Title: Systematic Review of the Diagnostic Accuracy of the Non-English Versions of Addenbrooke's Cognitive Examination—Revised and III

Running Head: Diagnostic accuracy of Non-English Versions of ACE-R and ACE-III

Noor Habib[a] Joshua Stott[a*]

[a] Department of Clinical, Educational and Health Psychology

University College London

London, United Kingdom

*Full address of the corresponding author:

Dr. Joshua Stott

Department of Clinical, Educational and Health Psychology

University College London, 1-19 Torrington Place

London WC1E 7HB (UK)

Tel: 0207-679 5950, E-Mail: j.stott@ucl.ac.uk

Sponsorship: Joshua Stott's involvement and time spent on this review funded by a fellowship generously awarded to him by the Alzheimer's Society, Grant number 236 (AS-CTF-14-005).

Abstract

Objective: This systematic review aims to review the evidence for the diagnostic accuracy of the non-English updated versions of Addenbrooke's Cognitive Examination (ACE)—the ACE-Revised (ACE-R) and the ACE-III—in the diagnosis of dementia. **Methods:** A systematic search resulted in 16 eligible studies evaluating the diagnostic accuracy of ACE-R and ACE-III in ten different languages. Most studies were assessed as of medium to low quality using Standards For Reporting of Diagnostic Accuracy (STARD) guidance. **Results:** The findings of excellent diagnostic accuracy are compromised by the methodological limitations of studies. While studies generally reported excellent diagnostic accuracy across and within different languages, optimal cut-offs even within particular language versions, varied. **Conclusion:** There is a need for future research to address these limitations through adherence to STARD guidelines. The ACE-III is particularly under-evaluated and should be a focus of future research. The variance in obtained optimal cut-offs within language versions is an issue compromising clinical utility and could be addressed in future work through use of a-priori defined thresholds.

Key Words: Dementia. Cognitive assessment. Screening. Memory.

Introduction

An accurate clinical diagnosis of dementia at an early stage and an early intervention that slows the progression of the disease can lead to a better prognosis (NICE, 2006).

Good quality screening tools with high sensitivity and specificity can facilitate this process (Prince, Renata & Ferri, 2011).

The Addenbrooke's Cognitive examination-revised (ACE-R) (Mioshi, Dawson, Mitchell, Arnold & Hodges, 2006) and the Addenbrooke's Cognitive Examination III (ACE-III) (Hsieh, Schubert, Hoon, Mioshi & Hodges, 2013) are two brief (15-minute) screening tools that have been developed to screen for dementia (Velayudhan et al., 2014). Both tools provide sub-scale scores for the cognitive domains of attention and orientation, memory, language, visuospatial functioning and verbal fluency. The English versions of these tools have high diagnostic accuracy at the recommended cut-off of 88; the ACE-R has a sensitivity of 91.8% and specificity of 87.5% (Mioshi et al., 2006) and the ACE-III has a sensitivity of 1 and specificity, 0.96 (Hsieh et al., 2013). The ACE-III was developed following potential licensing issues with the ACE-R, and is a very similar tool, they differ only in minor ways and in studies examining them together, scores have an almost perfect correlation ($r = 0.99$) (Hsieh et al., 2013). Thus while the authors have suggested that the ACE-III be used instead (Hsieh et al., 2013), results are likely to be very similar to those of the ACE-III, and will help inform future practice with this tool. Furthermore, as the ACE-R has been validated in more languages than the ACE-III, it may still have utility in assessing those with dementia who present to services who do not speak English fluently, despite the potential licensing issues. In support of this approach a number of studies validating the ACE-R in different languages have been published since the licensing issues arose (Fang et al., 2014; Sobreira et al., 2015; Gonçalves et al., 2015). Since the advent of the ACE-R and ACE-III, a number of non-English language versions have been developed. This is important in widening access to dementia diagnosis and hence care in countries where English is the majority languages but have significant non-English speaking minorities such as the UK, Australia and USA (Grypma, Mahajani & Tam, 2007). It is also important in increasing access to

dementia care in countries where dementia prevalence is increasing and non-English language speakers may be in the majority (Alexopoulos et al., 2010). It is not possible to use the cut-off thresholds and diagnostic accuracy metrics of the English versions for non-English versions as the equivalence on these aspects of screening tools across cultures and language cannot be assumed (Borsa, Damasio & Bandeira, 2012). Consequently, the aim of the current paper is to review the evidence as to the diagnostic accuracy of the non-English versions of the ACE-R and the ACE-III in diagnosing dementia with reference to diagnostic accuracy metrics derived from receiver operating characteristic (ROC) analysis as recommended in the Standards for Reporting of Diagnostic Accuracy (STARD) (Bossuyt et al., 2015). Despite the licensing issues, the majority of the published non-English versions of this measure is on ACE-R. Therefore, a secondary aim of this review is to shade the light on common methodological problems in the non-English versions of ACE-R to be considered in future translations of ACE-III.

Methods

Design

A systematic search and appraisal (using STARD guidelines) of published evidence on the diagnostic accuracy of non-English language versions of the ACE-R and ACE-III was undertaken. The search and appraisal were in line with systematic principles (Moher, Liberati, Tetzlaff & Altman, 2009).

Search Strategy

Three electronic databases were searched: PsychINFO, MEDLINE, and EMBASE. The following search words were used: *Addenbrooke's Cognitive Examination-Revised*, *Addenbrooke's Cognitive Examination-III*, *ACE-R*, *ACE-III*, and *foreign language translation*. Terms were combined using Boolean operators *OR* and *AND*. Because the

original English language versions of the ACE-R and the ACE-III were published in 2006 and 2013, respectively, only studies published from 2006 to September 2016 were included in the search. Titles and abstracts were first screened for eligibility, with full articles accessed for review on the basis of screening. These full text articles were reviewed to assess their eligibility in light of the inclusion and exclusion criteria of the current review. Reference lists of included studies were reviewed to identify further articles. The inclusion criteria were: 1) Studies investigating the diagnostic accuracy of non-English versions of the ACE-R and the ACE-III; 2) If more than one study translated the ACE-R and the ACE-III to the same language, all studies were included.

The exclusion criteria were: 1) Studies not in English; 2) Studies on the English versions of the ACE-R and the ACE-III; 3) Studies that used non-English versions of the ACE-R or the ACE-III to track changes in cognitive functioning over time rather than assessing diagnostic accuracy; 4) Studies that used non-English versions of the ACE-R or the ACE-III as part of a wider cognitive assessment without providing information on diagnostic accuracy; 5) Abstracts, response letters, reviews and guides.

Data Extraction

Identified abstracts were exported to EndNote and screened against eligibility criteria by N.H. Potential articles for inclusion were accessed and reviewed again against eligibility criteria, again by N.H. In uncertain cases J.S. was consulted and a decision on exclusion/inclusion was reached through discussion. The final list of studies included was reviewed by J.S.

Summarising Findings and Quality Appraisal

Summary data as to aims, participants, study design, analyses, and diagnostic accuracy findings was extracted from each included study. Study quality was critically appraised

and scored by N.H. using the STARD quality appraisal checklist (Bossuyt et al., 2015). STARD was used as it is an international consensus instrument for evaluating the diagnostic accuracy of psychometric measures. This consists of 31 items in total with items relating to the title (1 item), abstract (1 item), introduction (2 items), methods (17 items), results (8 items) and discussion (2 items). Each item has a maximum score of 2 indicating information is present and a minimum of 0 indicating an absence of information, with 1 indicating information is present but with inadequate details. Scores were summed to give an overall quality score for the study. As recommended by the National Institute for Clinical Excellence (NICE, 2014) overall scores were combined with an assessment of how likely the issues identified were to alter the conclusion of the study to give an overall quality rating of high (++), medium (+), or low (-) quality.

Results

721 articles were initially identified from the three databases searched. Figure 1 provides a flowchart summarising the steps taken in excluding and including studies for this review. Reasons for exclusion included using the English version of ACE-R or ACE-III, using the measure as part of a wider assessment, article not written in English or the study focused only on normative data. A total of 16 papers were included in this review.

-Insert Figure 1-

Summary of Results

A summary of study characteristics of all eligible studies is reported in Table 1. The studies reported on French, German, Greek, Italian, Japanese, Korean, Mandarin, Cantonese, Portuguese and Spanish versions of the ACE-R and a Spanish and Thai language version of ACE-III. The mean age (±SD) of participants with dementia in the

studies ranged between 66.20 (± 8.96) (Konstantinopoulou et al., 2011) and 80.9 (± 3.6) (Pigliautile, 2012). Years of education ranged between 3.7 (± 4.2) (Wong et al., 2013) and 18 (± 4) (Bastide et al., 2012), male:female ratio varied between 8:8 (Konstantinopoulou et al., 2011) and 34:92 (Kawata et al., 2012). The types of dementia included across the studies were Alzheimer's Disease, Frontotemporal lobar degeneration (FTLD), Behavioural variant frontotemporal lobar degeneration (bvFTD), Subcortical vascular dementia.

- Insert Table 1-

Information about diagnostic accuracy including cut-off scores, sensitivity, specificity, and other ROC analysis metrics only reported values of dementia and normal cognition as shown in Table 2. The cut-off scores for the translated measures included in this review ranged between 60 (Pigliautile, 2012) and 89 (Bastide et al., 2012). The sensitivity of the measures ranged between 82% (Pigliautile, 2012) and 100% (Carvalho, Barbosa & Caramelli, 2010 ; Raimondi et al., 2012) and specificity values ranged between 68% (Sobreira et al., 2015) and 100% (Pigliautile, 2012; Raimondi et al., 2012) Of the 16 studies included in the current review, 12 studies were judged to be of low (-) quality, and four were judged to be of medium (+) quality (see Table 2). There were some issues general to all or most studies: Across all studies, no information was given on how indeterminate scores were handled and the frequency of such scores might have inflated or deflated the estimation of the diagnostic accuracy depending on whether they occurred more frequently in people with or without dementia (Bossuyt et al., 2003). There was also no indication of the time interval between the index and the reference tests in 15 studies; consequently, there may have been changes in the target condition over time that might have influenced the diagnostic accuracy of the measure (Bossuyt et al., 2003). Power calculation and intended sample size was not reported in 12 studies, despite the importance of determining the sample size needed to identify

clinically relevant findings (Machin, Campbell, Fayers & Pinol, 1997). Furthermore, nine studies did not include information on the sampling process; thus, it was difficult to assess the population to whom the study was generalisable (Knottnerus & Muris, 2002). Similarly, in nine studies, assessors were aware of the clinical diagnosis of the participants while administrating the ACE-R or the ACE-III. Non-blinding might mean that more people were accurately diagnosed in studies than they would be in clinical practice because assessors already know who has dementia (Philbrick, Horwitz & Feinstein, 1980). Below we discuss each language version individually, where the individual studies were rated as low quality it was generally because all or many of the above issues were present. Where these issues were (at least partially) addressed and the study was of higher quality it is specified in individual language summaries below.

- Insert Table 2-

Results of Individual Studies Categorised by the Language of Translation

French Translation. A retrospective study investigated the diagnostic accuracy of the French version of the ACE-R (Bastide et al., 2012). When differentiating those with dementia from healthy controls, the sensitivity of the test was 98.4% and the specificity was 98.6%, for a threshold of 89. The findings suggested that the test is a good tool to identify people with and without dementia (type not specified). However, the study was rated as of low quality due to many of the issues discussed above, and consequently the level of bias in the findings is unclear.

German Translation. One study investigated the diagnostic accuracy of a German translation of the ACE-R (Alexopoulos et al., 2010). The results of the study suggested that the measure could be used to discriminate between people with Alzheimer's disease

(sensitivity = 92%, specificity = 96%) at the threshold of 82 or frontotemporal lobar degeneration (sensitivity = 88%, specificity = 96%) at the threshold of 83 from healthy controls. However, the study was rated as low quality due to many of the issues discussed above, and the ability of the authors to draw an unbiased conclusion about sensitivity and specificity may have been compromised.

Greek Translation. The article on the Greek translation of the ACE-R (Konstantinopoulou et al., 2011) investigated its diagnostic accuracy in differentiating those with dementia of Alzheimer's type and frontotemporal lobar degeneration from those without. The findings from this study suggested that the ACE-R had excellent ability to discriminate those with from those without dementia (sensitivity = 97.1%v and specificity = 81.7% at a threshold of 85). However, the study was rated as of low quality due to the issues discussed above, possibly biasing conclusions drawn.

Italian Translation. One study evaluated the diagnostic accuracy of the Italian ACE-R (Pigliautile, 2012) in young-old adults aged (70.8 ± 3.6) and old-old adults aged (80.9 ± 3.6) in detecting dementia (type not specified). The findings from the study suggested a sensitivity of 90% and a specificity of 80% with a cut-off of 79 for young-old adults and a sensitivity of 82% and a specificity of 100% with a cut off of 60 for old-old adults, demonstrating good diagnostic accuracy across age groups. However, the study was rated as low quality. Consequently, the validity of conclusions as to diagnostic accuracy may be compromised.

Japanese Translation. In two studies, the authors investigated the diagnostic accuracy of the Japanese translation of the ACE-R in diagnosing dementia (type not specified). Both articles Kawata et al., 2012; Yoshida et al., 2012) suggested that the ACE-R is an

excellent tool to identify people with dementia (sensitivity = 94%, specificity = 94%) at a cut-off of 80 (Kawata et al., 2012) and (sensitivity = 99%, specificity = 99%) at a cut-off of 82 (Yoshida et al., 2012). In both cases the researchers were blind to clinical status of participants, enhancing study validity (Philbrick et al., 1980) but in both cases issues such as the lack of clear indication of the time interval between the reference and index test means that it is somewhat hard to know to whom the results are generalizable and it is notable that different optimal cut-offs were found although there was no significant difference in demographics of both studies. Unmeasured differences in the samples may account for the different cut offs found here.

Korean Translation. In the article examining the diagnostic accuracy of the Korean ACE-R (Kwak, Yang & Kim, 2010). the authors focused on ability both to detect dementia and to differentiate between Alzheimer's disease and Subcortical vascular dementia In detecting dementia, test sensitivity was 93% and specificity was 95% at a cut-off of 78, suggesting an excellent diagnostic accuracy. However, sensitivity and specificity of the test to differentiate between Alzheimer's disease and Subcortical vascular dementia was less accurate. The study was rated as of low quality, perhaps compromising ability to draw valid conclusions as to diagnostic accuracy.

Mandarin/Cantonese Translation. One article investigated the diagnostic accuracy of the Chinese (Mandarin) translation of the ACE-R (Fang et al., 2014) in diagnosing Alzheimer's disease. The results suggested that the ACE-R was an excellent tool (sensitivity = 92%, specificity = 86%) at a threshold of 67. The study was rated as low quality and the validity of conclusions may be compromised due to many of the issues discussed above. The second article reported on the diagnostic accuracy of the Chinese (Cantonese) translation of the ACE-R (Wong et al., 2013). The results indicated

excellent diagnostic accuracy of the ACE-R (sensitivity = 93%, specificity = 95%) in diagnosing dementia (type not specified) with cut-off of 73. This study was rated as of medium quality, as those administering the ACE-R were blind to diagnostic status and there was a specified time interval of a week between index and reference tests. Therefore, the validity of the study for drawing a conclusion about sensitivity and specificity was probably reasonably high, however lack of clarity regarding sampling procedures may affect ability to generalise the results.

Portuguese Translation. The diagnostic accuracy of the Portuguese (Brazilian) translation of the ACE-R was assessed in two studies. The findings from the first study Carvalho et al., 2010) suggested 100% sensitivity and 82.26% specificity at a cut-off of 78, indicating excellent ability to detect Alzheimer's disease. The findings from the second article on the Brazilian translation of the ACE-R (Sobreira et al., 2015) indicated less accuracy in distinguishing between people with or without dementia among those who had Parkinson's disease (sensitivity = 88%, specificity = 68%) at a cut-off of 76. However, both studies were rated as low quality. Consequently, while good diagnostic accuracy was reported, particularly to detect dementia of Alzheimer's type, the validity of the sensitivity and specificity figures may be compromised.

In addition to the Brazilian Portuguese versions, one study examined a European Portuguese version of the ACE-R study (Gonçalves et al., 2015) and suggested an excellent ability to detect those with Alzheimer's disease (sensitivity = 100%, specificity = 97%) and subcortical vascular dementia (sensitivity = 97%, specificity = 92%) with cut-offs of 72. The study was rated as medium quality because although many of the methodological issues affecting other studies were also present here, the blinding of the assessors to clinical information when administering ACE-R makes the conclusions more robust.

Spanish Translation. Two articles were identified on the diagnostic accuracy of the Spanish (Argentinian) translation of the ACE-R, and one article on the Spanish (European) translation of the ACE-III. In the first article on the Spanish (Argentinian) ACE-R (Raimondi et al., 2012), people who had Alzheimer's disease or subcortical vascular dementia were compared with healthy individuals who participated as study controls. The results suggested that the ACE-R is an excellent tool to discriminate between people with or without dementia (type not specified) (sensitivity = 100%, specificity = 100%) with cut-off of 88. However, the study was rated as low quality. Thus the validity of the study to draw a conclusion about sensitivity and specificity may have been compromised.

In the second article, the ability of the Spanish (Argentinian) version of the ACE-R to differentiate between Healthy controls and those with dementia (A mixed population of Alzheimer's disease and behavioural variant FTD) was evaluated (Torralva et al., 2011). The ACE-R showed excellent sensitivity = 97% and specificity = 88% with a cut-off of 85. However, due to the issues discussed above the study was rated as of low quality, possibly biasing conclusions as to diagnostic accuracy.

The third article (Matias-Guiu et al., 2015) evaluated diagnostic accuracy of the European Spanish translation of the ACE-III the only article in this review examining the ACE-III. The results suggested that the ACE-III is a good tool in distinguishing those with dementia with a sensitivity of 83%, specificity of 80% with cut-off of 65.6. However, the study was rated as of low quality, and ability of the study to draw a conclusion about the sensitivity and specificity may have been compromised.

Thai Translation. One study investigated the diagnostic accuracy of the Thai translation of ACE-III (Charernboon, Jaisin & Lerthattasilp, 2016). People with early dementia

(Alzheimer's disease, mixed type and vascular dementia) were compared with people with mild cognitive impairment and healthy controls. The results revealed excellent sensitivity = 100% and specificity = 97% with an optimal cut-off score of 61 to differentiate people with and without dementia., The study was rated as of medium quality meaning conclusions may be more robust than some other studies reported here.

Discussion

The aim of the current review was to investigate the diagnostic accuracy of the translated versions of the ACE-R and the ACE-III in detecting dementia. In general, the translated versions of the ACE-R and ACE-III showed good to excellent sensitivity and specificity for detecting dementia. This may vary across subtype as while values were high for those with Alzheimer's disease, FTD and vascular dementia, the one study evaluating dementia in Parkinson's indicated poorer diagnostic accuracy. Optimal cut-off thresholds were less consistent and varied across types of dementia (Torralva et al., 2011) and also across language versions. This highlights the need to evaluate the ACE-R/ACE-III in each new population and the need to use ROC analysis to develop cut-offs for each new language version and to be clear as to diagnostic subtypes within a sample (Borsa, Damasio & Bandeira, 2012). When two studies assessed the same language version, optimal cut-offs also varied between studies (Sobreira et al., 2015; Carvalho et al., 2010) and (Kawata et al., 2012; Yoshida et al., 2012). While in some cases this may be due to methodological problems leading to inter-study variability in sensitivity and specificity values (Sobreira et al., 2015; Carvalho et al., 2010) due to bias (Bossuyt et al., 2003) this was also the case when both studies were of higher methodological quality (Kawata et al., 2012; Yoshida et al., 2012). One possible explanation is that unmeasured differences in populations across studies, such as mean age and years of education, affect results. The range of the years of education varies between studies. It

is possible that without education-adjusted cut-offs to reduce biases (Kittner et al., 1986) differences between studies are due to educational variation rather than (or as well as) cultural variation. These findings also emphasise the need for further research on population characteristics other than language, years of education, and age that affect ACE-R/ACE-III scores. Premorbid IQ and ethnicity, have been found to be associated with other screening tool performance (Whitney, Maoz, Hook, Steiner & Bieliauskas, 2007; Pedrazaa et al., 2012), and should be measured in relation to ACE-III/ACE-R. It should also be noted that while only two studies evaluated the ACE-III, they showed very similar results to those of the ACE-R and issues raised in this review in relation to the ACE-R will apply equally to the ACE-III.

Methodological Problems and Limitations

There were some limitations to the review process. Thorough assessment of the identified articles against inclusion and exclusion criteria was carried out by the first author alone. Although the second author was consulted in relation to queries and experts in the field were consulted as to any missing articles, this is a limitation. Another limitation was that most studies included in the review were on ACE-R. Similar limitations apply to quality assessment of articles. We did not search grey literature as we wanted to uphold the quality of included studies, but had we done so we may have identified further relevant articles. Additionally, it was a limitation that we did not include studies not written in English. This only led to a small number of articles (N=4) relating to diagnostic accuracy not being included, but did mean that three languages/cultural variants of a language (Turkish, Czech and Chilean Spanish) were not included in our review. Another limitation was that we appraised a screening tool that was inherently culturally biased (Dodge et al., 2009) before the process of translation and validation in different culture. Additionally, it was a limitation that we only reported the diagnostic

accuracy values for dementia without elaborating on mild cognitive impairment in studies that examined both.

Areas for Future Research and clinical practice

Given variability in cut-off thresholds found here within and across languages, clinicians should be cautious in applying them. Research on specific language versions should also be replicated, seeking to evaluate not only the ROC derived optimal cut off in their sample but the performance of cut-offs found in previous studies to build up an evidence base for specific cut-offs (Grypma et al., 2007) in specific languages. It general it is important for future research to adhere to STARD guidelines for diagnostic studies in order to reduce bias in conclusions about diagnostic accuracy. Finally, given that only two paper examined the ACE-III were eligible for the current review, future research could focus on the translated versions of the ACE-III (Hsieh et al., 2013) and should consider using established processes for cross cultural adaptation (Hambleton, 2005) prior to assessment.

Conclusion

The findings of excellent diagnostic accuracy are compromised by the methodological limitations of studies. There is a need for future research to address the limitations discussed in the current review through adherence to STARD guidelines.

Conflict of Interest Declaration

The authors have no conflict of interest to report.

Description of Authors' roles

NH jointly conceived the study and wrote up the final draft. She approved the final draft for publication. JS jointly conceived the study and revised it for critically important content.

Funding

Joshua Stott's time on this project was funded as part of a fellowship awarded to Joshua Stott by the Alzheimer's society. Grant number 236 (AS-CTF-14-005). This funding source had no involvement in the study design, collection, analysis and interpretation of data, writing the manuscript and in the decision to submit the manuscript for publication.

Table 1 Study Characteristics

Language	ACE-R/ ACE-III	Participants (type)	Gender male:female	Age in years Mean ± SD	Years of Education Mean ± SD	ACE-R/ACE-III Mean ± SD
French (Bastide et al., 2012)	ACE-R	Dementia (n= 128) MCI (n= 118) Healthy (n= 73)	47:81 47:71 17:56	75 ±11 72 ±9 68 ±11	18 ± 4 18 ± 4 20 ± 4	70 ± 10 83 ± 8 93 ± 4
German (Alexopoulos et al., 2010)	ACE-R	Alzheimer's (n= 56) FTLD (n= 22) MCI (n= 75) Healthy (n= 76)	20:36 13:9 45:30 29:47	72.00 ± 8.18 69.64 ± 6.18 67.83 ± 8.01 69.64 ± 7.53	11.02 ± 2.63 11.70 ± 3.52 12.00 ± 3.27 11.78 ± 2.51	64.80 ± 11.32 64.50 ± 17.82 81.34 ± 9.09 90.37 ± 4.99
Greek (Konstantinopou-lou et al., 2011)	ACE-R	Alzheimer's (n= 16) FTLD (n= 19) Healthy (n= 60)	8: 8 6:13 30:30	71.69 ± 5.50 67.47 ± 6.87 66.20 ± 8.96	7.75 ± 3.98 9.89 ± 4.12 10.60 ± 4.22	55.63 ± 17.14 61.00 ± 17.82 89.13 ± 7.54
Italian (Pigliautile, 2012)	ACE-R	Young-old Dementia (n= 40) Healthy (n= 41) Old-old Dementia (n= 67) Healthy (n= 31)	 16:24 18:23 25:42 11:20	 70.8 ± 3.6 69.6 ± 2.8 80.9 ± 3.6 80.7 ± 3.6	 7.1 ± 3.7 8.9 ± 4.6 7.1 ± 4.8 7.7 ± 3.9	 63.3 ± 13.2 87.1 ± 9.3 53.6 ± 12.2 80.5 ± 10.7
Japanese (Kawata et al., 2012)	ACE-R	Dementia (n= 126) Healthy (n= 85)	34:92 34:51	77.3 ± 7.6 71.5 ±9.1	10.6 ± 2.5 12.3 ± 2.6	58.4 ± 16.4 90.8 ± 6.9
Japanese (Yoshida et al., 2012)	ACE-R	Dementia (n= 130) MCI (n= 39) Healthy (n= 73)	42: 88 17: 22 27: 46	75.4 ± 7 71.4 ± 9.2 66.3 ± 10	11.1 ± 2.7 11.4 ± 2.1 12.7 ± 2.3	61.5 ± 12.9 82.2 ± 6.4 93.3 ± 3.9

Korean (Kwak et al., 2010)	ACE-R	AD (n= 30) SVD (n= 42) Healthy (n= 84)	13:17 20:22 40:44	73.1 ± 11.2 70.1 ± 10.2 67.8 ± 9.3	8.9 ± 4.2 8.6 ± 3.9 10.1 ± 4.1	52.5 ± 15.1 53.2 ± 17.0 80.7 ± 6.0
Mandarin (Chinese) (Fang et al., 2014)	ACE-R	AD (n= 25) MCI (n= 75) Healthy (n= 51)	11:14 37:38 23:28	73.32 ± 8.13 69.52 ± 9.69 68.16 ± 8.18	9.68 ± 5.01 10.07 ± 4.41 11.77 ± 3.46	55.72 ± 9.20 76.56 ± 10.31 87.59 ± 7.68
Cantonese (Chinese) (Wong et al., 2013)	ACE-R	Dementia (n= 54) MCI (n= 50) Healthy (n= 43)	19:35 21:29 21:29	79.2 ± 6.6 76.9 ± 7.3 72.8 ± 7.5	3.7± 4.2 4.2 ± 4.2 5.6± 4.3	50.8 ± 15.4 68.2 ± 15.7 86.4 ± 8.9
Portuguese (Brazilian) (Carvalho et al., 2010)	ACE-R	AD (n= 31) Healthy (n= 62)	13:18 22:40	78.03 ± 6.74 77.82 ± 6.58	9.97 ± 5.19 10.05 ± 4.98	63.10 ± 10.22 83.63 ± 7.90
Portuguese (Brazilian) (Sobreira et al., 2015)	ACE-R	Dementia (n= 17) MCI (n= 32) Healthy (n= 30)	3:13 16:15 10:20	72.5 (53-81)* 57 (37-77) 61 (28-79)	5.50 (2-18)* 10 (0-20) 4 (1-20)	67 (32-85)* 80 (41-98) 80.5 (53-95)
Portuguese (Goncalyes et al., 2015)	ACE-R	SVD (n= 18) AD (n= 36) Healthy (n= 38)	11:7 16:20 17:21	75.50 ± 5.29 75.14 ± 4.12 76.95 ± 6.92	3.22 ± 1.73 4.64 ± 3.16 5.61 ± 2.81	55.06 ± 9.19 55.53 ± 10.16 82.11 ± 1.29
Spanish (Raimondi et al., 2012)	ACE-R	AD (n= 25) SVD (n= 32) Healthy (n= 26)	12:13 16:16 13:13	77.64 ± 5.3 75.59 ± 6.4 73.23 ± 8.9	14.48 ± 3.6 12.97 ± 4.3 14.46 ± 2.2	
Spanish (Torralva et al., 2011)	ACE-R	AD (n= 46) bvFTD (n= 41) Healthy (n= 40)	12: 34 9: 32 11: 29	73.4 ± 5.7 70.0 ± 9.3 71.5 ±5.6	12.9 ± 4.6 12.8 ± 5.1 13.0 ± 3.8	78.1 ± 9.4 64.2 ±16 94.3 ± 4.2
Spanish (Matias-Guiu et al., 2015)	ACE-III	Dementia (n= 87) Healthy (n= 130)	34:53 46:84	77.3 ± 8.4 71.0 ± 11.0	7.5 ± 4.6 9.8 ± 5.9	50.4 ± 16.0 81.8 ± 12.7

Thai (Charernboon et al., 2016)	ACE-III	Dementia (n= 30) MCI (n= 29) Healthy (n= 48)	76.9 ± 7.4 70.7 ± 7.4 65.6 ± 6.3	7.7 ± 4.2 8.6 ± 5.5 10.5 ± 5.2	43.5 ± 11.1 67.8 ± 7.4 86.1 ± 6.8

* Only the median (min-max) was reported in the article.

A blank space indicates no information is available.

Abbreviations: Dementia, Dementia type not specified or mixed subtype sample; MCI, Mild cognitive impairment; Alzheimer's, Alzheimer's disease; FTLD, Frontotemporal lobar degeneration; bvFTD, Behavioural Variant Frontotemporal lobar degeneration; SVD, Subcortical vascular dementia, SVD.

Table 2 Diagnostic Accuracy Information and Quality assessment of included studies

Language	Cut-off score	Sensitivity	Specificity	ROC curve (AUC)	STARD score/number of items	Main limitations	Rating of overall quality
French (Bastide et al., 2012)	89	98.4	98.6	0.99	32/62	No rational for the cut-off point of the reference standard, non-blinded, indeterminate data was not reported, power not calculated, time interval not stated, poorly defined sample	_
German (Alexopoulos et al., 2010)	AD 82 FTLD 83	92 88	96 96	0.99 0.97	35/62	Power not calculated, indeterminate data was not reported, no rational for the cut-off point of ACE-R, time interval not stated, poorly defined sample	_
Greek (Konstantinopoulou et al., 2011)	85	97.1	81.7	0.96	21/62	No rational for the cut-off point of ACE-R or the reference standard, indeterminate data was not reported, non-blind, power not calculated, time interval not stated	_
Italian (Pigliautile, 2012)	Young-old 79 Old-old 60	90 82	80 100	0.94 0.93	31/62	No rational for the cut-off point of ACE-R or the reference standard, non-blinded, indeterminate data was not reported, time interval not stated	_

Language	Cut-off score	Sen-sitiv-ity	Speci-ficity	ROC curve (AUC)	STARD score/number of items	Main limitations	Rating of overall quality
Japanese (Kawata et al., 2012)	80	94	94	0.98	37/62	Poorly defined sample, no rational for the cut-off point of ACE-R or the reference standard, indeterminate data was not reported, power not calculated, time interval not stated	_
Japanese (Yoshida et al., 2012)	82	99	99	0.99	44/62	Power not calculated, indeterminate data was not reported, no rational for the cut-off point of ACE-R, time interval not stated	+
Korean (Kwak et al., 2010)	78	93	95		35/62	Poorly defined sample, No rational for the cut-off point of ACE-R or the reference standard, indeterminate data was not reported, power not calculated, time interval not stated	_
Mandarin (Chinese) (Fang et al., 2014)	67	92	86	0.95	33/62	Non-blinded, indeterminate data was not reported, power not calculated, time interval not stated, poorly defined sample	_
Cantonese (Chinese) (Wong et al., 2013)	73	93	95	0.98	40/62	Indeterminate data was reported, time interval not stated, poorly defined sample	+
Portuguese (Brazilian) (Carvalho et al., 2010)	78	100	82.26	0.95	31/62	Insufficient details about the ACE-R or rational for the cut-off point, non-blinded, time interval not stated, indeterminate data was not reported, power not calculated	_
Portuguese (Brazilian) (Sobreira et al., 2015)	76	88	68	0.84	34/62	Non-blinded, power not calculated, indeterminate data was not reported,	_
Portuguese (Goncalyes et al., 2015)	SVD 72 AD 72	SVD 100 AD 97	SVD 97 AD 92	SVD 0.99 AD 0.98	34/62	No rational for the cut-off point of ACE-R or the reference standard, poorly defined sample, indeterminate data was not reported, time interval not stated	+
Spanish (Raimondi et al., 2012)	88	100	100	1.0	31/62	Poorly defined sample, no rational for the cut-off point of ACE-R, non-blinded, indeterminate data was not reported, power not calculated,	_

Language	Cut-off score	Sen-sitiv-ity	Speci-ficity	ROC curve (AUC)	STARD score/number of items	Main limitations	Rating of overall quality
						time interval not stated	
Spanish (Torralva et al., 2011)	85	97	88		32/62	Poorly defined sample, non-blinded, power not calculated, indeterminate data was not reported, time interval not stated	_
Spanish (Matias-Guiu et al., 2015)	65.6	83	80	0.92	32/62	No rational for the cut-off point of ACE-R or the reference standard, non-blinded, indeterminate data was not reported, time interval not stated	_
Thai (Charern-boon et al., 2016)	61	100	97	0.99	42/62	Time interval was not stated, poorly defined sample, indeterminate data was not reported.	+

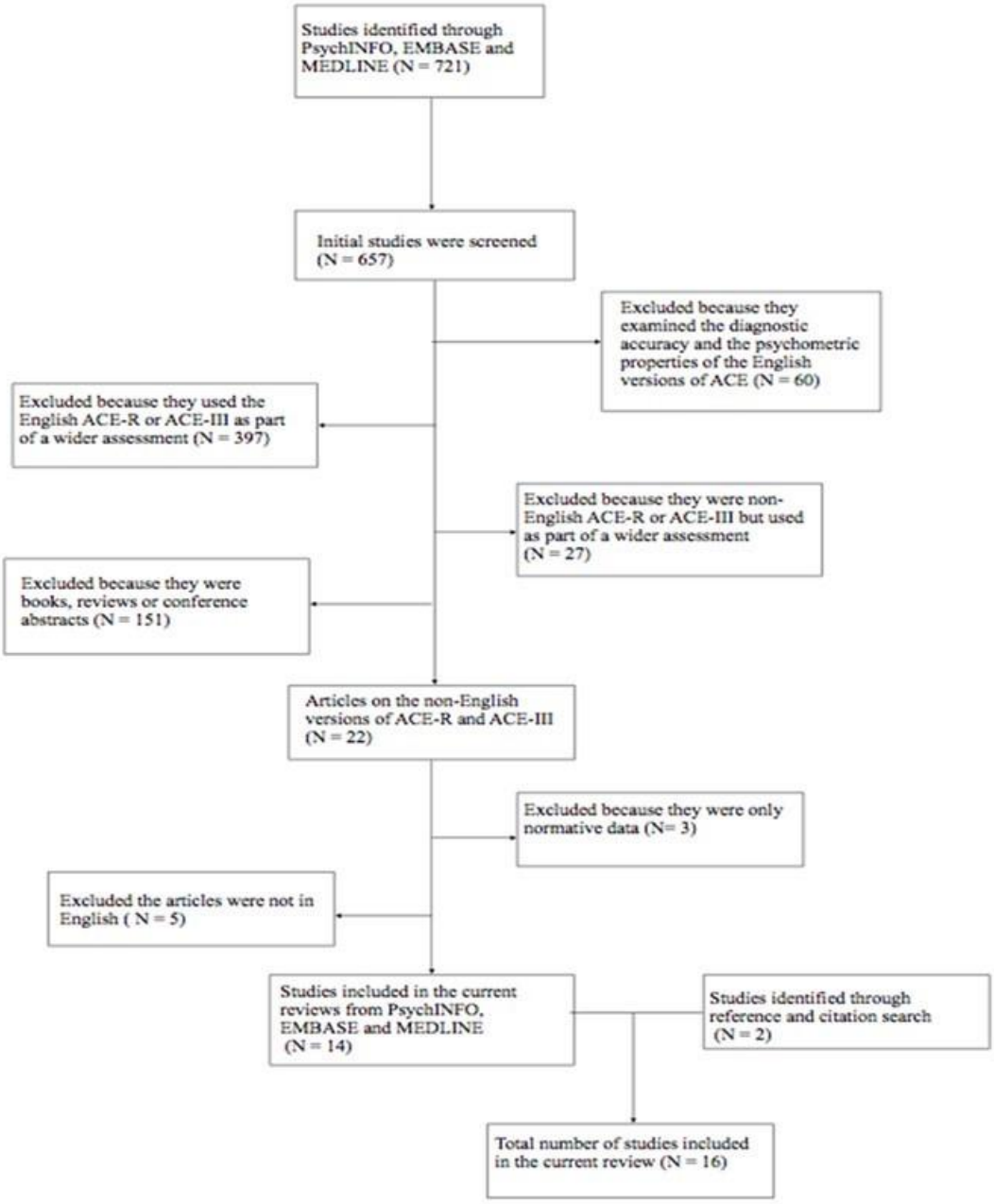
* A blank space indicates no information is available.

**ROC curve; Area under the receiver operating characteristic (AUC) curve, Alzheimer disease (AD), Frontotemporal lobar degeneration (FTLD), Subcortical vascular dementia (SVD).

*** ++ = High quality; + = medium quality and - = low quality

Figure Legend: Modified PRISMA flowchart showing search process.

Figure 1: Summary of search results



\

References:

Alexopoulos, P., Ebert, A., Richter-Schmidinger, T., Scholl, E, Natale, B., Aguilar, C. A., Gourzis, P., Weih, M., Perneczky, R., Diehl-Schmid, J., Kneib, T., Forstl, H., Kurz, A., Danek A., & Kornhuber, J. (2010). Validation of the German Revised Addenbrooke's Cognitive Examination for Detecting Mild Cognitive Impairment, Mild Dementia in Alzheimer's Disease and Frontotemporal Lobar Degeneration. *Dementia and Geriatric Cognitive Disorders*, 29, 448-456.

Bastide, L., Breucker, S., Vam den Berge, M., Fery, P., Pepersack, T., & Bier, J. C. (2012). The Addenbrooke's Cognitive Examination Revised Is as Effective as the Original to Detect Dementia in a French-Speaking Population. *Dementia and Geriatric Cognitive Disorders*, 34, 337-343.

Borsa, J. C., Damasio, F. B., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: some consideration. *Paideia,* 53, 423-432.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glaszio, P. P., Irwig, L., Moher, D., Rennie, D., de Vet, H. C., & Lijmer, J. G. (2003). Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry,* 138, W1–W12.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glaszio, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C., Kressel, H. Y., Rifai, N., Golub, R.M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *British Medical Journal,* 351-h5527.

Carvalho, V., Barbosa, M.T., & Caramelli, P. (2010). Brazilian Version of the Addenbrooke Cognitive Examination-revised in the Diagnosis of Mild Alzheimer Disease. *Cognitive and Behavioural Neurology*, 23, 8-13.

Charernboon, T., Jaisin, K., & Lerthattasilp, T. (2016). The Thai Version of the Addenbrooke's Cognitive Examination III. psychiatry Investigation, 13, 571- 573.

Dodge, H. H., Meguro, K., Ishii, H., Yamaguchi, S., Saxton, J. A., & Ganguli, M. (2009). Cross-cultural comparisons of the Mini-mental State Examination between Japanese and US cohorts. *International psychogeriatrics*, *21*(1), 113-122.

Fang, R., Gang, W.G., Huang, Y., Zhuang, J., Tang, H., Wang, Y., Deng, Y., Xu, W., Chen, S., & Ren, R. (2014). Validation of the Chinese Version of Addenbrooke's Cognitive Examination-Revised for Screening Mild Alzheimer's Disease and Mild Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, 37, 223-231.

Gonçalves, C., Pinho, M.S., Cruz, V., Pais, j., Gens, H., Oliveira, F., Santana, I., Rente, J., & José M.S. (2015). The Portuguese version of Addenbrooke's Cognitive Examination–Revised (ACE-R) in the diagnosis of subcortical vascular dementia and Alzheimer's disease. *Aging Neuropsychology and Cognition,* 22, 473-485.

Grypma, R., Mahajani, S., & Tam, E. (2007). Screening and diagnostic assessment of non-English speaking people with dementia. *Alzheimer's Australia-National Cross Cultural Dementia Network 2007.*

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. *Adapting educational and psychological tests for cross-cultural assessment, 1*, 3-38.

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., & Hodges, J.R. (2013). Validation of the Addenbrooke's Cognitive Examination-III in frontotemporal dementia and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 36, 242–250.

Kawata, K., Hashimoto, R., Nishio, Y., Hayashi, A., Ogawa, N., Kanno, S., Hiraoka, K., Yokoi, K., Lizuka, O., & Mori, E. (2012). A validation study of the Japanese version of the Addenbrooke's cognitive examination-Revised. *Dementia and Geriatric Cognitive Disorders*, 2, 29-37.

Kittner, S., White, L., Farmer, M., Wolz, M., Kaplan, E., Moes, E., Brody, J., & Feinleib, M. (1986). Methodological issues in screening for dementia: the problem of education adjustment. *Journal of Chronic Diseases, 39, 163-170.*

Knottnerus, J., & Muris J. (2002). Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: *British Medical Journal Publishing Group*, 39–59.

Konstantinopoulou, E., Kosmidis, M., Ioannidis, P., Kiosseoglou, G., Karacostas, D., & Taskos, N. (2011). Adaptation of Addenbrooke's Cognitive Examination-Revised for the Greek population. *European Journal of Neurology*, 18, 442-447.

Kwak, Y., Yang, Y., & Kim, G.W. (2010). Korean Addenbrooke's Cognitive Examination Revised (K-ACER) for differential diagnosis of Alzheimer's disease and subcortical ischemic vascular dementia. *Geriatrics and gerontology international,* 10, 295-301.

Machin, D., Campbell, M., Fayers, P., & Pinol, A. (1997). *Sample size tables for clinical studies*. 2nd ed. Blackwell Science: Oxford.

Matias-Guiu, J.A., Fernández de Bobadilla, R., Escudero, G., Pérez-Pérez, J., Cortés, A., Morenas-Rodríguez, E., Valles-Salgado, M., Moreno-Ramos, T., Kulisevsky, J., & Matías-Guiu, J. (2015). Validación de la versión española del test Addenbrooke's Cognitive Examination III para el diagnóstico de demencia." *Neurología*, 30, 545-551.

Mioshi, E., Dawson, K., Mitchell, J,, Arnold, R., & Hodges, J.R. (2006). The Addenbrooke's Cognitive Examination Revised: a brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry,* 21, 1078–1085.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2009). For The PRISMA Group: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine,* 151, 264–269.

National Institute for Health and Clinical Excellence (NICE). (2006). Dementia: supporting people with dementia and their carers in health and social care. *Social Care Institute for Excellence* www.nice.org.uk/Guidance/CG42

National Institute for Health and Clinical Excellence (NICE). (2014). Developing NICE guidelines: the manual 2014.

Pedrazaa, O., Clark, J. H., O'Bryant, S. E, Smith, G.E., Ivnik, R. J., Graff-Radford, N. R., Willis, F. B., Petersen, R. C., & Lucas, J. A. (2012). Diagnostic Validity of Age and Education Corrections for the Mini-Mental State Examination in Older African Americans. *The International neuropsychological Society,* 60, 328-331.

Philbrick, J., Horwitz, R., & Feinstein, A. (1980). Methodological problems of exercise testing for coronary disease: groups, analysis and bias. *American Journal of Cardiology,* 46, 807–12.

Pigliautile, M. (2012). Validation study of the Italian Addenbrooke's cognitive examination revised in a young-old and old-old population | Studio di validazione dell'ACE-R in lingua italiana nella popolazione degli young-old e degli old-old." *Giornale di gerontologia,* 60, 134-141.

Prince, M., Renata, B. R. & Ferri, C. (2011). The World Alzheimer Report 2011: The benefits of early diagnosis and intervention. *Alzheimer's Disease International*.

Raimondi, C., Gleichgerrcht, E., Richly, P., Torralva, T., Roca, M., Camino, J., & Manes, F. (2012). The Spanish version of the Addenbrooke's Cognitive Examination — Revised (ACE-R) in subcortical ischemic vascular dementia. *Journal of Neurological Science,* 322, 228-231.

Sobreira, E., Pena-Pereira, M. A., Eckeli, A. L., Sobreira-Neto, M. A., Chagas, M. H., Foss, M. P., Cholerton, B., Zabetian, C. P., Mata, I. F. & Tumas, V. (2015). Screening of cognitive impairment in patients with Parkinson's disease: diagnostic validity of the

Brazilian versions of the Montreal Cognitive Assessment and the Addenbrooke's Cognitive Examination-Revised. *Arquivos de neuro-psiquiatria*, 73, 929-933.

Torralva, T., Roca, M., Gleichgerrcht, E., Bonifacio, A., Raimondi, C., & Manes, F. (2011). Validación de la versión en español del Addenbrooke's Cognitive Examination-Revisado (ACE-R). *Neurología*, 26, 351-356.

Velayudhan, L., Ryu, S. H., Raczek, M., Philpot, M., Lindesay, J., Critchfield, M. & Livingston, G. (2014). Review of brief cognitive tests for patients with suspected dementia. *International Psychogeriatric*, 26, 1247–1262.

Whitney, K. A., Maoz, O., Hook, J. N., Steiner, A. R., & Bieliauskas, L. A. (2007). IQ and Scores on the Mini-Mental State Examination (MMSE): Controlling for Effort and Education Among Geriatric Inpatients. *Aging Neuropsychology and Cognition*, 5, 545-552.

Wong, L., Chan, C., Leung, J., Yung, C., Wu, K., Cheung, S., & Lam, C. (2013). A validation study of the Chinese-Cantonese Addenbrooke's Cognitive Examination Revised (C-ACER). *Neuropsychiatry Disorders and Treatment*, 9, 731-737.

Yoshida, H., Terada, S., Honda, H., Kishimoto, Y., Takeda, N., Oshima, E., Hirayama, K., Yokota, O., & Uchitomi, Y. (2012). Validation of the revised Addenbrooke's Cognitive Examination (ACE-R) for detecting mild cognitive impairment and dementia in Japanese population. International Psychogeriatric, 24, 28-37.