Diagnostic accuracy of digital solutions to screen for cognitive impairment: a systematic review and meta-analysis

Marisa Magno  
University of Aveiro

Ana Isabel Martins  
University of Aveiro

Joana Pais  
University of Porto

Anabela G. Silva  
University of Aveiro

Nelson Pacheco Rocha (✉ npr@ua.pt)  
University of Aveiro

Article

Keywords:

Posted Date: August 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3160170/v1

License: ☐  This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Abstract

Early detection of cognitive impairment is essential to initiate intervention and guarantee access to healthcare services. Digital solutions are emerging in literature as an alternative approach to cognitive screening.

Our primary goal is to synthesize the evidence on digital solutions' diagnostic ability/accuracy to screen for cognitive impairment. A secondary goal is to distinguish whether the ability to screen cognitive impairment varies as a function of the type of digital solution: paper-based or innovative digital solutions.

A systematic review and meta-analysis of digital solutions' diagnostic accuracy were conducted, including 25 studies. Digital solutions presented a variable diagnostic accuracy range. Innovative digital solutions offered at least 0.78 of sensitivity but showed lower specificity levels than the other subgroup. Paper-based digital solutions revealed at least 0.72 of specificity, but sensitivity started at 0.49.

Most digital solutions do not demand the presence of a trained professional and include an automatic digital screening system and scoring, which can enhance cognitive screening and monitoring.

Digital solutions are adequate for cognitive screening in the community and clinical practice showing sensitivity and specificity levels similar to traditional widely used paper and pencil tests. A detailed assessment of the accuracy levels of each digital solution is recommended.

Introduction

There are several known risk factors (e.g., diabetes, hypertension, hypercholesterolemia, depression, physical frailty, low education level, or low social support level) contributing to neurodegenerative diseases such as Alzheimer, Parkinson, Huntington, or frontotemporal dementia, but ageing is the strongest one\(^1\)\(^–\)\(^6\). Therefore, the prevalence of these diseases increases as our society ages.

Mild cognitive impairment (MCI), an intermediate stage between normal ageing and dementia, is characterized by an objective cognitive decline in one or more cognitive domains (e.g., memory, attention, language, or executive function) without any significant impairment in daily-life activities\(^7\) and may be associated with a variety of underlying causes, including Alzheimer's pathophysiology\(^7\)\(^–\)\(^9\). In turn, dementia is a major neurocognitive disorder that is characterized by a significant decline in one or more cognitive domains that interferes with a person's independence in daily activities\(^10\). Although there is evidence that patients with MCI may experience reversion to cognitive normality\(^10\)\(^,\)\(^11\), there is a high probability that this condition will progress to dementia. Therefore, early detection of MCI is critical to effectively initiate the intervention (including counseling, psychoeducation, cognitive training, and medication\(^12\)), and guarantee to both patients and relatives access to relevant healthcare services\(^13\).

However, MCI is significantly misdiagnosed due to a diverse set of barriers, namely the high prevalence of comorbidities among older adults, lack of expertise or limited confidence of the practitioners, the short duration of most primary care visits, limitations of the assessment instruments, or the inadequacy of
Despite these barriers, there are many screening tests that provide a quick evaluation of cognitive and functional aspects. At present, two of the most well-known cognitive screening tests are the Mini-Mental State Examination (MMSE)\(^{14}\) and the Montreal Cognitive Assessment (MoCA)\(^{15}\), which include tasks to assess multiple cognitive domains. In addition to MMSE and MoCA, other currently available cognitive tests also encompass multiple cognitive domains, including the Neuropsychiatry Unit Cognitive Assessment Tool (NUCOG)\(^{16}\), the Saint Louis University Mental Status examination (SLUMS)\(^{17}\), the Self-Administered Gerocognitive Examination (SAGE)\(^{18}\), or the Addenbrooke’s Cognitive Examination III (ACE-III)\(^{19}\). In turn, screening tests such as Alzheimer Quick Test (AQT)\(^{20}\), Scenery Picture Memory Test (SPMT)\(^{21}\), Memory Impairment Screen (MIS)\(^{22}\), Mini-Cog\(^{23,24}\) or Clock Drawing\(^{25}\) measure one or two cognitive domain (i.e., attention for the AQT, episodic memory for the SPMT, memory and orientation for the MIS, Memory and visuospatial abilities for the Mini-Cog, or executive functions and visuospatial abilities for the Clock Drawing) but require less than five minutes to be applied\(^{26}\).

Computerized solutions to support neuropsychological tests have existed for several decades and might use different types of interaction devices, be it computers or handheld devices, or virtual reality\(^{27}\). Some solutions offer adaptations of paper-based tests to evaluate specific cognitive domains\(^{28}\) (e.g., Trail Making Test or Simple and Complex Reaction Time) or multiple cognitive domains\(^{27}\) (e.g., MoCA\(^{29}\), MMSE\(^{30}\), SAGE\(^{31}\)), while other solutions (e.g., Memoro\(^{32}\), NutriNet-Santé Cognitive Test Battery (NutriCog)\(^{33}\), Cambridge Neuropsychological Test Automated Battery (CANTAB)\(^{34}\)) were specifically developed to be applied using electronic means.

Furthermore, the scientific literature reports the development of new instruments able to monitor individuals in their residential environments\(^{27}\). In this respect, smart devices (e.g., smartphones, smartwatches, or smart-home devices) may collect data on individuals’ habits and patterns, which can be analysed to detect subtle changes that may indicate a decline in cognitive performance\(^{35}\). Moreover, serious games and virtual reality are alternative approaches to cognitive screening and may also reduce feelings of test anxiety\(^{35,36}\).

The primary goal of the current review is to synthesize the evidence on digital solutions’ diagnostic ability/accuracy to screen for cognitive impairment. A secondary goal is to distinguish whether the ability to screen cognitive impairment varies as a function of the type of digital solution: (1) based, in essence, on pre-existing traditional tests named paper and pencil tests (abbreviated as paper-based digital solutions throughout this article); (2) developed from inception to be applied by electronic means (abbreviated as innovative digital solutions throughout this article).

**Methods**
Protocol registration

This systematic review was conducted considering Cochrane’s recommendations for systematic reviews of diagnostic test accuracy\textsuperscript{37}, and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines\textsuperscript{38}. The protocol was registered in PROSPERO\textsuperscript{39}.

Search strategy and study eligibility criteria

The search was performed in Scopus, Web of Science, and Pubmed in September 2022. Databases were searched from inception to August 2022 using a combination of the following keywords: “cognitive screening”, “cognitive test”, “memory screening”, “memory test”, “attention screening”, “attention test”, “computer”, “game”, “gaming”, “virtual”, “online”, “internet”, “mobile”, “app”, and “digital”.

To be included in this review, studies had to: i) focus on any digital solution (the index test) that can be used as a generic community-based screening tool for cognitive impairment, and that was self-administrated; ii) include a sample of adults (≥ 18 years old) or older adults (≥ 65 years old); iii) compare the digital solution with a reference standard (i.e., other instrument, a clinical assessment or a combination of these); iv) be written in English; v) follow a case-control, cross-sectional, or cohort designs that at some point allow for the identification of two groups (with and without cognitive impairment); vi) report at least one diagnostic accuracy property, namely sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), or, alternatively, provided enough data to calculate these indicators. Studies that included participants that have any acute neurological condition, cognitive impairment, or were institutionalized were excluded. In addition, studies that reported on digital solutions used as a monitoring tool for patients with an existing cognitive impairment diagnosis were also excluded.

Study selection process and data extraction

All retrieved references were imported to the Mendeley Desktop software, Version 1.19.8, and checked for duplicates. Then, one author (NPR) screened the titles and abstracts of all citations according to predefined study selection criteria. Full texts of potentially relevant articles were retrieved and independently screened by two randomly chosen authors from a set of three authors (AGS, AIM, and NPR), to verify if the inclusion and exclusion criteria were met. If a consensus could not be reached between the two authors, the third author was consulted.

Data were extracted by two authors (AIM and MM) using an electronic form developed for this purpose. The extracted information was double-checked. The information extracted from each study was: author(s) and year of publication; sample sizes and characteristics (e.g., sex, age); type and name of digital solution (index test); type and name of reference standard test; and diagnostic accuracy property (e.g., estimates of sensitivity and specificity). For each study, the information to construct a 2 by 2 contingency table for each index test, including the number of True Positive (TP), False Positive (FP), True Negatives (TN), and False Negatives (FN) was also extracted. When these counts were missing from the
study, the data needed (e.g., sample size, number of participants with the target condition, estimates of sensitivity and specificity, and estimates of PPV and NPV) were extracted.

The results presented in this review consider the best cut-off reported for each index test in each study for achieving the best diagnostic ability to screen for cognitive impairment. If more than one index test result were presented (e.g., different thresholds), we chose the results given by the better cut-off reported, considering the reference standard test. Sensitivity and specificity depend on the cut-off value considered positive to identify the target condition (i.e., generally, the higher the sensitivity, the lower the specificity, and the higher the specificity, the lower the sensitivity) \(^\text{40,41}\).

**Methodological quality assessment**

Each manuscript was independently assessed by two randomly chosen authors from a set of three authors (AGS, AIM, and NPR). Disagreements were solved by consensus or discussion with the third author. The assessment of the eligible studies’ methodological quality was performed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2). QUADAS-2 is a validated tool used to evaluate the quality of diagnostic accuracy studies \(^\text{42}\), and comprises 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of the risk of bias through signalling questions that can be answered “yes”, “no”, or “unclear”. The first 3 domains (i.e., patient selection, index test, reference standard) are also assessed based on applicability concerns. Overall concerns about the risk of bias and applicability per domain are then rated as “high”, “low” or “unclear”. These results are defined for each domain \(^\text{42}\). A pilot test for bias risk assessment was conducted, using studies that were not eligible for this review.

**Data analysis**

For each study, a 2 by 2 contingency table was constructed, including the TP, FP, TN, and FN for the index tests. If these values were not reported in the manuscript, they were calculated from the data extracted from each study (sample size, number of participants with the target condition, estimates of sensitivity and specificity or estimates of PPV and NPV), following the Cochrane’s recommendations for calculation of TP, FP, TN, and FN \(^\text{43}\). Approximations and rounding were made, if necessary. Calculations were double-checked and cross-checked against the accuracy measures presented in the study.

For the meta-analysis, we used hierarchical random-effects models and Receiver Operating Characteristic (ROC) analysis. Hierarchical Summary Receiver Operating Characteristics (HSROC) models were implemented for the estimation of a Summary Receiver Operating Characteristic (SROC) curve. This method provides information on the test performance, describing variations in sensitivity and specificity \(^\text{40,44}\), considering Cochrane’s recommendations for systematic reviews of diagnostic test accuracy \(^\text{37}\).

A SROC plot was developed, presenting the results of each study in the ROC space.
Data was subdivided into two analyses considering the type of index test used: (i) paper-based digital solutions; (ii) innovative digital solutions.

A summary point for sensitivity or specificity can't be calculated in this meta-analysis because both reference standard and index tests presented differ between studies, i.e., the studies need to be more homogeneous to calculate an estimative of several studies\(^ {40,45}\).

To perform the analysis, a web application developed in R using Shiny, the MetaDTA, was used\(^ {45,46}\). Among other features, MetaDTA allows to incorporate into the graphical representation the data obtained by the QUADAS-2 tool.

**Results**

**Study selection**

The results of the search performed on databases are presented in Fig. 1. A total of 8557 articles were identified. In the first step, 3452 duplicate articles, 311 reviews or surveys, 171 references without abstract or without authors, 141 articles not written in English, and one article retracted, were removed. After that, 4481 articles remained for screening based on title and abstract. Of these, 4373 articles were excluded because they did not meet the outlined inclusion criteria, whereas 108 full-text articles were thought potentially eligible. Twenty-five studies were included in this systematic review according to eligibility criteria (Fig. 1).

**Methodological quality**

Results of QUADAS-2 assessment are summarized in Table 1 and displayed in Fig. 2. The risk of bias in the flow and timing domain is low in 17 out of the 25 diagnostic accuracy studies evaluated. The risk of bias in the reference standard and index text domains is unclear in 16 out of the 25 studies. Eighteen studies present high risk of bias in the patient selection domain. Concerns about risk applicability for the patient selection domain were rated as high in 12 studies, low in 12 studies, and unclear in 1 study. Concerns on the applicability in the domains of reference standard and the index test were rated as low for most of the studies. The exceptions were 3 studies\(^ {47–49}\) that scored high in the domain of reference standard and 1 study\(^ {50}\) that scored high in the domain of the index test.

(Table 1 about here)

**General overview of included studies**

The studies included in this review adopted distinct definitions of the target condition. The majority of them (18 out of the 25) defined the target condition as Mild Cognitive Impairment (MCI), including two studies that used different terminologies, namely, Subtle Cognitive impairment (SCI) and Mild Cognitive Dementia (MCD). One study considered amnestic Mild Cognitive Impairment (aMCI) as the target condition. Three studies included MCI or Mildly Impairment (MI) and other clinical conditions (e.g., MCI &
Dementia, MCI & Mild Alzheimer's Disease, MI & Impairment). Three studies specified Cognitive Impairment (CI) as the target condition; one of them included a significant percentage of severe cases of dementia and was excluded from the meta-analysis\textsuperscript{51}, as the other studies didn't have a substantial proportion of severe cases of dementia in their samples.

Twenty-three studies out of the 25 studies included used distinct instruments and/or clinical assessment processes as reference standards.

The index tests differed across all the included studies, except for 2 studies that used the MemTrax test (MTX)\textsuperscript{52,53}, and 2 studies that used the Brain Health Assessment test (BHA)\textsuperscript{54,55}. Subgrouping the studies, there were 16 studies reporting on paper-based digital solutions and 9 studies reporting on innovative digital solutions. The characteristics of the included studies can be found in Supplementary Information.

**Meta-analysis results**

The meta-analysis included 24 studies. As previously reported, one study was excluded because the sample included many individuals with severe cases of dementia\textsuperscript{51}. The HSROC models for estimation of a summary ROC (SROC) curve project the results of each of the 24 studies in the ROC space, with the covariate of the index test subgroup (Fig. 3). The results indicated that sensitivity values for index tests vary between 0.49 and 0.95 and specificity values between 0.50 and 0.91. Innovative digital solutions presented values for sensitivity that vary between 0.78 and 0.94 and specificity that vary between 0.50 and 0.90. The sensitivity of paper-based digital solutions varies between 0.49 and 0.95 and the specificity varies between 0.72 and 0.91.

Figures 4 and 5 display the sensitivity and specificity values for index test subgrouped according to the type of test (paper-based digital solutions or innovative digital solutions). Each study is represented with a circle in the ROC space. The forest plots of sensitivity and specificity are presented per study.

The results of each of the 24 studies in the ROC space with quality assessment obtained by the QUADAS-2 tool, namely concerning the risk of bias and quality concerns, are present in Supplementary Information.

**Discussion**

This systematic review assessed the diagnostic accuracy of digital solutions used for cognitive screening, further analysing whether these were paper-based digital solutions or innovative digital solutions. Overall, results suggest that digital solutions are adequate for cognitive screening. The results of the 24 studies projected in the ROC space showed that index tests were above the line of zero discrimination\textsuperscript{41}, demonstrating that the accuracy of all these digital index tests was better than random classification.
The index tests assessed were quite variable, with sensitivity levels varying between 0.49 and 0.95 and specificity levels between 0.50 and 0.91. The index tests classified as innovative digital solutions offered at least 0.78 of sensitivity but showed lower specificity levels than the other subgroup (between 0.50 and 0.90). The index tests classified as paper-based digital solutions revealed at least 0.72 of specificity, but sensitivity started at 0.49 (and eight studies out of 15 reported sensitivity values below 0.78).

The study that reported higher sensitivity values among those tests classified as paper-based digital solutions reported on the Beijing version of the MoCA (sensitivity = 0.95; specificity = 0.87). This performance was similar to the MoCA paper and pencil version of the instrument for detecting mild cognitive impairment in elderly Chinese living in communities (sensitivity = 0.81; specificity = 0.83) suggesting that both versions are equivalent. For the subgroup of tests classified as innovative digital solutions, the Digital Screening System showed the highest sensitivity/specificity levels (sensitivity = 0.85, specificity = 0.90). These two index tests were assessed against robust reference tests (a clinical assessment performed by a team of health professionals including a neurologist, a geriatrician and a psychiatrist (MoCA-CC), and experienced doctor and neuropsychologist (Digital Screening System)).

The assessment by a team of specialists is the gold standard for cognitive evaluation. The Digital Screening System aims to assess visuospatial constructional capabilities, visual memory function, and cognitive functions, such as visuospatial abilities, visual episodic memory, organization skills, attention, and visuomotor coordination. It is based on the neuropsychological test Rey-Osterrieth Complex Figure and used a data-driven convolutional neural network architecture through transfer learning and deep learning methods. Despite being developed from inception to be applied by electronic means, most innovative digital solutions are inspired by traditional neuropsychological tests. In the innovative digital solutions subgroup, the index test Virtual Supermarket Program (VSP) stands out for using virtual reality game–based tests in screening for MCI in older adults, showing an attempt to develop a test that uses a task from daily life, potentially increasing its ecological validity. Interestingly, this test showed high sensitivity and specificity values (sensitivity = 0.85; specificity = 0.80), suggesting that there is value in exploring the use of game-based tests to screen for cognitive impairment.

The index test that presented lower sensitivity/specificity in the subgroup of index tests based on paper and pencil tests is the MemTrax test (MTX) (sensitivity = 0.49, specificity = 0.78). This index test was based on the Continuous Recognition Task (CRT) paradigm. Among the index tests developed from inception to be applied as digital solutions, the Cognivue, and the CogEvo, showed the lowest sensitivity/specificity levels (Cognivue: sensitivity = 0.78, specificity = 0.50; CogEvo: sensitivity = 0.78, specificity = 0.54). These three index tests that presented the lowest sensitivity/specificity levels were compared against a reference standard consisting of only brief cognitive screening instruments (e.g., MoCA, SLUMS, and MMSE, respectively). The MoCA paper and pencil test demonstrated a sensitivity of 90% and a specificity of 87% for detecting MCI. The MMSE paper and pencil test showed a pooled sensitivity of 85% and a specificity of 86% in a non-clinical community setting. The SLUMS paper and pencil test for detecting MCI in patients with less than High School Education had a sensitivity of 92% and a specificity of 81%, and in patients with High School Education or more, a sensitivity of 95% and a
specificity of 76%\(^1\)\(^7\). Despite the relatively high sensitivity and specificity levels, these instruments are not the gold standard for cognitive assessment and, therefore, their use might have affected the sensitivity and specificity calculations of the index test and, certainly, undermines the confidence in the reported results.

Early detection of cognitive impairment is critical to an early intervention\(^1\)\(^2\),\(^1\)\(^3\). Index tests with high sensitivity levels are essential when the goal is to identify a serious disease with available treatment\(^4\)\(^1\),\(^6\)\(^4\). Digital solutions emerge as a valid alternative for cognitive screening, potentially enhancing cognitive screening and monitoring in the general and clinical population since most do not require the presence of a trained professional and have an automatic digital screening system and scoring\(^4\)\(^7\),\(^5\)\(^6\),\(^5\)\(^8\), decreasing the costs associated with its use and facilitating the screening for high numbers of individuals. We found in the current review that there are digital solutions with levels of sensitivity and specificity that are similar to widely used paper and pencil tests, such as the MMSE\(^1\)\(^4\) or MoCA\(^1\)\(^5\), suggesting that digital solutions can be used for cognitive screening.

The quality of the included studies as evaluated by the QUADAS-2 suggest risk of bias in the patients’ selection domain, including for those studies presenting the digital index tests with higher sensitivity/specificity values. A test accuracy study with a high risk of bias in the participant selection domain can give inflated estimates of sensitivity and specificity\(^6\)\(^5\). Despite the different definitions used by the studies, we found relative homogeneity in the target condition, as they all focus on diagnostic ability/accuracy to screen for cognitive impairment. Nevertheless, the reference standards display substantial methodological heterogeneity. This heterogeneity was due to significant variations in the instruments adopted and/or the clinical assessment process followed across the studies. A similar reference standard, preferentially a gold standard, should be applied across studies to facilitate accuracy comparisons and increase the confidence on the results\(^3\)\(^7\).

Considering the heterogeneity in reference standards and index tests across studies, a meta-analysis with a summary point is not recommended\(^4\)\(^0\). A meta-analysis should be performed when studies are homogeneous in terms of the reference and index tests\(^4\)\(^0\),\(^4\)\(^5\).

Future studies should adopt more rigorous, at-random sampling procedures to reduce the probability of risk of bias from patient recruitment. Also, future diagnostic should consider adopting a similar gold standard as a reference test to facilitate comparisons and increase the confidence on the results. Gold standards involve the assessment of multiple cognitive domains, including memory, by qualified professionals\(^5\)\(^9\). However, investigators and practitioners must consider the diagnostic properties of the different digital solutions and the reference test against which the accuracy values were calculated to make an informed choice.

**Conclusion**
Digital solutions can be used for cognitive screening in the community and clinical practice with levels of sensitivity and specificity similar to existing paper-based questionnaires. A careful assessment of the accuracy levels of each digital solution is needed before its use.

**Declarations**

**Acknowledgements**

This study was funded by the PRR – Plano de Recuperação e Resiliência and by the NextGenerationEU funds at Universidade de Aveiro, through the scope of the Agenda for Business Innovation “HfPT – Health from Portugal” (Project no. 41 with the application C644937233-00000047)

**Author contributions statement**

A.I.M., J.P., A.G.S., and N.P.R. contributed to the conception of the work, and all authors contributed to the design; M.M. wrote most of the manuscript, while A.G.S. and N.P.R. also made significant contributions during the writing and reviewing processes; M.M. conducted the data analysis, and prepared tables and figures; M.M., A.I.M., A.G.S., and N.P.R. contributed to acquisition, analysis, and interpretation of data; all authors reviewed the final manuscript.

**Data availability statement**

All data needed to evaluate the conclusions in the paper are present in the paper. Additional data related to this paper are available on request from the corresponding author, N.P.R..

**Additional information**

**Competing interests statement**

The authors declare no competing interests.

**References**


58. Cheah, W., Hwang, J., Hong, S., Fu, L. & Chang, Y. A Digital Screening System for Alzheimer Disease Based on a Neuropsychological Test and a Convolutional Neural Network: System Development and Validation Corresponding Author : 10, 1–16 (2022).


60. Yan, M. et al. A Virtual Supermarket Program for the Screening of Mild Cognitive Impairment in Older Adults: Diagnostic Accuracy Study Corresponding Author : 9, 1–12 (2021).


Table

Table 1 is are available in the Supplementary Files section
Figures

Identification of studies via databases and registers

Identification

Records identified from:
- Scopus (n = 3733);
- Web of Science (n = 1244);
- PubMed (n = 3580).

Records removed before screening:
- Duplicate records removed (n = 3452)
- Records marked as ineligible by automation tools (n = 0)
- Records removed for other reasons (n = 624)

Screening

Records screened (n = 4481)

Records excluded (n = 4373)

Reports sought for retrieval (n = 108)

Reports not retrieved (n = 0)

Reports assessed for eligibility (n = 108)

Reports excluded:
- Unreported or incomplete accuracy measures (n = 32)
- Comparison between a computerized test and its paper version (n = 11)
- Not a screening tool (n = 10)
- Designed for a specific health condition (n = 12)
- Out of scope (n = 7)
- Not self-administered (n = 6)
- Comparison between a computerized versions for iPad and PC (n = 2)
- Not a digital solution (n = 2)
- Repeated sample (n = 1)

Included

Studies included in review (n = 25)
Reports of included studies (n = 25)

Figure 1

PRISMA flow diagram.
**Figure 2**

QUADAS-2 assessment results – graphical display.
Figure 3

Random-effects meta-analysis – summary ROC (SROC) curve with covariate of the index test subgroup.
Figure 4

(a) Random-effects meta-analysis – summary ROC (SROC) curve of innovative digital solutions; (b) Forest plot of sensitivity and specificity.
Figure 5

(a) Random-effects meta-analysis – summary ROC (SROC) curve of paper-based digital solutions; (b) Forest plot of sensitivity and specificity.

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation31072023.pdf
- Table1.docx