ambiguity. Advice from the third author, J.S., was sought once again for ambiguous or disagreed upon articles. A consensus had to be reached for articles that were to be used in the review.

Neurodegenerative disease generic and specific study grouping

The final articles were then grouped based on whether the study population assessed for apathy was neurodegenerative disease generic (e.g. containing a mixture of AD and PD patients) or specific (e.g. containing only AD patient) as to determine the utility in each group.

Quality Assessment

The final articles selected for review were subjected to the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) checklist (Whiting *et al.*, 2003, see Table

1). This was used to determine the quality of each article. The target condition was apathy, the index test was the method by which apathy was assessed and the reference standard was the method by which apathy was diagnosed and, therefore, the validation method for the index test.

Items 1 to 5, 7 to 9 and 13 to 14 from the QUADAS were found to be applicable for determining quality of methodology and used for this particular review (see Table 1). Some QUADAS items (6 and 10 to 12) were not included due to their lack of applicability for assessing bias in psychiatric symptom scale validation studies, such as apathy. Items 1 and 2 relate to variability of the studies in the sample assessed and which impacts on the generalizability of findings to the study population. Items 3 to 5, 7 and 14 examine methodological rigour and bias, which influences the validity of the study results and, therefore, the validity of the scale. Finally, items 8, 9 and 13 determined the quality of reporting of methodology. Items were answered 'yes', 'no' or 'unsure' after close examination of the articles by author R.R.

Table 1. QUADAS items for quality review

Items	
1	Was the spectrum of patients representative of the patients who will receive the test in practice?
2	Were selection criteria clearly described?
3	Is the reference standard likely to correctly classify the target condition?
4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?
8	Was the execution of the index test described in sufficient detail to permit replication of the test?
9	Was the execution of the reference standard described in sufficient detail to permit its replication?
13	Were uninterpretable/ intermediate test results reported?
14	Were withdrawals from the study explained?
Target condition= Apathy; Index test= Apathy assessment method; Reference standard= Apathy diagnosis method	

There is no standardized scoring system for the QUADAS. Therefore, we categorised each study based on the number of QUADAS items that were answered 'yes' to. The categorisation was as follows; 9 or 10 items was classed as Excellent, 7 or 8 was Good, 4 to 6 was Adequate, 2 to 3 items was Poor and 0 to 1 was considered Unacceptable.

Interpretation of psychometric properties

We have provided a table of guidelines for interpreting internal consistency and correlation coefficients in relation to validity and reliability statistics (see table S1). This was adapted based on Hermans, van der Pas and Evenhuis' (2011) review paper of anxiety scales. However, we have updated the interpretation of internal

consistency (George and Mallery, 2003) and Kappa statistic (Viera and Garrett, 2005).

Results

Systematic review

Figure 1 illustrates the review process in a flowchart. From the databases listed and using the search criteria specified earlier, a total of 168 studies were identified. When the eligibility criteria were applied prior to the first stage, 21 articles were removed, leaving 147 articles. After identifying and removing 31 duplicate articles, this left 116 studies eligible for the first stage of study selection; title and abstract screening. These included a variety of neurodegenerative populations that assessed apathy using various methods.

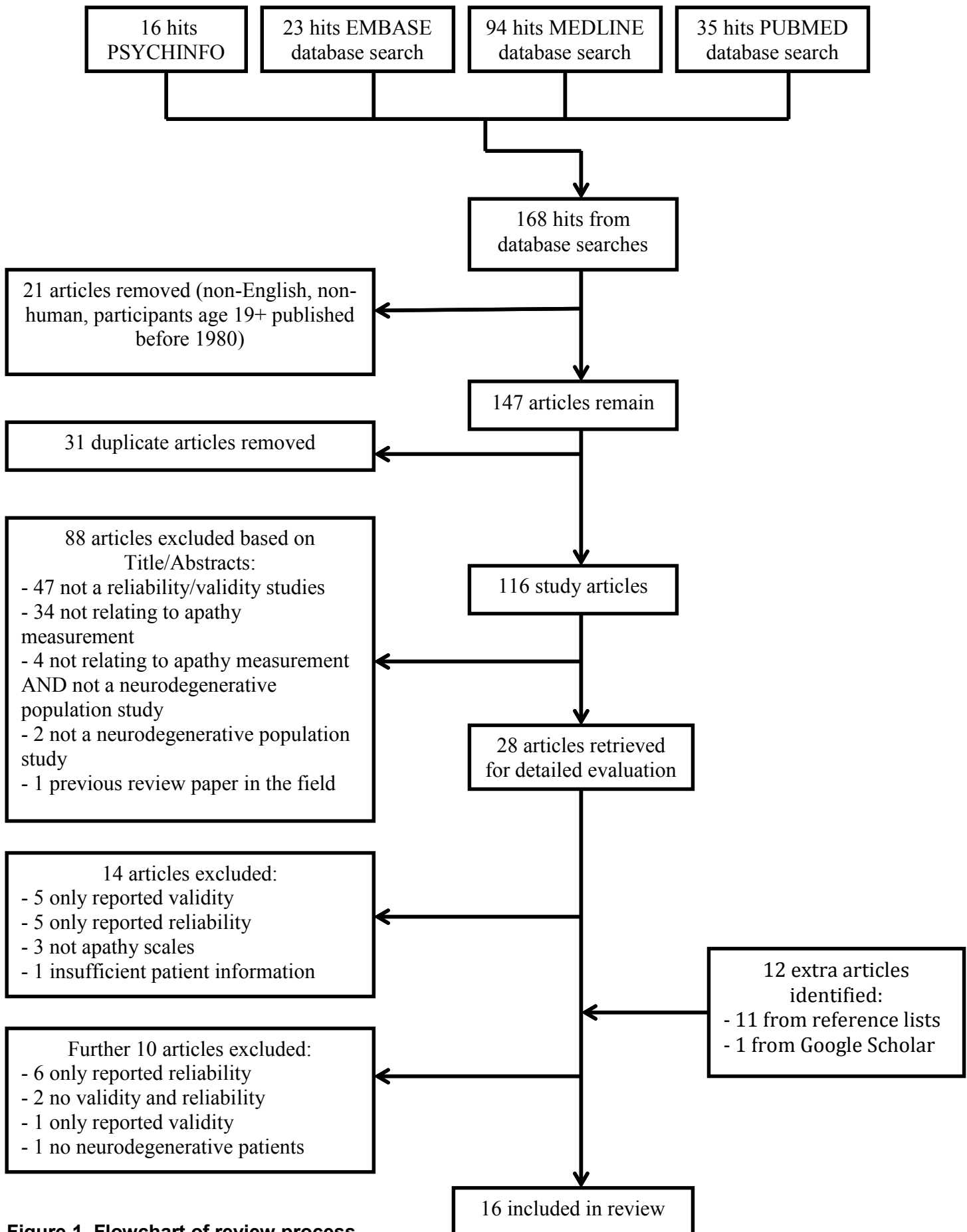


Figure 1. Flowchart of review process

In the first stage of study selection, of the 116 studies, 88 were not relevant or excluded: 47 were not reliability/validity studies, 34 were not studies that related to apathy measurement, four were not studies that related to apathy measurement and were not conducted in a neurodegenerative population, two were not conducted in a neurodegenerative population and one was a previous review paper in the field (Clarke *et al.*, 2008). This left 28 studies to be included for detailed evaluation. A further 12 articles were identified through Google Scholar and reference lists and included for detailed evaluation, resulting in a total of 40 studies for detailed evaluation.

The second stage of study selection, of the 40 studies, 24 articles were excluded: 11 only reported validity for the apathy measurement, six only reported reliability for the apathy measurement, three were not apathy scales, two did not report validity and reliability of the apathy measurement, one had insufficient patient information and one was not conducted in a neurodegenerative population.

Table 2 shows, in total, 16 studies were included in the review. Seven of these articles validated English versions of a scale (Starkstein *et al.* 1992, Cummings *et al.* 1994, Kaufer *et al.* 2000, Strauss *et al.* 2002, Clarke *et al.* 2007, Zahodne *et al.* 2009, de Medeiros *et al.* 2010).

The review of methodological quality of each study by the chosen QUADAS criteria is shown in Table 2. There was a range of Poor to Excellent quality studies with the modal quality rating being Adequate.

No studies stated explanations for patient withdrawals; a large proportion of which were unclear about their withdrawals and the remainder did not state any explanations. However, almost all of the studies used a reference standard that was independent of index test for their validation study (see figure S2). There were five studies that examined apathy scales or subscales in a generic neurodegenerative disease sample and 11 that examined them in specific neurodegenerative disease samples.

Table 2. Methodological quality ratings of review studies

Author (Year)	Instrument	Generic/ Specific	1	2	3	4	5	7	8	9	13	14	Overall Quality
Clarke et al. (2007)	Self, Informant and Clinical Apathy Evaluation Scale (AES-S/I/C)	Generic	N	U	Y	N	Y	Y	Y	Y	N	N	Adequate
Cummings et al. (1994)	Neuropsychiatric Inventory- Apathy (NPIa)	Generic	Y	Y	U	U	U	Y	Y	U	U	N	Adequate
de Medeiros et al. (2010)	Neuropsychiatric Inventory- Clinical Rating Scale- Apathy (NPIa-C)	Specific	Y	Y	Y	N	Y	Y	U	Y	U	U	Adequate
Dujardin et al. (2008)	Informant based Lille Apathy Rating Scale (LARS-i)	Specific	Y	U	Y	U	Y	Y	Y	Y	U	N	Adequate
Hsieh et al. (2012)	Clinical Apathy Evaluation Scale (AES)	Specific	Y	Y	Y	U	Y	Y	Y	Y	U	N	Good
Kaufer et al. (2000)	Brief Clinical Neuropsychiatric Inventory- Apathy (NPIa-Q)	Specific	U	U	Y	U	Y	N	Y	Y	U	U	Adequate
Leontjevas et al. (2012)	Abbreviated Apathy Evaluation Scale (AES-10)	Generic	N	U	Y	N	Y	N	Y	Y	U	U	Adequate
Lueken et al. (2007)	Abbreviated Apathy Evaluation Scale (AES-10)	Generic	N	U	Y	N	Y	N	Y	Y	U	U	Adequate
Malakouti et al. (2012)	Farsi Neuropsychiatric Inventory- Apathy (F-NPIa)	Specific	N	Y	N	U	U	Y	Y	Y	U	U	Adequate
Pedersen et al. (2012)	Starkstein Apathy Scale (SAS)	Specific	Y	N	N	U	N	U	Y	U	Y	U	Poor
Politis et al. (2004)	Hellenic Neuropsychiatric Inventory- Apathy (H-NPIa)	Specific	Y	Y	N	N	Y	Y	Y	Y	Y	U	Good
Robert et al. (2002)	Apathy Inventory (AI)	Generic	Y	Y	U	U	Y	Y	Y	Y	U	N	Adequate
Sockeel et al. (2006)	Lille Apathy Rating Scale (LARS)	Specific	Y	Y	Y	U	Y	Y	Y	U	Y	U	Good
Starkstein et al. (1992)	Apathy Scale (SAS)	Specific	Y	N	U	N	N	Y	Y	N	N	U	Poor
Strauss et al. (2002)	Dementia Apathy Interview and Rating (DAIR)	Specific	Y	Y	Y	Y	Y	Y	Y	Y	Y	U	Excellent
Zahodne et al. (2009)	Lille Apathy Rating Scale (LARS)	Specific	U	Y	Y	U	Y	Y	Y	Y	U	U	Adequate

Generic neurodegenerative disease scales

Table 3. Apathy scale psychometric properties in generic neurodegenerative disease studies

Apathy Measure	Author (Year)	Type of measure	Patient Group	Patient N	Internal Consistency Reliability	Validity Statistic	Reliability Statistic	Quality Rating
AES-S/I/C	Clarke et al. (2007)	Self Report, Informant Report and Clinical Patient Interview	55.2% AD, 14.3% mixed dementia (possible AD and possible lewy body dementia), 9.5% lewy body dementia, 5.7% VaD, 5.7% mixed dementia (AD and VaD), 4.8% FTD, 4.8% other dementia	121 (diagnosis ascertained for 105 participants)	Cronbach's alpha (of factors derived from FA): AES-S Apathy factor = 0.884 Other factor = 0.410 AES-I Apathy factor = 0.895 Interest factor = 0.881 AES-C Apathy factor = 0.864 Interest factor = 0.908	Convergent (NPIa Frequency x Severity): AES-S $r = 0.22$ ($p < 0.05$) AES-I $r = 0.49$ ($p < 0.01$) AES-C $r = 0.27$ ($p < 0.01$) Discriminant (NPIa): AES-S $r = 0.23$ ($p < 0.01$) AES-I $r = 0.22$ ($p < 0.01$) AES-C $r = 0.18$ (NS) Discriminant (NPIax): AES-S $r = 0.20$ (NS) AES-I $r = 0.30$ (NS) AES-C $r = 0.17$ (NS)	-	Adequate
NPIa	Cummings et al. (1994)	Informant Interview	AD, Probable VaD and other dementias	Concurrent validity study = 40	Cronbach's alpha (Frequency and severity):	Content (Delphi Panel: 10 national and international experts in geriatric psychiatry, behavioural neurology)	Between rater agreement (2 trained raters, from a pool of 4 behavioural	Adequate

				Inter-rater reliability study = 45 Test-retest reliability study = 20	Within range of 0.87 to 0.88	and neuropsychology rating each subdomain between 1 and 4; well assessed and poorly assessed, respectively): NPIa screening question = 1.3 NPIa subquestions = 1.4	neurology fellows, a geriatric psychiatry fellow, premedical student and behavioural neurologist): Frequency- 97.9% Severity- 89.4%	
							Test-retest ("within 3 weeks") Frequency $r = 0.74$ ($p < 0.001$) Severity $r = 0.68$ ($p < 0.01$)	
AES-10	Leontjevas et al. (2012)	Informant Interview	Nursing Home Residents: 42% no dementia, 17% AD, 11% VaD, 8% mixed dementia, 4% other dementia (Korsakov syndrome, Lewy Body dementia, dementia with PD, FTD and HD), 18% non specified dementia	100	Cronbach's -alpha = 0.95 -standardized alpha = 0.95	Congruent (NPIa Frequency x Severity): Spearman's rho = 0.62 ($p < 0.001$) Discriminant (CSDD): Spearman's rho = 0.27 ($p < 0.05$)	-	Adequate
AES-10	Lueken et al. (2007)	Informant Report	Nursing Home Residents: 7% no dementia, 69% AD, 7% VaD, 4% mixed dementia, 1% FTD,	356 (Subsample A = 178, Subsample B = 178)	Subsample A+B Cronbach's alpha = 0.92	Subsample A+B Concurrent (AES-I): $r = 0.97$ Subsample A Concurrent (NPIa): $r = 0.61$	-	Adequate

			1% dementia with PD, 3% other nonclassified dementia, 8% MCI			Discriminant (NPId): $r = 0.07$		
						(GDS): $r = -0.51$		
						Subsample B Concurrent (NP1a): $r = 0.62$		
						Discriminant (NPId): $r = 0.09$		
						(GDS): $r = -0.58$		
AI	Robert et al. (2002)	Informant Interview AND Patient Interview	AD, PD, MCI	115	Informant Interview Cronbach's alpha = 0.92	Concurrent (NP1a Frequency x Severity): Informant interview Emotional Blunting $r = 0.01$ (NS) Lack of Initiative $r = 0.23$ ($p < 0.01$) Lack of Interest $r = 0.63$ ($p < 0.001$) Global score $r =$ not recorded Patient Interview Emotional Blunting $r = 0.04$ (NS) Lack of Initiative $r = 0.01$ (NS) Lack of Interest $r = 0.26$ (NS)	Between rater agreement (26 psychiatrists, psychologists and medical students rating one videotaped response): Kappa coefficient for item scores and global score = 0.99 Test-retest ("same day"): Kappa coefficient emotional blunting = 0.99 Kappa coefficient lack of initiative = 0.97	Adequate

Global score r = not recorded	Kappa coefficient lack of interest = 0.99
Discriminant (NPId Frequency x Severity):	Kappa coefficient global score = 0.99
Informant interview	
Emotional Blunting r = not recorded	
Lack of Initiative r = 0.32 (p<0.05)	
Lack of Interest r = 0.40 (p<0.001)	
Global score r = 0.37 (p<0.01)	
Patient Interview	
Emotional Blunting r = not recorded	
Lack of Initiative r = 0.37 (p<0.01)	
Lack of Interest r = 0.31 (p<0.05)	
Global score r = 0.42 (p<0.001)	

NS= Not Significant; NPId= NPI depression/dysphoria; NPIanx= NPI anxiety; CSDD= Cornell Scale for Depression in Dementia

Table 3. shows the descriptors and quality of apathy scale validation studies in generic neurodegenerative disease patient samples. All five studies were of Adequate methodological quality.

The Apathy Evaluation Scale (AES) was originally developed and validated in English by Marin *et al.* (1991). It is made up of 18 items that assess emotional, behavioral and cognitive aspects of apathy through a 4-point Likert scale. The responses are coded to give one apathy score. The AES is available in Self report (S), Informant report (I) or Clinical patient interview (C) forms. The validation study by Clarke *et al.* (2007) looks at all three forms of the AES in a generic neurodegenerative dementia sample group. Convergent validity was measured through correlations with the NPIa, all of which were positive ranging from little or no correlation for the AES-S and AES-C to a low correlation with the AES-I. Discriminant validity for AES-S, AES-I and AES-C when compared to the NPId showed little or no positive correlation and with the NPIanx, ranging from low positive correlation for the AES-S and C to a low correlation for the AES-I. This study does not provide internal consistency values for each form of the AES. However, it has subdivided each available form of the AES in to two factors each, resulting in numerous internal consistency values. The two factors for the AES-I and AES-C were labelled Apathy and Interest. For the AES-I the Interest factor showed good internal consistency and the Apathy Factor showed excellent internal consistency. The AES-C showed a good internal consistency for the Apathy Factor and an excellent internal consistency for the Interest factor. The AES-C showed a differing factor structure labelled Apathy and Other, with internal consistencies that are good and unacceptable, respectively.

Lueken *et al.* (2007) developed the abbreviated Apathy Evaluation Scale (AES-10) in German as a shortened, 10-item version of the AES-I. The items were developed and, as stated in the study, translated in close cooperation with Dr Robert Marin, the original author of the AES. It is scored almost identically as the original AES, on a 4-point Likert scale. It was validated in a sample of Nursing home residents, of various neurodegenerative diseases, some of who did not have dementia. The internal consistency reliability was excellent. A high positive correlation between the AES-10 and NPIa in both subsample A and B showed good concurrent validity. This was further supported by a very high, positive correlation with the AES-C. The discriminant validity was shown to be good through little or no positive correlation between the AES-10 and the NPId in both subsamples. Furthermore, a moderate, negative correlation was observed between the AES-10 and the GDS.

A more recent study also looked at the Dutch version of the AES-10 in Nursing home residents, almost half of which were non demented (Leontjevas *et al.*, 2012). As with the previous study, the internal consistency was shown to be excellent. For congruent validity, the AES-10 and NPIa showed a moderate positive correlation and when compared to a depression measure (CSDD), there was a little or no positive correlation, showing good discriminant validity.

The Neuropsychiatric Inventory (NPI; Cummings *et al.*, 1994) is a 12 domain Informant interview that assesses behavioral disturbances in various diseases. Originally developed in English, it measures apathy via the NPI apathy/indifference (NPIa) domain firstly with a screening question to determine the presence of these behaviors (Yes/No/NA answers). This is followed by eight subquestions to determine

the characteristics of apathy (Yes/No answers). The frequency of the NPIa domain is rated on a 4-point Likert scale along with a 3-point Likert scale for severity.

Frequency and severity are then multiplied to give a final NPIa domain score. The internal consistency was difficult to determine specifically for the NPIa due to lack of detailed reporting in the Cummings *et al.* (1994) study. However, based on the range of values provided in the article, the internal consistency of the NPIa was good. A Delphi panel determined content validity by rating whether screening question and sub-questions captured the essential elements of behavior. For the NPIa, they rated both as close to being ‘assessed well’. Inter-rater agreement and test-retest reliability for frequency was also found to be high with a further moderate test retest for severity.

The Apathy Inventory (AI; Robert *et al.*, 2002) was developed and validated in French to assess global and subdomain apathy (emotional blunting, lack of initiative and lack of interest). It is a three item scale (one item for each domain) and is has been validated as a Self report and Informant interview. The Self report AI firstly has the patient assess their own behavior for each item (Yes/No) and then asks them to bisect a line in relation to the severity of their behavior that spans from mild to severe on a covert 12 point scale only know to the scorer. These three item severities are then added to give a composite apathy score. The Informant report is similar to the NPIa, recording whether the behaviors are displayed (Yes/No), frequency of behavior on a 4-point Likert scale along and severity of behavior on a 3-point Likert scale. Again, frequency and severity are multiplied to give a numeric apathy level for each subdomain, which are then added to give a composite apathy level. The internal consistency for the informant interview was excellent. The NPIa

generally held little or no to low positive correlations with AI subdomains, with only positive, moderate correlation with the informant Lack of Interest subdomain. There were no record of the correlation between the NPIa and the AI global scores.

Discriminant validity determined by the relationship the NPId and the AI (global and subdomains), which showed low to moderate positive correlations. The between rater and test retest reliability was almost perfect for the global and subdomain scores.

Specific neurodegenerative disease scales

Table 4. Apathy scales psychometric properties in specific neurodegenerative disease patient populations

Apathy Measure	Study	Type of measure	Patient Group	Patient N	Internal Consistency	Validity Statistic	Reliability Statistic	Quality Rating
NPIa-C	de Medeiros et al. (2010)	Informant Interview	Mild to severe Probable AD	128	-	Convergent validity (AES-I): Frequency x Severity Pearson's $r = 0.22$	Inter-rater (ICC; 2 independently trained raters): NPIa-C subquestions range from 0.74 to 0.89	Adequate
LARS-i	Dujardin et al. (2008)	Informant Interview	Probable PD	60	Cronbach's between-item $\alpha = 0.872$ -standardized $\alpha = 0.877$	Concurrent (AES-I): Global score $r = 0.850$ ($p < 0.001$) Concurrent (AES-C): Global score $r = 0.811$ ($p < 0.001$) Concurrent (LARS): Global score $r = 0.814$ ($p < 0.001$)	Inter-rater (ICC, 2 clinician interviewers): Global score = 0.966 ($p < 0.001$) Test-retest ("few days later"): Global score $r = 0.960$	Adequate
AES-C	Hsieh et al. (2012)	Informant Interview	Mild to moderate AD	144	Cronbach's $\alpha = 0.85$	Convergent (NPIa Frequency x Severity): Pearson's $r = 0.61$ ($p < 0.001$) Discriminant (NPId and NPIe, both Frequency x Severity): NPId Pearson's $r = 0.16$ (NS) NPIe Pearson's $r = -0.46$ ($p < 0.001$)	Test-retest ("3 day period"): Pearson's $r = 0.89$	Good
NPIa-Q	Kaufer et al. (2000)	Informant Report	Possible and probable AD	60	-	Convergent (NPIa): Spearman's $\rho = 0.85$ ($p < 0.0001$)	Test-retest ("same day"): Pearson's $r = 0.80$ ($p < 0.0001$)	Adequate

F-NPIa	Malakouti et al. (2012)	Informant Interview	Dementia	100	Cronbach's alpha (severity) = 0.82	Concurrent (PANSS): Frequency x Severity Pearson's r = 0.6-0.7 (p<0.001)	Inter-rater (ICC, 2 trained psychiatrist raters): Frequency Pearson's r = 0.85 (p<0.001) Severity Pearson's r = 0.83 (p<0.001) Frequency x Severity Pearson's r = 0.87 (p<0.001) Test-retest ("two weeks"): Frequency Pearson's r = 0.90 (p<0.001) Severity Pearson's r = 0.90 (p<0.001) Frequency x Severity Pearson's r = 0.76 (p<0.001)	Adequate
SAS	Pedersen et al. (2012)	Self Report	PD	194	Cronbach's alpha = 0.69	Discriminant (MADRS): Spearman's rho = 0.25 (p<0.01)	-	Poor
H-NPIa	Politis et al. (2004)	Informant Interview	AD	29	Cronbach's alpha (Frequency x Severity): Within range of 0.69 to 0.72	Concurrent (BPRS): Frequency x Severity Pearson's r = 0.48 (p<0.001)	-	Good
LARS	Sockeel et al. (2006)	Patient Interview	Probable PD	159	Cronbach's standardized alpha	Concurrent (AES): Global score r = 0.87	Inter-rater (ICC, 2 clinician raters): Global score r = 0.95	Good

SAS	Starkstein et al. (1992)	Self Report	Idiopathic PD	50	<p>-between item = 0.80 -between subscale = 0.74</p> <p>Cronbach's alpha = 0.76</p>	<p>Intellectual curiosity subscale r = 0.84 Action Initiation subscale r = 0.65 Emotion subscale r = 0.44 Self awareness subscale r = 0.15 Convergent (Neurologist blind rating): Rated 6 apathetic and 6 apathetic with apathetic patients having significantly higher SAS score (apathetic SAS mean=14.8, SD=5.7; non-apathetic SAS mean=5.5, SD=2.2; t(10)=3.70, p<0.001) Convergent (Clinical ratings; Nurse, Physician and Neuropsychologist technician): r = 0.40, 0.31 and 0.46, respectively</p> <p>Discriminant (BRSD-d): r = 0.08</p>	<p>Test-retest ("approximately four months"): Global score r = 0.98 Inter-rater (2 raters): r = 0.81 (p<0.01)</p> <p>Test-retest ("1 week apart"): r = 0.90 (p<0.01)</p>	Poor
DAIR	Strauss et al. (2002)	Informant Interview	Possible and probable AD	100	<p>Cronbach's alpha = 0.89</p>	<p>Convergent (SAS): Global score r = 0.62 (p<0.001) Intellectual curiosity subscale r = 0.61 (p<0.001) Action Initiation subscale r = 0.42 (p<0.001) Emotion subscale r = 0.33 (p<0.01) Self awareness subscale r = 0.229 (NS)</p> <p>Inter-method agreement (ICC; SAS): Global score = 0.75</p> <p>Divergent (BDI-II):</p>	<p>Inter-rater (2 trained raters, one interviews one rates audiotape of interview): 100% agreement</p> <p>Test-retest ("an average of 56 days"): r = 0.85 (p<0.001)</p>	Excellent
LARS	Zahodne et al. (2009)	Patient Interview	Probable idiopathic PD	71	<p>Cronbach's standardized alpha = 0.82</p>	<p>Convergent (SAS): Global score r = 0.62 (p<0.001) Intellectual curiosity subscale r = 0.61 (p<0.001) Action Initiation subscale r = 0.42 (p<0.001) Emotion subscale r = 0.33 (p<0.01) Self awareness subscale r = 0.229 (NS)</p> <p>Inter-method agreement (ICC; SAS): Global score = 0.75</p> <p>Divergent (BDI-II):</p>	-	Adequate

Global score $r = 0.45$ ($p < 0.001$)

Inter-method agreement (ICC; BDI-II):

Global score $r = 0.62$ ($p < 0.001$)

NS= Not Significant; ICC= Intra-class correlation NPI_d= NPI depression/dysphoria; NPI_{anx}= NPI anxiety; NPI_e= NPI elation/euphoria; PANSS= Positive and Negative Symptoms Scale; MADRS= Montgomery-Aasberg Depression Rating Scale; BPRS= Brief Psychiatric Rating Scale; BRSD-d= Behavior Rating Scale for Dementia- Depression subscale; BDI-II= Beck Depression Inventory II

Table 4 shows the different apathy scale validation studies in specific neurodegenerative disease patient samples, with descriptors and quality rating for each. The 11 studies ranged from Poor to Excellent quality.

The NPI has been translated in to Greek (Hellenic; Politis *et al.*, 2004) and a Persian language (Farsi; Malakouti *et al.*, 2012) and validated in AD and generic dementia populations. Both are administered and scored in the same way as the original NPIa (Cummings *et al.*, 1994). The Hellenic NPIa (H-NPIa) was translated by two bilingual psychiatrists from English to Greek and then independently translated back to English by one other psychiatrist. The quality of the study was classed as Good. The H-NPIa frequency x severity internal consistency fell within a range of questionable and acceptable with no further reliability reported. The concurrent validity between the apathy factor derived from the BPRS and the H-NPIa showed a low positive correlation.

The Farsi NPIa (F-NPIa) was also translated by two bilingual psychiatrists from English to Farsi and then back-translated to English independently by two other bilingual psychiatrists. It was then validated in an unspecified dementia patient sample. The methodological quality of the study was classed as Adequate. However, the internal consistency was reported of the F-NPIa was good. The concurrent was determined against the PANSS, which was found to be of a high correlation. The reliability of the scale was good, with an excellent inter-rater intra-class correlation coefficient (ICC) and the test retest correlation coefficient was good.

In addition to the NPI translations above, several revised versions have been developed. The Brief Clinical NPI (NPI-Q; Kaufer *et al.*, 2000), developed in English, is scored identically to the NPI but differs in administration method in that it is an informant report measure. The informants are asked to fill in the NPI-Q about the patient. The original validation study by Kaufer *et al.* (2000) recruited AD patients and was found to be of Adequate quality. While there was no report of the internal consistency for the NPI-Q apathy (NPIa-Q), the test-retest reliability correlation coefficient and convergent validity against the NPIa were both high positive correlations.

The NPI Clinical rating scale (NPI-C; de Medeiros *et al.*, 2010) is a revised version of the NPI with expanded domains and questions. It recognises the importance of the domain subquestions and rates frequency and severity through an item-by-item basis. It was validated in mild to severe probable AD patients by de Medeiros *et al.* (2010). This was a large-scale, cross-sectional validation study with 8 participating countries (Argentina, Brazil, Canada, France, Greece, Hungary, Italy and America) and the NPI-C was translated to 7 languages (English, French, Greek, Hungarian, Italian, Portuguese and Spanish). The quality of this study was Adequate. It does not report internal consistency NPI-C apathy (NPIa-C) and the convergent validity against the AES-I showed little or no positive correlation. The inter-rater ICC for each of the subquestions of the NPIa-C ranged from good to excellent.

The Starkstein Apathy Scale (SAS) was developed in English to assess apathy and validated in idiopathic PD patients (Starkstein *et al.*, 1992). It is a condensed version of the AES, with only 14 items, scored by same 4-point Likert scale. It

defines itself as less comprehensive than the AES as to be less demanding for PD patients to complete. This study was of Poor quality with only an acceptable internal consistency for the SAS. The convergent validity was determined through a blind neurologist rating participants as apathetic or non-aphathetic and then examining if their SAS scores are significantly different. They were found to be different, based on a sample size of 12 (six rated as apathetic, six rated as non-aphathetic by the neurologist). The inter-rater reliability correlation coefficient was high and the test-retest reliability correlation was very high. In a later study, the SAS was back-translated in to Norwegian, the translation critically appraised by four of the authors and used in a Norwegian PD patient sample study (Pedersen *et al.*, 2012). This study was also of Poor quality. This study reported a questionable internal consistency for the SAS with good discriminant validity due to little or no positive correlation with the MADRS.

The Lille Apathy Rating Scale was originally developed in French and validated in patients with probable PD by Sockeel *et al.* (2006). It is a 33-item scale scored by a mixture of simple (Yes/No/NA) and Likert-type responses. It is composed of factor analytically derived subscales (Intellectual curiosity, Action initiation, Emotional and Self awareness), which are also summed to provide a global apathy score. The original validation study (Sockeel *et al.*, 2006) was of Good quality. It showed an acceptable (between subscale) to good (between item) internal consistency. Concurrent validity between the AES and LARS global score was high positive correlation, with LARS subscales being positive and ranging from little or no to high correlations. Inter-rater ICC and test retest reliability was found to be excellent. A study by Zahodne *et al.* (2009) of a slightly less Adequate quality further

validated LARS in English in probable idiopathic PD patients. It was showed to have a good internal consistency. For convergent validity, the SAS and the LARS global score correlated positively and moderately, with the subscales positively correlating from little to none to moderately. Additionally, the inter-method agreement (ICC) for the SAS and LARS global score was excellent. Divergent validity between the BDI-II and LARS global score was moderate and the ICC between the two was fair.

Similarly to the LARS, the informant version of the LARS (LARS-i; Dujardin *et al.*, 2008) was also developed in French and for validation in a PD population. Some items have been reworded to be suitable as a question for a patient's informant. The administration and scoring methods are the same as the LARS. This particular study was of Adequate quality. The internal consistency for the LARS-i was shown to be good. The concurrent validity was good due to high positive correlations with the AES-I, AES-C and LARS. Inter-rater ICC and test retest reliability was excellent.

The AES-C was further validated in mild to moderate severity AD patient sample (Hsieh *et al.*, 2012). It was back-translated to Taiwanese by two bilingual individuals. This was done until an agreement on the translation was reached between the two. In terms of quality, this study was classed as Good. The internal consistency for the AES-C was found to be good, with a moderate positive convergent validity against the NPIa. Discriminant validity against the NPId showed little to no correlation while against the NPLe it showed a low negative correlation. Test retest reliability correlation coefficient for the AES-C was good.

The Dementia Apathy Interview and Rating (DAIR) is an informant based interview, developed and validated in English to assess apathy in probable and possible cases on AD (Stuss *et al.*, 2002). The assessment is a 16-question interview of the primary caregiver to assess the patient's initiation behavior, interest and engagement in the environment through a 4-point Likert scale. This was the only study with Excellent methodological quality. The internal consistency of the DAIR was good with a generally low positive correlation in relation to clinician's ratings, indicative of average convergent validity. It had little or no positive correlation with the BDRS, showing good discriminant validity against depression. It had a perfect inter-rater agreement with a high positive test-retest reliability correlation coefficient.

Discussion

This review systematically examines and provides an overview of the quality and psychometric properties of apathy scales, which have been used within neurodegenerative disease populations. Generally, most of the validation studies in this review were found to be of Adequate quality. Studies did not reach higher quality scores due to under described selection criteria, poor reporting of withdrawals from studies and a lacking of a good reference standard for diagnosing apathy. Nevertheless, their merits are that both the reference standards and the index tests for apathy were described in sufficient detail and, mostly, were independent of each other.

Generic neurodegenerative scales

The methodological quality of all the validation studies for the neurodegenerative generic scales was Adequate. This could have been improved with

better, defined selection criteria and comprehensive sampling to form more generic neurodegenerative patient groups. Therefore, it is important to examine the psychometric properties of each as to determine their practicality.

The widely used NPI, developed and validated in a generic dementia sample by Cummings *et al.* (1994), is often used as a reference standard for other apathy measures. In the original validation study by Cummings *et al.* (1994), the design of the NPI was done using a Delphi panel of experts as to also determine content validity. They reported apathy as ‘well assessed’ but did not examine any other statistical validity criteria for the NPIa. Specific NPIa internal consistency was not reported but it was deduced as good, with a high between-rater agreement and good test-retest reliability. Most of the NPI subdomains were concurrently validated against existing instruments, whereas the NPIa domain was not. This lack of concurrent validity for the NPIa and a methodological rating of Adequate for this study create uncertainty of the robustness of this subscale’s ability to assess apathy. It is observed that many studies in this review have used the NPIa as a reference standard to validate their own methods of apathy assessment. As a consequence, perhaps these and any future validation studies using the NPIa as a reference standard should be considered and interpreted with caution.

Similarly to the NPI, the AES is also considered as a validation reference standard. The original Marin *et al.* (1991) AES development and validation study was excluded in the earlier stage of this review due to a small sample size of the AD patients (N = 21 out of 123/ less than 20%) and included a very mixed patient sample of AD, stroke (left and right hemisphere), major depression and healthy controls.

Additionally, there was no specific AD reliability or validity information reported for the AES. However, a later study by Clarke *et al.* (2007) validated all forms of the AES against a generic dementia sample. The Clarke *et al.* (2007) study does not report reliability statistics and only factor analytically derived internal consistencies, making it difficult to interpret the scales as a whole. The convergent and discriminant validity was variable for all forms of the AES, resulting in unclear psychometric properties and, once again, therefore caution should be taken in considering its application as a reference standard for a neurodegenerative group.

Two studies validating the AES-10 (Lueken *et al.*, 2007, Leontjevas *et al.*, 2012), which is a shortened form of the AES, showed excellent internal consistency reliability and improved validity. Similar to the Clarke *et al.* study, the AES-10 validation studies documented no other reliability statistics. It should, however, be noted that both studies recruited Nursing Home residents some of whom did not have a diagnosis of a neurodegenerative disease. The Lueken *et al.* (2007) was more representative of a neurodegenerative generic dementia population. However, the Leontjevas *et al.* (2012) study was not very representative of a neurodegenerative population because almost half of the participants had no signs of dementia. So while the AES-10 seems psychometrically strong, it could still benefit from further validation in a generic neurodegenerative disease population.

Robert *et al.*'s (2002) AI validation study utilised a neurodegenerative generic patient sample containing AD, PD and mild cognitive impairment patients. While the AI was generally reliable, the inter-rater agreement, internal consistency, test-retest reliability being excellent, the validity was unclear. The concurrent validity was

underreported for the global score and showed an irregular relationship for the subscale scores. The association between the AI global and subscale scores with depression were stronger than that of the apathy measure. This stronger relationship with depression could indicate that the AI is not measuring the intended apathy, but rather some component associated with both apathy and depression, resulting in a lack of scale validity.

Based on the results of this review, it is difficult to establish which scale is best for assessing apathy in a generic neurodegenerative disease population. There are various methodological flaws with the validation studies that affect the validity and reliability of the each scale. The main issue being that a generic neurodegenerative sample should involve all forms of neurodegenerative disease, not just various dementias (Cummings *et al.*, 1994, Clarke *et al.*, 2007) or a mixture healthy individuals and neurodegenerative patients (Lueken *et al.*, 2007, Leontjevans *et al.*, 2012). Psychometrically, the AES-10 seems to be the most valid, but has not been applied in a purely neurodegenerative population.

Specific neurodegenerative scales

With regard to neurodegenerative specific apathy scales, the DAIR (Strauss *et al.*, 2002) validation study has Excellent methodological quality for AD patients, due to lack of bias, a superior quality of methodological reporting and high generalizability of the results. This was closely followed by Good quality validation studies of the AES-C (Hsieh *et al.*, 2012) and the H-NPIa (Politis *et al.*, 2004). Both had a representative sample and high generalizability of the results. However, the H-NPIa was validated against a combination of BPRS subscales that were labelled as an

apathy factor, which is not a direct or established measure of apathy but rather a new construct. Additionally, when compared psychometrically, the H-NPIa reliability and validity was inferior to both the DAIR and the AES-C, partly due to insufficient reliability criteria reporting. When comparing the DAIR and the AES-C, both are well constructed, informant interviews completed by trained assessor with matched psychometric properties. Both validation studies note good internal consistency reliability but vary in the reporting of reliability and validity. The AES-C reports more detailed convergent and discriminant validity that are marginally superior to the DAIR. Conversely, the DAIR has better inter-rater agreement and test-retest reliability in comparison to the AES-C. Based on the psychometric properties of both these scales and the quality of their validation studies, either of these methods would be recommended for assessing apathy in AD patient populations.

In relation to PD patients, both the SAS validation studies (Starkstein *et al.*, 1992, Pedersen *et al.*, 2012) scored Poor on quality mostly due to the lack of methodological reporting and selection bias and also showed unfavourable psychometric validity and reliability. Only the Sockeel *et al.* (2006) LARS validation study was methodologically classed as Good quality due to minimised bias, detailed methodology reporting and generalizability of results. The internal consistency reliability was found to be acceptable which was supported by very good validity and reliability properties of the LARS global score. The high convergent validity and excellent reliability make this a good measure for apathy in PD. However, it does not examine discriminant or divergent validity against depression, which has been shown to occur with apathy in PD (Lieberman, 2006). A later validation study showed average discriminant validity against depression (Zahodne *et al.*, 2009).

In ALS, apathy is continually assessed using the Frontal Systems Behavior Scale (Grossman *et al.*, 2007; Witgert *et al.*, 2010) despite this scale not being validated in this neurodegenerative disease population. Additionally, research in HD has been conducted using scales that assess uncorroborated apathy that is a derivation of factor analysis (Thompson *et al.*, 2012) or use scales that have not been validated in an HD sample (Hamilton *et al.*, 2003, van Duijn *et al.*, 2010). Therefore, no well-validated scales to assess apathy within these two neurodegenerative disorders have been found.

Strengths and limitations

This review has been undertaken using a systematic methodology (PRISMA), ensuring that eligibility criteria, inclusion and exclusion criteria were carefully outlined and implemented. Secondly, all studies were included that met the criteria, which could be seen to reduce publication bias. Thirdly, through rating the validation studies for the scales on methodological quality using the QUADAS, allowed us to determine the quality of the validation procedure for the apathy scale in question. Finally, interpreting the psychometric statistics for reliability and validity of the scales (adapted from Hermans *et al.*, 2011) allowed for a more standardized assessment. This resulted in more objectively driven recommendations for the most robust scale to be used in research or clinical practice to be made.

A limitation for this study could be that we chose to exclude studies that reported just validity or reliability of apathy scales. However, a diagnostic instrument should be both reliable and valid at detecting a symptom (Kimberlin and Winetrstein,

2008) and both of these need to be demonstrated in the same sample (i.e. there is little point in having a scale that is reliable but not valid in one sample, and valid but not reliable in another). Therefore, by including validation studies that examine both of these criteria, we attempted to avoid partially validated scales that might not be suitable for assessment. Additionally, a systematic review using the PRISMA method requires “a clearly formulated question that uses systematic and explicit methods...” of identification, selection and critical appraisal of “relevant research” (Moher *et al.*, 2009). So by clearly defining those studies reporting complete results (i.e. reliability and validity statistics of a scale) improve the quality of this systematic review, making it possible for more credible, scale-utility recommendations to be made.

Finally, it should be noted that apathy as a lack of motivation may manifest somewhat differently depending on the cultural context. Hence, caution is required when considering applying a scale validated in one country in another, especially if cultural norms are very dissimilar.

Conclusion

In conclusion, there are several recommended instruments to assess apathy for some specific neurodegenerative disease populations. The DAIR is a well-validated instrument in AD patients, shown to have very good psychometric characteristics, which is further supported by Excellent methodological quality of the validation study (Strauss *et al.*, 2002). In relation to PD patient, the LARS validation study (Sockeye *et al.*, 2006) showed that the global scoring of apathy was psychometrically strong, with further research needed to determine its subscale structure and relationship with

depression. Finally, scales should be developed and validated in diseases such as HD and ALS, which make apathy difficult to assess due to marked motor symptoms.

This review did not find any methodologically strong studies that validated a psychometrically robust, generic neurodegenerative disease apathy scale. This was partly due to the unrepresentative samples, lack of standardizing selection criteria in the validation studies for generic neurodegenerative disease. Many of the studies validated scales in samples that were biased to dementia patients, with no studies including ALS patients. It may be difficult to design or validate such a generic scale due to practical difficulties of forming a broad and mixed neurodegenerative disease patient sample group. However, further well-constructed validation studies or newly developed scales are needed to explore if it is possible to assess apathy in generic neurodegenerative disease.

Funding

R.R. is jointly funded by the Anne Rowling Regenerative Neurology Clinic and the Alzheimer Scotland Dementia Research Centre.

Conflict of interest

None.

Description of authors' roles

R.R. and J.S. devised the review. R.R. selected and condensed the material and wrote the paper with support of J.S. C.H. was the second reviewer for the paper, checked

material and co-rated papers. J.S. was the third reviewer and supervised R.R. S.A. supervised R.R. and provided intellectual input in to the manuscript.

Acknowledgements

The authors gratefully acknowledge the support of the Alzheimer Scotland Dementia Research Centre and the Anne Rowling Regenerative Neurology Clinic.

Supplementary material

Supplementary Table S1. Guidelines for interpreting psychometric statistics

(based on Hermans *et al.*, 2011)

Supplementary Figure S1. PRISMA Checklist

Supplementary Figure S2. Stacked bar chart of methodological quality for all the apathy scale validation studies by QUADAS item

References

Abrahams, S. (2011). Social cognition in amyotrophic lateral sclerosis.

Neurodegenerative Disease Management, 1(5), 397-405.

Chase, T. N. (2011). Apathy in neuropsychiatric disease: diagnosis, pathophysiology and treatment. *Neurotoxicity Research*, 19, 266-278.

Clarke, D. E., Ko, J. Y., Kuhl, E. A., van Reekum, R., Salvador, R., and Marin, R. S. (2011). Are the available apathy measures reliable and valid? A review of the psychometric evidence. *Journal of psychosomatic research*, 70(1), 73-97.

Clarke, D., van Reekum, R., Simard, M., Streiner, D., Freedman, M., and Conn, D. (2007). Apathy in dementia: an examination of the psychometric properties of the apathy evaluation scale. *The Journal of neuropsychiatry and clinical neurosciences*, 19(1), 57-64.

Clarke, D., et al. (2008). Apathy in dementia: clinical and sociodemographic correlates. *The Journal of neuropsychiatry and clinical neurosciences*, 20(3), 337-347.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Cummings, J. L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D. A., and Gornbein, J. (1994). The Neuropsychiatric Inventory comprehensive assessment of psychopathology in dementia. *Neurology*, 44(12), 2308-2308.

de Medeiros, et al. (2010). The Neuropsychiatric Inventory-Clinician rating scale (NPI-C): reliability and validity of a revised assessment of neuropsychiatric symptoms in dementia. *International Psychogeriatrics*, 22(6), 984-994.

Dujardin, K., Sockeel, P., Delliaux, M., Destée, A., and Defebvre, L. (2008). The Lille Apathy Rating Scale: Validation of a caregiver-based version. *Movement Disorders*, 23(6), 845-849.

Dujardin, K., Sockeel, P., Delliaux, M., Destée, A., and Defebvre, L. (2009).

Apathy may herald cognitive decline and dementia in Parkinson's disease. *Movement Disorders*, 24(16), 2391-2397.

Dujardin, K., et al. (2007). Characteristics of apathy in Parkinson's disease.

Movement Disorders, 22(6), 778-784.

George, D., and Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn and Bacon.

Girardi, A., MacPherson, S. E., and Abrahams, S. (2011). Deficits in emotional and social cognition in amyotrophic lateral sclerosis. *Neuropsychology*, 25(1), 53-65.

Goldberg, Y. K., Eastwood, J. D., LaGuardia, J., and Danckert, J. (2011).

Boredom: An emotional experience distinct from apathy, anhedonia, or depression. *Journal of Social and Clinical Psychology*, 30(6), 647-666.

Goldstein, L. H., and Abrahams, S. (2013). Changes in cognition and behaviour in amyotrophic lateral sclerosis: nature of impairment and implications for assessment. *The Lancet Neurology*, 12(4), 368-380.

Grossman, A. B., Woolley-Levine, S., Bradley, W. G., and Miller, R. G. (2007).

Detecting neurobehavioral changes in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis*, 8, 56-61.

- Hamilton, J. M., et al.** (2003). Behavioural abnormalities contribute to functional decline in Huntington's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 74(1), 120-122.
- Hermans, H., van der Pas, F. H., and Evenhuis, H. M.** (2011). Instruments assessing anxiety in adults with intellectual disabilities: A systematic review. *Research in Developmental Disabilities*, 32(3), 861-870.
- Hsieh, C. J., Chu, H., Cheng, J. J. S., Shen, W. W., and Lin, C. C.** (2012). Validation of apathy evaluation scale and assessment of severity of apathy in Alzheimer's disease. *Psychiatry and clinical neurosciences*, 66(3), 227-234.
- Jonsson, M., Edman, Å., Lind, K., Rolstad, S., Sjögren, M., and Wallin, A.** (2010). Apathy is a prominent neuropsychiatric feature of radiological white-matter changes in patients with dementia. *International journal of geriatric psychiatry*, 25(6), 588-595.
- Kaufer, D. I., et al.** (2000). Validation of the NPI-Q, a brief clinical form of the Neuropsychiatric Inventory. *The Journal of neuropsychiatry and clinical neurosciences*, 12(2), 233-239.
- Kimberlin, C. L., and Winetrstein, A. G.** (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65(23), 2276-2284.

Landes, A. M., Sperry, S. D., and Strauss, M. E. (2005). Prevalence of apathy, dysphoria, and depression in relation to dementia severity in Alzheimer's disease. *The Journal of Neuropsychiatry and Clinical Neuroscience*, 17, 342-349.

Leontjevas, R., et al. (2012). A comparative validation of the Abbreviated Apathy Evaluation Scale (AES-10) with the Neuropsychiatric Inventory Apathy Subscale against diagnostic criteria of apathy. *Journal of the American Medical Directors Association*, 13(3), e301-e306.

Levy, M. L., et al. (1998). Apathy is not depression. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 10(3), 314-319.

Lieberman, A. (2006). Depression in Parkinson's disease—a review. *Acta Neurologica Scandinavica*, 113(1), 1-8.

Lueken, U., Seidl, U., Völker, L., Schweiger, E., Kruse, A., and Schröder, J. (2007). Development of a short version of the Apathy Evaluation Scale specifically adapted for demented nursing home residents. *The American journal of geriatric psychiatry*, 15(5), 376-385.

Malakouti, S. K., Panaghi, L., Foroughan, M., Salehi, M., and Zandi, T. (2012). Farsi version of the Neuropsychiatric Inventory: validity and reliability study among Iranian elderly with dementia. *International Psychogeriatrics*, 24(2), 223-230.

Marin, R. S. (1996). Apathy: concept, syndrome, neural mechanisms, and treatment. *Seminars in Clinical Neuropsychiatry*, 1, 304-314.

Marin, R. S. Biedrzycki, R. C., and Firinciogullari, S. (1991). Reliability and validity of the Apathy Evaluation Scale. *Psychiatry research*, 38(2), 143-162.

Mendez, M. F., Lauterbach, E. C., and Sampson, S. M. (2008). ANPA Committee on Research. An evidence-based review of the psychopathology of frontotemporal dementia: a report of the ANPA Committee on Research. *The Journal of Neuropsychiatry and Clinical Neuroscience*, 20(2), 130-149.

Mioshi, E., Kipps, C. M., and Hodges, J. R. (2009). Activities of daily living in behavioral variant frontotemporal dementia: differences in caregiver and performance-based assessments. *Alzheimer Disease and Associated Disorders*, 23(1), 70-76.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151(4), 264-269.

Pedersen, K. F., Larsen, J. P., Alves, G., and Aarsland, D. (2009). Prevalence and clinical correlates of apathy in Parkinson's disease: a community-based study. *Parkinsonism and Related Disorders*, 15, 295-299.

Pedersen, K. F., Alves, G., Larsen, J. P., Tysnes, O. B., Møller, S. G., and Brønnick, K. (2012). Psychometric properties of the Starkstein Apathy Scale in patients with early untreated Parkinson disease. *The American journal of geriatric psychiatry*, 20(2), 142-148.

Politis, A. M., Mayer, L. S., Passa, M., Maillis, A., and Lyketsos, C. G. (2004). Validity and reliability of the newly translated Hellenic Neuropsychiatric Inventory (H-NPI) applied to Greek outpatients with Alzheimer's disease: a study of disturbing behaviors among referrals to a memory clinic. *International journal of geriatric psychiatry*, 19(3), 203-208.

Quaranta, D., Marra, C., Rossi, C., Gainotti, G., and Masullo, C. (2012). Different Apathy Profile in Behavioral Variant of Frontotemporal Dementia and Alzheimer's Disease: A Preliminary Investigation. *Current Gerontology and Geriatrics Research, Current Gerontology and Geriatrics Research*.
doi:10.1155/2012/719250

Robert, P. H., et al. (2002). The apathy inventory: assessment of apathy and awareness in Alzheimer's disease, Parkinson's disease and mild cognitive impairment. *International journal of geriatric psychiatry*, 17(12), 1099-1105.

Sockeel, P., Dujardin, K., Devos, D., Deneve, C., Destée, A., and Defebvre, L. (2006). The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: validation in Parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 77(5), 579-584.

Staekenborg, S. S., et al. (2010). Behavioural and psychological symptoms in vascular dementia; differences between small-and large-vessel disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 81(5), 547-551.

Starkstein, S. E., Mayberg, H. S., Preziosi, T., Andrezejewski, P., Leiguarda, R., and Robinson, R. G. (1992). Reliability, validity, and clinical correlates of apathy in Parkinson's disease. *The Journal of Neuropsychiatry and Clinical Neuroscience*, 4(2), 134-139.

Strauss, M. E., and Sperry, S. D. (2002). An informant-based assessment of apathy in Alzheimer disease. *Cognitive and Behavioral Neurology*, 15(3), 176-183.

Thompson, J. C., et al. (2012). Longitudinal evaluation of neuropsychiatric symptoms in Huntington's disease. *The Journal of neuropsychiatry and clinical neurosciences*, 24(1), 53-60.

van Duijn, E., Kingma, E. M., and van der Mast, R. C. (2007). Psychopathology in verified Huntington's disease gene carriers. *The Journal of neuropsychiatry and clinical neurosciences*, 19(4), 441-448.

van Duijn, E., Reedeker, N., Giltay, E. J., Roos, R. A. C., and van der Mast, R. C. (2010). Correlates of apathy in Huntington's disease. *The Journal of neuropsychiatry and clinical neurosciences*, 22(3), 287-294.

van Reekum, R., Stuss, D. T., and Ostrander, L. (2005). Apathy: why care? *The Journal of neuropsychiatry and clinical neurosciences*, 17(1), 7-19.

Viera, A. J., and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5), 360-363.

Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., and Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC medical research methodology*, 3(1), 25.

Witgert, M., et al. (2010). Frontal-lobe mediated behavioral dysfunction in amyotrophic lateral sclerosis. *European Journal of Neurology*, 17(1), 103-110.

Woolley, S. C., Zhang, Y., Schuff, N., Weiner, M. W., and Katz, J. S. (2011). Neuroanatomical correlates of apathy in ALS using 4 Tesla diffusion tensor MRI. *Amyotrophic Lateral Sclerosis*, 12(1), 52-58.

Zahodne, L. B., et al. (2009). Examination of the Lille apathy rating scale in Parkinson disease. *Movement Disorders*, 24(5), 677-683.

Supplementary Table S1. Guidelines for interpreting psychometric statistics (based on Hermans *et al.*, 2011)

Psychometric Statistic	Interpretation
Internal Consistency	
<0.50	Unacceptable
0.50-0.59	Poor
0.60-0.69	Questionable
0.70-0.79	Acceptable
0.80-0.89	Good
≥0.90	Excellent
Correlation Coefficients (Pearson's product-moment and Spearman rank)	
<0.29	Little or no correlation
0.30-0.49	Low correlation
0.50-0.69	Moderate correlation
0.70-0.89	High correlation
≥0.90	Very high correlation
Intra-class Correlation Coefficient	
<0.40	Poor
0.40-0.59	Fair
0.60-0.74	Good
≥0.75	Excellent
Kappa Statistic	
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Supplementary Figure S1. PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3 to 7
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	7 and 8
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	8
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	8
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	8 and 9
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	9 and 10
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	9 and 10
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	8 to 10
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	10 and 11
Summary	13	State the principal summary measures (e.g., risk ratio, difference in means).	10 and 11

measures			
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	10 and 11
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	10 and 11
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	10 and 11
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	11 to 13
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	13 to 21
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	13 to 21
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	13 to 21
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	13 to 21
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	13 to 21
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	13 to 21
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	21 to 26
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	26 and 27
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	27 and 28
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	28

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

Supplementary Figure S2. Stacked bar chart of methodological quality for all the apathy scale validation studies by QUADAS item

