scientific reports

Check for updates

OPEN

Optimizing mobile cognitive assessment reduces administration time while maintaining screening accuracy in older adults

Hyun Jeong Ko[1,8], Jin Sung Kim[2,3,7,8], Whani Kim[1], Byung Hun Yun[1,4], So Yoon Park[1,4], Dong Han Kim[1], Ui Jun Kwon[1], Sang Kwon Lim[1], Bo Ri Kim[5], Jee Hang Lee[2,3,7]✉, Geon Ha Kim[6]✉ & Jin Woo Kim[1,4]

Growing dementia prevalence underscores the need for efficient screening methods, but lengthy digital assessments often cause fatigue among older adults. To address this, we developed and validated the Digital Assessment of Cognitive Impairment (DACI), a brief mobile application designed to accurately identify cognitive impairment (CI). Initially, 304 older adults (272 healthy, 32 cognitively impaired) completed both a pencil-and-paper Cognitive Impairment Screening Test (CIST) and a full-length DACI. The best-performing CatBoost model achieved an area under the curve (AUC) of 0.813, with sensitivity of 0.903, requiring an average completion time of 321 s. Subsequent feature selection identified two essential subtests for a compact DACI. This compact version was then validated with an additional 297 participants (227 healthy, 70 cognitively impaired), achieving improved diagnostic performance (AUC = 0.871) in only 91 s. DACI effectively distinguishes cognitively impaired older adults with minimal test duration, potentially reducing fatigue and improving adherence. By providing rapid, accurate remote assessment without additional equipment, DACI may significantly enhance accessibility of cognitive screening among older adults. Future studies are warranted to validate DACI's feasibility and performance in unsupervised home settings.

Keywords Remote cognitive assessment, Cognitive impairment, Mobile device, Optimal test design, Digital health

The global prevalence of dementia is projected to reach 152 million by 2050[1]. Early detection and timely intervention for cognitive impairment (CI) can delay significant functional decline in Alzheimer's disease (AD)[2–11]. Recent medical advances, including FDA-approved treatments, have shown promising outcomes in slowing cognitive decline[12], making early screening crucial for effective intervention[13].

Typical cognitive assessments require qualified neuropsychologists, appropriate facilities, and accessible locations[14]. However, the growing aging population and anticipated healthcare professional shortage might limit access to in-person assessments[4,15]. Mobility challenges, economic constraints, and infection risks further restrict service availability[16–18]. Additionally, the expertise required for scoring and interpretation limits daily assessment capacity[19].

Researchers have explored digital technologies for cognitive screening to address these challenges[20–25]. The COVID-19 pandemic accelerated the development of remote cognitive assessments and app-based testing[8,26,27]. These digital tools enable remote participation through web or mobile applications with minimal staff assistance. Digital biomarkers in addition enhance assessment resolution, increasing sensitivity and accuracy[3,23,28–37]. In conjunction with this, these tools facilitate rapid result delivery through AI and data science, analyzing both

[1]Department of Digital Therapeutics Research and Development, HAII Corp, Seoul, Republic of Korea. [2]Department of Human-Centred AI, Sangmyung University, Seoul, Republic of Korea. [3]Graduate School of AI and Informatics, Sangmyung University, Seoul, Republic of Korea. [4]Graduate School of Cognitive Science, Yonsei University, Seoul, Republic of Korea. [5]Ewha Medical Research Institute, Ewha Womans University, Seoul, Republic of Korea. [6]Department of Neurology, Ewha Womans University Mokdong Hospital, Ewha Womans University School of Medicine, Seoul, Republic of Korea. [7]Institute for Advanced Intelligence Study, Daejeon, Republic of Korea. [8]Hyun Jeong Ko, Jin Sung Kim have contributed equally to this work and share first authorship. ✉email: jeehang@smu.ac.kr; geonha@ewha.ac.kr

active user inputs and passive digital biomarkers automatically[38–45]. These advances improve both accessibility and resource utilization efficiency.

However, these digital tools face early-stage challenges, particularly in usability. Poor user experience design and small interfaces impede adoption among older adults[46,47]. Lower digital literacy in older populations compounds these difficulties, whilst prolonged engagement can increase cognitive and physical fatigue[48,49]. Consequently, large-scale controlled studies have been reluctant to adopt digital tools, leaving their effectiveness largely unvalidated[3]. Whilst digital applications are recognised for conserving resources[41], concerns persist about performance inconsistencies and variable task duration across age groups[50].

Developing optimised digital tools that maximise screening accuracy whilst minimising test duration becomes crucial within this context. The potential for digital assessment tools to distinguish cognitive impairment from healthy controls more accurately in reduced time offers significant promise for consistent performance across age groups. In this regard, this paper aims to develop and validate the Digital Assessment of Cognitive Impairment (DACI)—a cost-effective, mobile-based tool for remote cognitive impairment screening in older adults. Using machine learning, DACI detects cognitive decline without additional sensor equipment. We also introduce a compact DACI version with fewer tasks that maintains accuracy whilst reducing administration time. We anticipate this solution will demonstrate consistent performance and high validity for cognitive impairment assessment.

Methods

Figure 1 illustrates the overall procedure of the study. In summary, the study was initiated with the design of DACI, followed by an initial human subjects experiment involving 304 participants using the full-length, comprehensive version. A predictive machine learning (ML) model was subsequently developed based on data collected during this first experiment. This process brought about the model with the highest area under the curve (AUC) among several predictive models trained on data derived from the full set of tests using various ML algorithms.

Building on this model, we conducted an optimisation process to identify the minimal set of tasks within DACI by assessing the relative importance of each test. This involved enumerating the feature importance from the collected data. We then applied constrained optimisation, utilising an exhaustive search algorithm aimed at minimising assessment time. The ultimate objective was to identify the optimal set of features that could achieve predictive performance statistically equivalent to that of the best AUC model.

Finally, the compact version of DACI, compiled from the selected tests, was subsequently employed in a follow-up experiment involving an additional 297 participants. Data collected from this experiment were used to develop a new predictive model using machine learning (ML) algorithms. Afterwards, we evaluated the performance of the most accurate predictive model by comparing its screening accuracy and completion time to those of the original model, which was trained on data from the full-length DACI.

Designing a full-length, comprehensive version of DACI

The DACI is, aforementioned, a mobile-based screening tool for cognitive impairment. It primarily consists of four cognitive tasks, each designed to evaluate key cognitive domains, including executive function, associative memory, working memory, and calculation ability. We designed a dedicated task for each domain as described below. Participants performed the multiple sessions each task (Table 1).

Design principle

Based on a comprehensive and systematic review, we developed DACI as an expanded digital version of the pencil-and-paper-based CIST test. We digitized the original test and enhanced each task according to findings from our literature review. Through our analysis, we identified executive function, associative memory, working

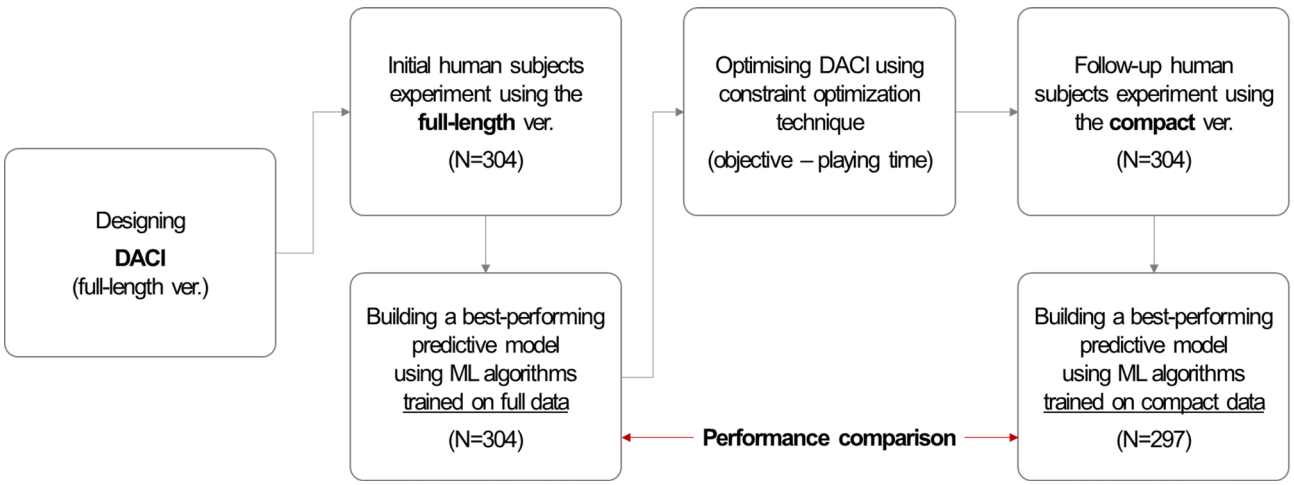


Fig. 1. An overview of the whole experiment procedure.

DACI task name	Cognitive domain	References
Stroop test	Executive function	Scarpina and Tagini (2017), Kasai et al. (2020)
Symbol association	Associativememory	Wechsler (1997)
Self-ordered pointing task	Working memory	Geva et al. (2016)
Quantitative comparison task	Calculation	Martin et al. (2003)

Table 1. Research background for DACI task design.

memory, and calculation as the cognitive functions most relevant for early detection of Alzheimer's disease (AD). We subsequently selected and digitized appropriate cognitive tasks for each domain. These tasks included the Stroop test, symbol association, self-ordered pointing task, and arithmetic tasks (cognitive domains and DACI tasks in Alzheimer's disease are shown in Supplementary Table 1).

The National Institute on Aging-Alzheimer's Association reported that pathophysiological changes in the medial temporal lobe, including the hippocampus, might be associated with probable AD and dementia[51]. We therefore focused first on addressing cognitive domains relevant to hippocampal functions, then designing neuropsychological tests based on these functions.

A previous study found that the newly developed CQ score, which incorporates the Stroop test and simple calculation, showed significant correlation with hippocampal volume[52]. Supporting this finding, researchers demonstrated that Stroop interference effects were significantly greater for the MCI group compared to the cognitively normal (CN) group[53,54]. Based on these findings, we selected and digitized the numeric Stroop test for inclusion in DACI.

We also considered that AD is highly correlated with the diagnosis of amnestic mild cognitive impairment (aMCI). According to multiple studies, many individuals develop AD within 3–6 years after receiving an aMCI diagnosis[55–57]. Early cognitive changes in aMCI can be particularly sensitive to associative recall[58], which led us to include the symbol association test.

Researchers also found that prodromal decline in working memory preceded decline in other cognitive domains[59,60]. Based on this evidence, researchers have used the SOPT to examine changes in working memory associated with normal aging[61–64]. These findings led us to include SOPT in DACI.

Early manifestations of AD include calculation deficits. Compared to patients with mild AD, patients with moderate AD performed worse on written arithmetic tasks ($P<0.002$). Patients with mild AD showed preservation of single- and multiple-digit addition and subtraction operations, as well as single-digit division operations, on WRAT-3 problems. In contrast, only single-digit addition was preserved in patients with moderate AD[65].

To be effective as a screening test for older adults, the tool was designed to be completed in approximately 5 min, considering the importance of minimizing fatigue[48,49]. Additionally, we employed an auditory instruction technique that has been proven effective for older populations[66].

Executive function

To assess executive function, we developed a Numeric Stroop task adapted from the past research on Stroop test[52,67]. Mismatched digits and the number of digits served as the main stimuli in the Numeric Stroop task (e.g., 1111, 111, 444, 3, 44, 2), although matched stimuli (e.g., 333, 22, 4444) were occasionally included.

The task consisted of two phases, each comprising ten questions. In the first phase, participants were asked to select a number that matched the one presented on the screen. For example, if '7' was displayed, participants were required to select the same number, '7', as the correct response. During the first phase, we designed the task to elicit the reverse Stroop effect. This effect occurs when the unattended dimension (i.e., the number of digits) interferes with the processing of the attended dimension (i.e., digit identity)[68]. To elicit this interference, the task was designed so that participants required more time to read incongruent digit stimuli (e.g., 111, 33, 4, 2), thereby demonstrating the reverse numeric Stroop effect.

In the second phase, participants were asked to count the total number of items displayed and select the matching quantity. For example, if '111' appeared on the screen, participants were expected to choose '3' representing the total count of the displayed digits (Fig. 2A). During this phase, we designed the task to elicit the Stroop effect. For this effect to emerge, the unattended dimension (i.e., digit identity) must interfere with the processing of the attended dimension (i.e., numerosity)[68]. Accordingly, the task was designed to induce longer response times when participants counted incongruent digit stimuli (e.g., 111, 33, 444, 2), thus producing a numeric Stroop effect.

Associative memory

The *Remembrance of paired symbols task* was designed to assess associative memory function[69]. The task consisted of two phases: an associative learning phase and a quiz phase. During the associative learning phase, participants were instructed to memorise six associations sequentially displayed on the screen, each consisting of two symbols (Fig. 2B-left). Six pairs of symbols were presented in the initial session, with each pair shown for 4 s. In the quiz phase, participants were asked to select the symbol in the right panel that was associated with the symbol displayed in the left panel (Fig. 2B-right). This phase involved six matching tasks, corresponding to the six symbol pairs shown in the learning phase. In this regard, participants were required to answer six questions, each assessing their capacity to recall the previously memorised symbol associations.

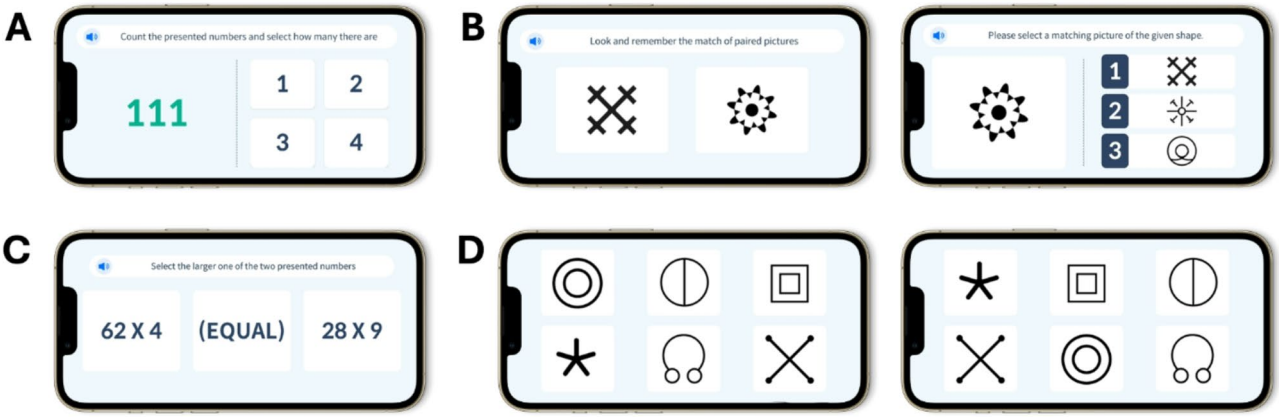


Fig. 2. Snapshot of the DACI tasks. (**A**) Numeric Stroop task; a screen of part A. (**B**) Remembrance of paired symbols task. (**C**) Quantitative Comparison task. (**D**) Find an Unselected Symbol task.

Variables	HC	MCI	*P* value
Age (years) (Range, SD)	67.03	71.59	< 0.0001
	(51–86, 7.05)	(56–91, 7.08)	
Education (years) (Range, SD)	10.47	10.47	n.s.
	(0–20, 4.5)	(0–20, 3.92)	
Gender (All [Male])	272	32	n.s.
	(109)	(12)	
CIST score (Range, SD)	26.58	23.15	< 0.0001
	(18–30, 2.87)	(10–26, 3.78)	

Table 2. Demographic characteristics of the study participants. *For Age, Education, and CIST score, homogeneity of variance was first assessed with Levene's test ($\alpha = 0.05$); when variances were equal we applied the standard independent-samples t-test, and when unequal we used Welch's t-test. Gender differences were evaluated with a χ^2 test. "n.s." indicates non-significant findings.

Calculation

To assess calculation capacity, we designed a *quantitative comparison task* based upon the *multi-digit written arithmetic test* in the Wide-Range Achievement Test-3 (WRAT-3)[65]. Here, participants were asked to select the equation with the larger outcome expected from two presented options. If the outcome of both equations were identical, participants were instructed to select the 'equal' button (Fig. 2C). The task included three levels of difficulty, with each level consisting of four questions, which resulted in a total 12 questions.

Working memory

The Find an Unselected Symbol task was designed to assess working memory based upon the Self-Ordered Pointing Task (SOPT)[70,71]. We note that Petrides et al. and Geva et al. indicate that only high memory load trials are useful in distinguishing between individuals with and without hippocampal damage[70,71]. In line with this, the DACI does not include memory loads exceeding 6. In this task, a set of six-item stimuli was presented in a 2×3 array. Whilst the layout of the stimuli remained constant across screens, their positions were randomised (Fig. 2D). The participants were asked to select a stimulus item on each screen that had not been selected in previous. The participants were not allowed to select the same location on three consecutive screens to prevent them from employing a strategy that simplified the task. The task comprised three trials, each containing six questions.

Initial human subject experiment incorporating predictive model using comprehensive version of DACI

Initial human subjects experiment using the full-length DACI

To conduct experiments using the DACI, we recruited 304 older adults (272 with normal cognitive function and 32 with cognitive impairment) from four domestic institutions (see Supplementary Sect. I for more details on the institutions). Ethical approval for this study was obtained from the Institutional Review Board (IRB) of Ewha Womans University Mokdong Hospital (EUMC 2021-08-011) and that of Yonsei University (IRB #7001988-202110-HR-1397-03). All study procedures were conducted in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki and applicable local ethical standards. Table 2 shows the demographic information of the participants.

All participants were aged 50 years or older and were able to read and write. The exclusion criteria were defined as the presence of any of the following characteristics: (1) suspected or diagnosed major neurological or psychiatric illnesses, (2) severe visual or hearing impairments that interfere with an appropriate response, or (3) any other significant medical conditions. Table 2 summarises the demographic characteristics of the participants.

Written informed consent was obtained from all participants. Under the guidance and supervision of medical staff, all participants were first asked to complete a pencil-and-paper-based Cognitive Impairment Screening Test (CIST) for preclinical validation. The CIST was developed by National Institute of Dementia in South Korea, which is recognised for its culturally and ethnically tailored adaptation of the Mini-Mental State Examination (MMSE) for the South Korean population (see Supplementary Sect. II for more details).

Based on the results of the CIST, neurologists categorised the participants as either cognitive impairment (CI) or healthy controls (HC). The cutoff score for the categorisation is as follows—participants were categorised as CI if their CIST scores fell 1.5 standard deviations below the age- and education-adjusted norms[72].The normative data used for this classification were organized by both age and education level. This ensured that participants were compared against age- and education-appropriate norms, thereby controlling for demographic factors known to influence neuropsychological performance when determining CI vs. HC classification. (see Supplementary Table 2 for the age-and education-adjusted CIST norms).

All participants subsequently completed the digitised cognitive assessment using DACI on the same day. DACI automatically recorded all behavioural and performance data, including response time, lead time, correct or incorrect status. Behavioural data refer to task interaction measures such as response time for each question and time spent on tasks, while performance data indicate response accuracy, including the total number of correct answers and question-level correctness. These data were later used to construct datasets for developing predictive models with machine learning (ML) algorithms. The behavioural and performance data were labelled according to the classifications provided by neurologists (e.g., CI or HC).

We note that the CIST consisted of 13 questions. It was expected to take approximately 10 min whereas the DACI took approximately 6 min. On average, both assessments were completed within 16 min. The duration of DACI for each participant was computed based on the system logs. It is, on average, 321 s (approximately 6 min).

Building predictive models using data collected from DACI full-length version

Using the combined logs and test results labelled with participant classifications by neurologists, we developed predictive models to screen for cognitive impairment through various ML algorithms. The primary objective was to identify the model that achieved the highest classification accuracy based on participants' data collected from DACI.

Given assessment tests in DACI, it collected a total of 84 features to classify participants as either CI or HC. In the first phase of the *Numeric Stroop task*, 18 features were extracted, including lead time, the number of correct responses, reaction times for all 10 questions, and descriptive statistics of the reaction times (e.g., sum, mean, standard deviation, minimum, maximum, and median). The same set of features was extracted during the second phase of the task. For the *Remembrance of Paired Symbols task*, 14 features were extracted, comprising lead time, the number of correct responses, and reaction time for each of the six items. Additionally, we used descriptive statistics on the reaction times for the questions. For the *quantitative comparison task*, we used 31 features, including the total number of correct responses, correctness status for each of 12 questions, and reaction time for all 12 questions. The same set of descriptive statistics on the reaction time was used, consistent with the previous tasks. Lastly, we extract three features from the *Find an Unselected Symbol task*. The lead-time for each of the three trials was computed for each feature.

We used scikit-learn (v0.24.2) to train and test the models, experimenting with ten ML algorithms commonly employed in research[73,74]. The algorithms included *CatBoost*, *Bagging*, *Logistic Regression*, *Light GBM(LGBM)*, *Naïve Bayes*, *XGBoost*, *Random Forest*, *Gradient Boosting*, *Support Vector Machine (SVM)* with two kernels: *Radial Basis Function (RBF)* and a *Linear kernel*.

For the descriptive features (inputs), we used all 84 features. The total number of samples was 304, identical to the number of participants. To train and evaluate the predictive models, we randomly split the entire dataset into training and test sets, based on a *Repeated Stratified K-fold* approach. Overall, 80% of the data (HC = 218; CI = 26) were allocated for training, and the remaining 20% (HC = 54, CI = 6) were used for the test set. All features were standardised (see Supplementary Sect. III for more details on standardization).

We set the number of folds (K) to five and the number of repetitions to four. This configuration enabled the construction of 20 classifiers for each algorithm, resulting in 200 (20 models per algorithm × 10 algorithms) simulations to determine the best model. To evaluate the classification performance, we used two distinct measures: the (i) *Area Under Curve* (AUC) score and (ii) accuracy (see Supplementary Sect. IV for more details on evaluation metrics).

Finding the minimal set of subtests to develop the compact DACI

A primary objective was to identify the minimal set of subtests required to reduce the duration of cognitive assessment whilst preserving classification accuracy. To achieve this, additional computational experiments were conducted to explore the feasibility of minimising the number of features needed to develop a compact version of DACI.

In the process, we identified a minimal feature set by examining the degree of feature importance in relation to the objective described above. We estimated the importance of all features, which resulted in the selection of the minimal set and its associated subtests. The final step involved validating the subtests that best aligned with this goal.

Estimating the degree of feature dominance

To reduce the number of features, we considered two metrics for each feature in order to measure the dominance of each one (i.e., the degree to what extent a feature influences classification performance). These metrics included the feature importance in relation to prediction accuracy, and p value indicating the statistical significance of a feature when comparing the mean values between HC and CI groups through a statistical test.

To that end, we first estimated a feature importance of all 84 features, which were used in the predictive model (for the screening of HC and CI) as described in Sect. "Finding the minimal set of subtests to develop the compact DACI". We subsequently measured the importance of each feature using the best-performing classifier (*CatBoost* in this case; see Sect. Performance on predictive models trained on the full-length DACI data for more details).

The feature importance represents the extent to which a prediction changes, on average, when value of a specific feature is changed. The higher the value of importance is, the greater the change in the prediction is, on average, when the feature is changed. We computed the feature importance using Eq. (5) in Supplementary Sect. V.

For each feature, we conducted Levene's test ($\alpha = 0.05$) to assess homogeneity of variances between the HC and CI groups; where Levene's test was non-significant ($p > 0.05$) we applied Student's independent two-sample t-test, and where it was significant ($p \leq 0.05$) we used Welch's t test. Given that both HC and CI groups exceed $n = 30$, we relied on the central limit theorem to justify approximate normality of the sampling distribution of the mean. This procedure allowed us to determine whether any significant differences existed between feature values of the HC and CI groups.

Finding the minimal set of subtests

To identify the minimum number of subtests required to maintain an optimal classification performance, we constructed various combinations using the original 84 features. Rather than employing all those features, we selected 41 features that demonstrated statistically significant differences in feature value between the HC and CI groups. We applied a baseline p value threshold of 0.01 to each feature. As a result, we successfully identified 41 features that were significantly different between the two groups.

Next, we generated candidate feature sets using a mathematical combination technique applied to the selected features. This approach involved systematically exploring all possible combinations rather than relying on random selection. Subsequently, candidate feature sets were generated using mathematical combination techniques applied to the 41 selected features. This process involved systematically exploring all possible combinations rather than relying on random selection. A cumulative approach was instead employed, with taking into account the p value of each feature into the selection process.

Initially, we sorted the features in ascending order based upon their p value. The generation of candidate feature sets began with the inclusion of the feature possessing the lowest p value. Therefore, the first candidate feature set contained only a single feature. The second candidate set was compiled by adding the feature possessing the next-lowest p value. This iterative process continued, with each subsequent candidate set including the feature directly following in the ranking.

For the illustrative example, let us assume that we have three features—f_1, f_2 and f_3 with p values of 0.001, 0.0001, and 0.01, respectively. We can then obtain a set $S = \{f_2, f_1, f_3\}$ which reflects the ascending p value. In this setting, the initial candidate feature set is $\{f_2\}$, followed by $\{f_2, f_1\}$, and finally $\{f_2, f_1, f_3\}$. Accordingly, the candidate feature sets were generated as $\{\{f_2\}, \{f_2, f_1\}, \{f_2, f_1, f_3\}\}$.

Given the candidate feature sets, we analysed them to identify the minimal feature set that could achieve the highest AUC score. We used the highest AUC score as the upper bound for classification performance. Once the optimal set with the highest AUC score was determined, we conducted further exploration to identify a feature combination that could significantly reduce the duration of cognitive assessment whilst preserving statistically equivalent classification performance (i.e., the performance of the optimal feature set vs., the upper bound).

We cast this problem as a constraint optimization task, aiming to solve the problem of minimising duration for cognitive assessment with the constraint that maintaining the upper bound, represented by the highest AUC score. The constraint optimisation problem is defined as:

$$\min \ f(x) \tag{1}$$

$$s.t. \ g(x) \approx g(U) \tag{2}$$

Let $x = (x_1, \cdots, x_n)$ represents the candidate feature sets, where the subscript of x, ranging from *1* to *n*, denotes the index of each set. The function $f(x)$ returns the duration required to complete the cognitive assessment for a given x, whilst $g(x)$ refers to the predictive model that returns the classification performance (e.g., the ROC-AUC score) based upon a set of features x. The set U represents the feature set that leads to achieving the highest classification performance, serving as the upper bound for classification accuracy. Equation (2), representing the constraint, means that statistical tests (e.g., t test, Wilcoxon rank-sum test) should indicate no significant difference between the classification performance of $g(x)$ and $g(U)$.

As outlined above, the final outcome of this constrained optimisation is the identification of the optimal feature set that maximises classification performance whilst minimising assessment duration. For example, let us consider $U = 21$; that is the predictive model with a candidate feature set containing 21 features. We denoted it to $g(21)$, which can achieve the highest classification performance, either CI or CH, we assumed in this context.

As an example, let us suppose $x = 13$, which represents one candidate feature set with 13 features. If the predictive model $g(13)$ can achieve statistically the same highest AUC score as that of $g(21)$, it is obvious

that the task involving 13 features will naturally require less time than the task involving 21 features, whilst maintaining statistically equivalent classification accuracy.

Algorithm 1 illustrates the constraint optimisation process, with the final output being the optimal feature set that meets these criteria.

Require: Playing time array T, AUC score array A, Upper limit U
Ensure: Optimal solution index S

Initialize $maxAUC \leftarrow -\infty$
Initialize $S \leftarrow -1$
for $i = 0$ to $U - 2$ **do**
  if $Shapiro_Wilk_test(A[i]) > 0.05$ **then**
    $P_{val} \leftarrow Paired_t_test(A[i], A[U-1])$
  else
    $P_{val} \leftarrow Wilcoxon_Rank_Sum_test(A[i], A[U-1])$
  end if
  if $T[i] < T[U-1]$ **and** $P_{val} > 0.05$ **then**
    if $A[i] > maxAUC$ **then**
      $maxAUC \leftarrow A[i]$
      $S \leftarrow i$
    end if
  end if
end for
return S

Algorithm 1. Find optimal solution

To test and validate the ML model, we applied a Repeated Stratified K-fold method. We set the folds (K) to five and repeated them four times to split the dataset into the training and test sets. To ensure fairness in the data splitting process, we fixed the random seeds across all operations. This allowed us to assess variations in the AUC score as the number of features changed within the same dataset.

Follow-up human subjects experiment incorporating a lightweight predictive model using the compact version of DACI

Follow-up human subjects experiment using the compact DACI

Following the identification of the minimal test set, we conducted the follow-up human subjects experiment using the compact version of DACI. Additional 297 older adults (227 with normal cognitive function and 70 with cognitive impairment) from the domestic institutions (See Supplementary Sect. I for more details) were recruited for this experiment. Here, we applied the same inclusion and exclusion criteria introduced in Sect. "Initial human subject experiment incorporating predictive model using comprehensive version of DACI". The experiment was performed under the same IRB approval. Table 3 shows the demographic information of the participants.

The experiment procedure is absolutely the same as that in Sect. Initial human subject experiment incorporating predictive model using comprehensive version of DACI. Obtaining the written consent form from all participants, we instructed them to complete a pencil-and-paper based CIST. Again, neurologists classified

	Age		Education level (years)		Gender	CIST score	
	Average	Standard deviation	Average	Standard deviation	All (Male)	Average	Standard deviation
Cognitive impairment	76.91	4.76	5.97	4.01	70 (16)	18.8	4.58
Healthy control	69.56	5.93	11.44	4.27	227 (66)	27.17	2.77

Table 3. Demographic characteristics of the follow-up study participants.

Model	AUC	Sensitivity	Specificity	Accuracy
CatBoost	0.813	0.903	0.733	0.751

Table 4. Performance of the predictive model trained by catboost on the classification of HC and CI.

Task	Feature	Statistical significance (p value)	Effect size	Feature importance (SHAP value)
Numeric stroop	Response Time_part A_question 5	2.20×10^{-02}	1.14	0.59
	Response Time_part A mean	3.21×10^{-04}	1.03	0.16
	Response Time_part A sum	3.21×10^{-04}	1.03	0.16
Remembrance paired symbols	Response Time_question 1	4.95×10^{-04}	1.0	0.61
	Lead Time_total	4.28×10^{-03}	1.0	0.17
	Response Time_question sum	4.28×10^{-03}	1.0	0.1
Quantitative comparison	Response Time_question 2	5.24×10^{-02}	0.86	1.00
	Response Time_question min	1.33×10^{-02}	0.74	0.11
	Response Time_question 4	6.52×10^{-02}	0.73	0.29
Find an unselected symbol	Lead Time_set 2	1.17×10^{-03}	0.61	0.15
	Lead Time_set 1	6.57×10^{-03}	0.51	0.23
	Lead Time_set 3	8.27×10^{-03}	0.50	0.16

Table 5. Summary of the feature dominance estimation.

the participants as either CI or HC on the basis of CIST results. Of course the cutoff score is the same as that in Sect. Initial human subject experiment incorporating predictive model using comprehensive version of DACI. The participants at this time performed the compact version of DACI afterwards instead of the full-length, comprehensive version on the same day.

Building predictive models using data collected from DACI compact version

As expected, we used the same methodology to build the predictive model in this experiment. After collecting the user logs and test results, coupled with participant classifications provided by neurologists, we built the predictive models for the screen of CI and HC. The identical set of ML algorithms, the same version of scikit-learn, and the same training and evaluation approach described in Sect. Initial human subject experiment incorporating predictive model using comprehensive version of DACI were used.

The total number of samples was 297 which was identical to the number of participants. The dataset was split into training and test sets in an 80:20 ratio, the same as that of the initial experiment. The number of folds (K) was again set to five, with four repetitions, resulting in the construction of 20 classifiers per algorithm, enabling 200 simulations (20 models per algorithm × 10 algorithms) to seek the best-performing model. We used two distinct measures to evaluate the performance: the (i) *Area Under Curve* (AUC) score and (ii) accuracy (see Supplementary Sect. IV for more details on evaluation metrics).

A key distinction in this experiment was that participants used the DACI compact version. As a result, we anticipated that the number of features collected from the DACI compact version would be lower than that of the DACI full-length version.

Results

Performance on predictive models trained on the full-length DACI data

Table 4 shows the AUC, sensitivity, specificity, and accuracy of the best-performing predictive model on average. In this case, the model trained using the CatBoost ML algorithm exhibited the highest performance among the ten candidate models (see Supplementary Fig. S1 for the AUC distributions across all models; Supplementary Fig. S2 for accuracy distributions; Supplementary Fig. S3 for confusion matrices; Supplementary Fig. S4 for ROC curves). We set the AUC 0.813 as an upper-bound used for finding a minimal set of subtests to develop the compact version of DACI (The full set of results for the ten candidate models are shown in Supplementary Table 3).

Constraint optimisation process successfully construct the compact DACI

Degree of feature dominance

Table 5 presents the results of the degree of feature dominance derived from the *CatBoost* model. The top three features are listed in descending order of dominance, with the complete all results available in Supplementary Table 4.

To ensure these results, we conducted a correlation test between feature importance and the p value across all features. The test results confirmed a significant correlation ($p < 0.05$, $R^2 = 0.1144$), suggesting that the lower the p value is, the higher the feature importance is. We note that p values were used as the primary criterion for selecting the minimal feature set as feature importance often varies depending on the ML algorithm.

The following provides definitions for each feature listed in Table 5:

- Response Time_Part A_Question 5—Refers to the reaction time for Question 5 of the *Numeric Stroop Task (Part A)*.
- Response Time_Part A Mean—Represents the average reaction time across all six questions in *Part A of the Numeric Stroop Task*.
- Response Time_Part A Sum—Indicates the total response time for the six questions in *Part A of the Numeric Stroop Task*.
- Lead Time_Total—Denotes the overall time required to complete a *Remembrance of Paired Symbols task*.
- Response Time_Question Sum—Represents the cumulative reaction time to complete all questions in the *Remembrance of Paired Symbols task*.
- Response Time_Question Min—Refers to the shortest reaction time recorded among all questions in the *Quantitative Comparison task*.
- Lead Time_Question 2—Reflects the total time required to complete one set of the *Find an Unselected Symbol task*.

In addition to the t-test–based feature selection and CatBoost-based feature importance, we further employed SHAP (SHapley Additive exPlanations) analysis to visualize how each feature contributes to the classification outcome. The SHAP summary plots are presented in Supplementary Fig. S5, highlighting the relative impact of key features.

Construction of the compact version of DACI

Although the best-performing model in the initial experiment included all 84 features, we actually used only 41 out of 84 as described in the Materials and Methods section. This approach resulted in an AUC of 0.813, raising the optimistic expectation that curating features based on inter-group differences might lead to higher AUC scores.

We thus combined various subsets of 41 features to identify the minimal feature set that could achieve the highest AUC score. The classifier trained with 21 features demonstrated the maximum AUC score of 0.8329. These were derived from three subtests: *Numeric Stroop*, *Remembrance paired symbols*, and *Quantitative Comparison*. It resulted in an average assessment time of 224 s. We set this as the upper bound for the classification performance.

Using Algorithm 1, we identified 13 features that constituted the optimal set which could minimise the cognitive assessment time whilst retaining performance being statistically the same as the upper bound. We subsequently conducted additional tests on assessment duration and AUC scores to ensure that the optimal feature set was selected appropriately. To identify features with no significant effect on AUC scores, we performed a paired sample *t* test between the upper bound and groups with fewer than 21 features (Fig. 3A). The results emphasised that 12 groups were not statistically different from the group utilising 21 features.

We further conducted a statistical test to identify significant differences in assessment duration across groups (Fig. 3B). When using 2 to 15 features, only two subtests—*Numeric Stroop* and *Remembrance paired symbols*—were selected. In this case, the assessment duration required was 91 s on average. When a single feature was used, the *Remembrance of Paired Symbols* subtest was selected exclusively, with an average assessment duration of 47 s.

In summary, seven feature groups—2, 4, 5, 6, 13, 14, and 15—showed no statistically significant difference in AUC score, but exhibited a statistically significant difference in cognitive assessment duration. These seven groups were designated as the best-performing groups (Fig. 3C). We then selected the 13-feature group, achieving the highest AUC score of 0.8232, was ultimately selected as the optimal feature set. Through this process, we validated the optimal solution determined by the proposed algorithms. These findings suggest that cognitive impairment (CI) screening can be conducted in significantly less time, as only two tasks are required within the compact DACI.

Performance on predictive models trained with the compact DACI data

To further validate the compact version of DACI, comprising only two subtests as derived in Sect. 3.2, an additional experiment was conducted with 297 new participants (227 HC and 70 CI). Following the same protocol as the earlier experiment, participants first completed the CIST, followed by the DACI. Using these newly collected data, we performed the same approach (Repeated Stratified K-fold) to evaluate as in the initial experiment.

Table 6 presents the performance of the CatBoost-based model, which once again demonstrated the highest prediction capacity. The results confirm that the compact version of DACI sustained statistically the same prediction performance as that of the full-length DACI. These findings suggest that the simplified DACI can serve as a reliable and user-friendly screening tool that retains essential sensitivity for detecting CI whilst significantly reducing assessment duration.

Discussion

The development of DACI was based on a comprehensive review of existing digitized cognitive assessment tools for detecting Alzheimer's disease symptoms. This review guided the selection of task components that were both theoretically sound and practically appropriate for rapid digital assessment. In designing DACI, we placed particular emphasis on incorporating tasks that are sensitive to early cognitive changes. These changes occur specifically in executive functioning[51–54,67,68], associative memory[58,69], calculation[52,65], and working memory[59–64,70,71].

After developing the tool, we evaluated its clinical validity by comparing DACI-derived performance outcomes with neurologist classifications using a standardized cognitive screening instrument (CIST). In this study, we found that data collected from the carefully designed DACI effectively distinguished older adults with CI from HC. Among the predictive models trained on all 84 features from 304 participants, the CatBoost

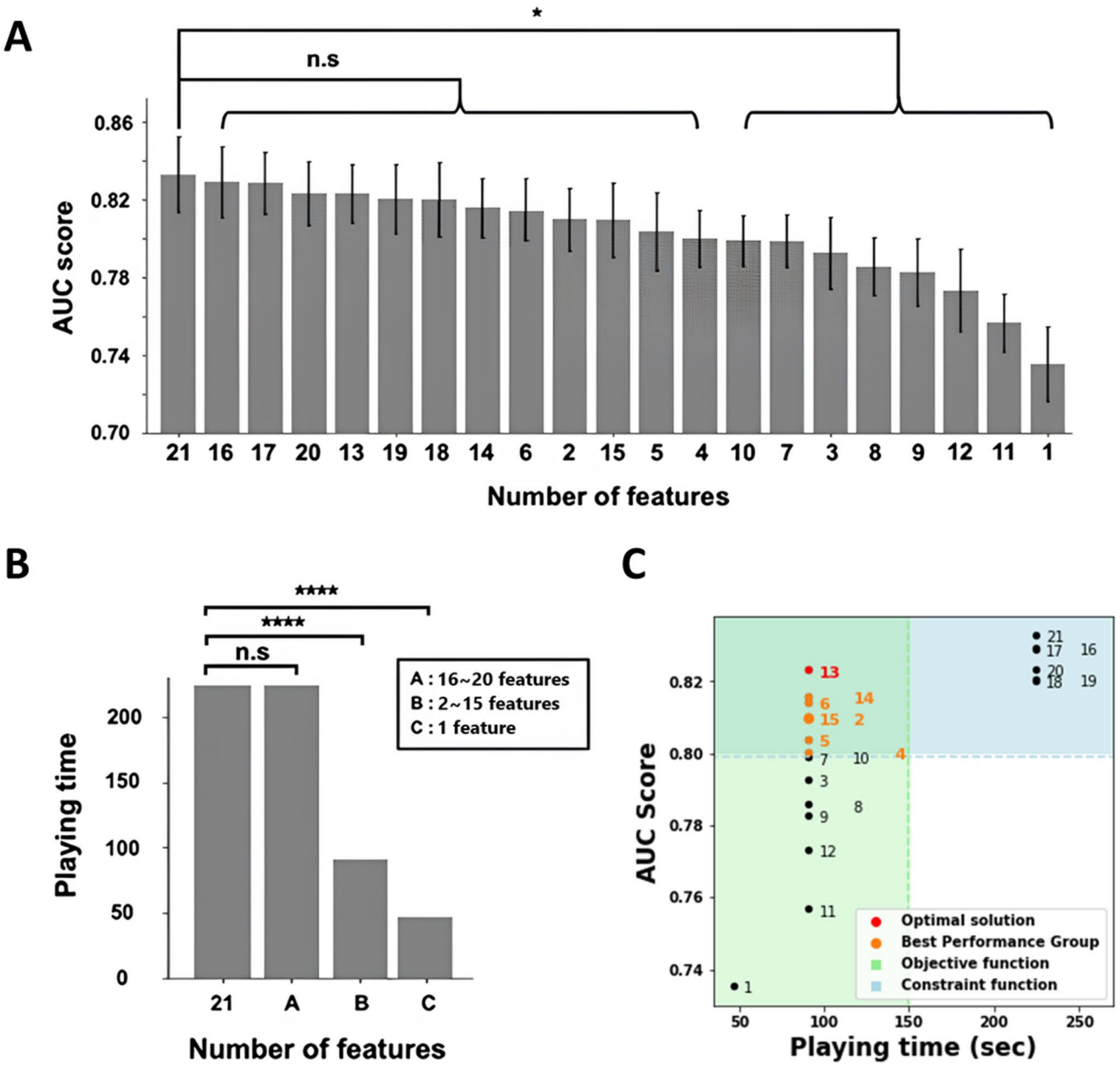


Fig. 3. Results of the Experiment on Designing a Minimal Subtest Set. (A) The plot displays AUC scores of predictive models trained with varying numbers of features. The AUC score of the best-performing group was significantly higher than that of the others. The X-axis represents the number of features, while the Y-axis indicates the AUC score. Error bars denote the standard error of the mean. (B) The plot illustrates task duration in relation to the number of features. The X-axis represents the number of features, and the Y-axis indicates task duration. Group A includes 16 to 20 features, Group B comprises 2 to 15 features, and Group C consists of a single feature. (C) The plot shows AUC scores in relation to task duration. The number next to each circle dot represents the number of features used. For overlapping values, the feature set with the higher AUC score is displayed to the right. The X-axis represents task duration, while the Y-axis shows the AUC score.

Model	AUC	Sensitivity	Specificity	Accuracy
CatBoost	0.871	0.864	0.783	0.802

Table 6. Prediction performance with a compact set of DACI on screening HC and CI ($N = 297$) using the catboost model.

algorithm achieved the highest AUC of 0.813, with a sensitivity of 0.903 and a specificity of 0.733 (Table 4). This result demonstrated that DACI effectively distinguished between the cognitively impaired and healthy individuals, supporting its diagnostic relevance[75–80].

We selected CatBoost as the primary predictive model for DACI based on its superior sensitivity—a critical factor in screening tests[81,82]. Previous studies have reported that remote digital cognitive assessments achieved an accuracy of 0.84, sensitivity of 0.86, and specificity of 0.83 for classifying CI from HC participants[83]. In comparison, DACI's performance using neuropsychological tests alone demonstrated moderate superiority in screening sensitivity for individuals at risk of cognitive impairment compared to methods proposed in previous studies[75,76], while achieving competitive accuracy with early screening approaches[77] that incorporate diverse modalities[78–80]. Our findings suggest that DACI represents an effective screening tool for cognitive impairment in older adults, with demonstrated high sensitivity. However, DACI's specificity was lower than that reported in previous studies, indicating that further research should be required to enhance this performance metric.

In addition, we developed a compact version of DACI to reduce completion time without compromising screening accuracy. The compact version exhibited strong potential for achieving superior classification performance (AUC 0.871) compared to previous studies while simultaneously reducing assessment duration. We employed optimization techniques to identify a minimal set of subtests that significantly reduced completion time. Through successive feature refinement, 13 significant features were identified from two subtests. We then conducted an additional round of human subject experiments using DACI-mini with 297 new participants. After collecting user behavioral data, we successfully validated our approach by developing an early screening model that reduced the average task duration to 91 s while achieving optimal AUC performance (0.871).

Given that this cognitive screening application targets older adults, shorter assessment duration enhances sustained attention during cognitive tasks while reducing fatigue, both of which are essential for accurately measuring cognitive function. Minimizing completion time is crucial when developing digital assessment tools for older populations due to the potential for fatigue resulting from prolonged digital device use[48,49]. In this study, we validated a digital assessment tool that maintains measurement accuracy while significantly reducing completion time. This enhancement improves the tool's practical utility in clinical and community settings, where time-efficient screening is essential[84–86].

However, a potential limitation is that the results from the compact DACI might be slightly different from those obtained with the full-length DACI. Behavioural and performance data during the assessment could be influenced by the order and combination of tasks. For instance, previously completed tasks might be likely to affect performance on subsequent tasks, resulting in a "carryover effect." Since the task order and composition in the full-length DACI was different from those in the compact DACI, this could bring about some variations in user experience and cognitive processes. Consequently, discrepancies could arise in the behavioural and performance data collected across different versions, raising concerns about feature set inconsistencies.

Mobile-based cognitive assessments offer several advantages over traditional screening methods and other digital devices, including reduced experimenter influence, greater accessibility, minimal need for invasive testing, faster administration, and improved cost- and time-efficiency[5]. Previous studies have in addition demonstrated that compact, mobile app-based remote cognitive assessments exhibit high validity and reliability in detecting cognitive impairment in older adults[29].

Given that usability is a significant limitation of digital cognitive assessments in older adults, particularly those with cognitive impairments, the application of remote cognitive assessments in this population is essential. Many digital screening tools require additional equipment[87], which can be challenging to implement in older adults without professional supervision. DACI, however, can be deployed using only participants' mobile devices that eliminates the need for supplementary equipment or external support. This system has the potential to provide not only cost-effective but also convenient means which can significantly contribute to cognitive assessment accessibility in older adults.

Limitations of the study

This study had several limitations. First, the relatively small sample size of participants with cognitive impairments is likely to have constrained the results. Larger studies involving more patients with cognitive impairment could enhance the resolution, precision and generalisability of DACI's screening performance.

In addition, the study was conducted in a semi-supervised setting. In other words, participants had assistance during the installation of the DACI mobile application. At the same time, they were provided with a concise guideline at the point of consent. Supervisors repeated instructions when necessary to address misunderstandings. Although DACI was designed for independent use by older adults in home environments, the study was conducted in an institutional setting to ensure a controlled environment. Despite minimising supervisors' intervention, it would be evident that assistance was often required. This highlights the need for further preparation in future research to assess DACI's feasibility in unsupervised environments. It is also necessary to validate the efficacy of the compact DACI in real-world settings.

As an additional usability assessment independent of the main study, we performed System Usability Scale (SUS) evaluations with 75 older adult participants. The average score was 57.53 (SD = 15.06), which falls below the commonly accepted threshold of 68 for acceptable usability[88,89]. This indicates that DACI's usability remains somewhat limited, particularly for older users. The relatively low score likely reflects age-related factors such as reduced digital literacy and unfamiliarity with similar technologies. Based on these findings, future research should prioritize improving the user experience to enhance overall usability.

Cognitive impairment was assessed using the CIST. Whilst CIST demonstrated an AUC of 0.83 for distinguishing normal participants from those with MCI[72], further research incorporating detailed neuropsychological assessments is warranted to classify cognitive impairment among healthy older adults using DACI.

Despite these limitations, the findings indicate that DACI is a valuable remote cognitive assessment tool, demonstrating high sensitivity in screening for cognitive impairment in older adults.

Data availability

The data and code that support the findings of this study are available from the corresponding author upon reasonable request. Please contact jeehang@smu.ac.kr for access to the research data.

Received: 14 March 2025; Accepted: 29 July 2025
Published online: 19 August 2025

References

1. World Health Organization. *Global Status Report on the Public Health Response To Dementia* (World Health Organization, 2021). https://apps.who.int/iris/handle/10665/344701
2. Prince, M., Bryce, R. & Ferri, C. *Alzheimer's Disease International World Alzheimer Report. The Benefits of Early Diagnosis and Intervention* (Institute of Psychiatry, King's College London, London, UK, 2011). https://www.alzint.org/u/WorldAlzheimerReport2011.pdf
3. Sabbagh, M. N. et al. Early detection of mild cognitive impairment (MCI) in an at-home setting. *J. Prev. Alzheimers Dis.* **7** (3), 171–178. https://doi.org/10.14283/jpad.2020.22 (2020).
4. Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N. & Rubino, I. Diagnosis of early alzheimer's disease: Clinical practice in 2021. *J. Prev. Alzheimers Dis.* **8** (4), 371–386. https://doi.org/10.14283/jpad.2021.23 (2021).
5. Ding, Z., Lee, T. L. & Chan, A. S. Digital cognitive biomarker for mild cognitive impairments and dementia: A systematic review. *J. Clin. Med.* **11** (14), 4191. https://doi.org/10.3390/jcm11144191 (2022).
6. Perez-Valero, E. et al. An automated approach for the detection of alzheimer's disease from resting-state electroencephalography. *Front. Neuroinform.* **16**, 924547. https://doi.org/10.3389/fninf.2022.924547 (2022).
7. Van Dyck, C. H. et al. Lecanemab in early Alzheimer's disease. *N Engl. J. Med.* **388** (1), 9–21. https://doi.org/10.1056/NEJMoa2212948 (2023).
8. Livingston, G. et al. Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. *Lancet.* **404** (10452), 572–628. https://doi.org/10.1016/S0140-6736(24)01296-0 (2024).
9. Deena, G., Karunakaran, N. B., Sharma, S., Pandey, V. & Nishant, N. Cognitive assessment and early detection of alzheimer's disease. In *Advances in Medical Technologies and Clinical Practice* (IGI Global, Hershey, 2024). https://doi.org/10.4018/979-8-3693-3605-2.ch025
10. Rehman, S. U. et al. AI-based tool for early detection of Alzheimer's disease. *Heliyon* **10** (5), e29375. https://doi.org/10.1016/j.heliyon.2024.e29375 (2024).
11. Lee, J. H. et al. A framework for best practices and readiness in the advent of anti-amyloid therapy for early Alzheimer's disease in Asia. *J. Alzheimers Dis.* **89** (2). https://doi.org/10.3233/JAD-240684 (2024).
12. McDade, E. et al. Lecanemab in patients with early Alzheimer's disease: Detailed results on biomarker, cognitive, and clinical effects from the randomized and open-label extension of the phase 2 proof-of-concept study. *Alzheimers Res. Ther.* **14**, 191. https://doi.org/10.1186/s13195-022-01124-2 (2022).
13. Liss, J. L. et al. Practical recommendations for timely, accurate diagnosis of symptomatic alzheimer's disease (MCI and dementia) in primary care: A review and synthesis. *J. Intern. Med.* **290** (2), 310–334. https://doi.org/10.1111/joim.13244 (2021).
14. Hong, Y. J. et al. Development of a home-based cognitive test for cognitive monitoring in subjective cognitive decline with high risk of Alzheimer's disease. *Med. (Baltim).* **102** (9), e33096. https://doi.org/10.1097/MD.0000000000033096 (2023).
15. Teipel, S. et al. Multimodal imaging in alzheimer's disease: Validity and usefulness for early detection. *Lancet Neurol.* **14** (10), 1037–1053. https://doi.org/10.1016/S1474-4422(15)00093-9 (2015).
16. Reckess, G. Z. et al. Screenifng by telephone in the Alzheimer's disease anti-inflammatory prevention trial. *J. Alzheimers Dis.* **36** (3), 433–443. https://doi.org/10.3233/JAD-130113 (2013).
17. De Marchi, F. et al. Telehealth in neurodegenerative diseases: Opportunities and challenges for patients and physicians. *Brain Sci.* **11** (2), 237. https://doi.org/10.3390/brainsci11020237 (2021).
18. Geddes, M. R. et al. Remote cognitive and behavioral assessment: Report of the alzheimer society of Canada task force on dementia care best practices for COVID-19. *Alzheimers Dement. (Amst).* **12** (1), e12111. https://doi.org/10.1002/dad2.12111 (2020).
19. Jeong, S. et al. A case report of Korean medicine treatment for a post-traumatic hydrocephalus patient complaining of gait disturbance and dementia. *J. Intern. Korean Med.* **44** (1), 77–85. https://doi.org/10.22246/jikm.2023.44.1.77 (2023).
20. Seo, E. H. et al. Validity of the telephone interview for cognitive status (TICS) and modified TICS (TICSm) for mild cognitive impairment (MCI) and dementia screening. *Arch. Gerontol. Geriatr.* **52** (1), e26–e30. https://doi.org/10.1016/j.archger.2010.04.008 (2011).
21. Velayudhan, L. et al. Review of brief cognitive tests for patients with suspected dementia. *Int. Psychogeriatr.* **26** (8), 1247–1262. https://doi.org/10.1017/S1041610214000416 (2014).
22. Bottiroli, S. et al. Smart aging platform for evaluating cognitive functions in aging: A comparison with the MoCA in a normal population. *Front. Aging Neurosci.* **9**, 379. https://doi.org/10.3389/fnagi.2017.00379 (2017).
23. Kourtis, L. C., Regele, O. B., Wright, J. M. & Jones, G. B. Digital biomarkers for alzheimer's disease: The mobile/wearable devices opportunity. *NPJ Digit. Med.* **2**, 9. https://doi.org/10.1038/s41746-019-0084-2 (2019).
24. Frank, C., St John, P. & Molnar, F. Screening tools for virtual assessment of cognition. *Can. Fam Physician.* **66** (7), 502–503 (2020).
25. Watt, J. A. et al. Diagnostic accuracy of virtual cognitive assessment and testing: Systematic review and meta-analysis. *J. Am. Geriatr. Soc.* **69** (6), 1429–1440. https://doi.org/10.1111/jgs.17190 (2021).
26. Brown, E. E., Kumar, S., Rajji, T. K., Pollock, B. G. & Mulsant, B. H. Anticipating and mitigating the impact of the COVID-19 pandemic on Alzheimer's disease and related dementias. *Am. J. Geriatr. Psychiatry.* **28** (7), 712–721. https://doi.org/10.1016/j.jagp.2020.04.010 (2020).
27. Hampel, H. et al. Designing the next-generation clinical care pathway for alzheimer's disease. *Nat. Aging.* **2** (8), 692–703. https://doi.org/10.1038/s43587-022-00269-x (2022).
28. Chen, R. et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2145–2155 (2019). https://doi.org/10.1145/3292500.3330690
29. Thabtah, F., Peebles, D., Retzler, J. & Hathurusingha, C. Dementia medical screening using mobile applications: A systematic review with a new mapping model. *J. Biomed. Inf.* **111**, 103573. https://doi.org/10.1016/j.jbi.2020.103573 (2020).
30. Genç, Y. E. Evaluating the use of digital biomarkers to test treatment effects on cognition and movement in patients with lewy body dementia. *J. Parkinsons Dis.* **12** https://doi.org/10.3233/jpd-213126 (2022).
31. Tsoi, K. K. et al. Applications of artificial intelligence in dementia research. *Camb. Prism. Prec Med.* **1**, e9. https://doi.org/10.1017/pcm.2022.10 (2023).

32. Kurland, D. B. et al. Digital biomarkers and the evolution of spine care outcomes measures: Smartphones and wearables. *Neurosurgery*. **92** (2), 2519. https://doi.org/10.1227/neu.0000000000002519 (2023).
33. Kim, W. et al. A preliminary feasibility study on screening cognitive impairment based on multi-modal biomarkers and stacking ensemble AI approach. *Alzheimers Dement.* **19** (S5), e082466. https://doi.org/10.1002/alz.082466 (2023).
34. Chudzik, A., Śledzianowski, A. & Przybyszewski, A. W. Machine learning and digital biomarkers can detect early stages of neurodegenerative diseases. *Sens. (Basel)*. **24** (5), 1572. https://doi.org/10.3390/s24051572 (2024).
35. Hoang, T. H., Zallek, C. M. & Minh, N. Smartphone-based digitized neurological examination toolbox for multi-test neurological abnormality detection and documentation. *IEEE J. Biomed. Health Inf.* https://doi.org/10.1109/jbhi.2024.3439492 (2024).
36. Zhang, Y., Folarin, A. A. & Dobson, R. Exploring digital biomarkers for depression using mobile technology. *Int. J. Popul. Data Sci.* **9** (4), 2422. https://doi.org/10.23889/ijpds.v9i4.2422 (2024).
37. Yun, B. H. et al. Development and effectiveness of an AI chatbot-based mobile cognitive screening and customized training application for preventing dementia: Older adults living in rural areas of South Korea. *Arch. Des. Res.* **37** (5), 77–90. https://doi.org/10.15187/adr.2024.11.37.5.77 (2024).
38. Roshanzamir, A., Aghajan, H. & Soleymani Baghshah, M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Med. Inf. Decis. Mak.* **21**, 92. https://doi.org/10.1186/s12911-021-01456-3 (2021).
39. Öhman, F., Hassenstab, J., Berron, D., Schöll, M. & Papp, K. V. Current advances in digital cognitive assessment for preclinical Alzheimer's disease. *Alzheimers Dement. (Amst).* **13** (1), e12217. https://doi.org/10.1002/dad2.12217 (2021).
40. Ferrar, J. et al. Developing digital tools for remote clinical research: How to evaluate the validity and practicality of active assessments in field settings. *J. Med. Internet Res.* **23** (6), e26004. https://doi.org/10.2196/26004 (2021).
41. Wang, Q. et al. Usability evaluation of mHealth apps for elderly individuals: A scoping review. *BMC Med. Inf. Decis. Mak.* **22**, 64. https://doi.org/10.1186/s12911-022-02064-5 (2022).
42. Arya, S. S., Dias, S. B., Jelinek, H. F., Hadjileontiadis, L. J. & Pappa, A. M. The convergence of traditional and digital biomarkers through AI-assisted biosensing: a new era in translational diagnostics? *Biosens. Bioelectron.* **225**, 115387. https://doi.org/10.1016/j.bios.2023.115387 (2023).
43. Javed, A. R. et al. Artificial intelligence for cognitive health assessment: State-of-the-art, open challenges and future directions. *Cogn. Comput.* **15** (6), 1767–1812. https://doi.org/10.1007/s12559-023-10153-4 (2023).
44. Chen, L., Zhen, W. & Peng, D. Research on digital tool in cognitive assessment: A bibliometric analysis. *Front. Psychiatry*. **14**, 1227261. https://doi.org/10.3389/fpsyt.2023.1227261 (2023).
45. Zierer, C., Behrendt, C. & Lepach-Engelhardt, A. C. Digital biomarkers in depression: A systematic review and call for standardization and harmonization of feature engineering. *J. Affect. Disord*. **355**, 233–246. https://doi.org/10.1016/j.jad.2024.03.163 (2024).
46. Castilla, D., Suso-Ribera, C., Zaragoza, I., Garcia-Palacios, A. & Botella, C. Designing ICTs for users with mild cognitive impairment: A usability study. *Int. J. Environ. Res. Public. Health*. **17** (14), 5153. https://doi.org/10.3390/ijerph17145153 (2020).
47. Wilson, S. A., Byrne, P., Rodgers, S. E. & Maden, M. A systematic review of smartphone and tablet use by older adults with and without cognitive impairment. *Innov. Aging*. **6** (2), igac002. https://doi.org/10.1093/geroni/igac002 (2022).
48. Sonderegger, A., Schmutz, S. & Sauer, J. The influence of age in usability testing. *Appl. Ergon.* **52**, 291–300. https://doi.org/10.1016/j.apergo.2015.06.012 (2016).
49. Wagner, N., Hassanein, K. & Head, M. The impact of age on website usability. *Comput. Hum. Behav.* **37**, 270–282. https://doi.org/10.1016/j.chb.2014.05.003 (2014).
50. Liu, M. A., Tseng, W. Y. I. & Chen, S. H. A. Comparing the performance of older and younger adults in a virtual reality-based virtual Morris water maze task. *Front. Psychol.* **10**, 91. https://doi.org/10.3389/fpsyg.2019.00091 (2019).
51. Jack, C. R. Jr et al. NIA-AA research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's Dement.* **14** (4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018 (2018).
52. Kasai, W., Goto, T., Aoyama, Y. & Sato, K. Association between cognitive quotient test score and hippocampal volume: A novel, rapid application-based screening tool. *Sci. Rep.* **10** (1), 16728. https://doi.org/10.1016/j.jalz.2018.02.018 (2020).
53. Bélanger, S., Belleville, S. & Gauthier, S. Inhibition impairments in alzheimer's disease, mild cognitive impairment and healthy aging: Effect of congruency proportion in a stroop task. *Neuropsychologia*. **48** (2), 581–590. https://doi.org/10.1016/j.neuropsychologia.2009.10.021 (2010).
54. Kramer, J. H. et al. Multiple cognitive deficits in amnestic mild cognitive impairment. *Dement. Geriatr. Cogn. Disord.* **22** (4), 306–311. https://doi.org/10.1159/000095303 (2006).
55. Fisk, J. D., Merry, H. R. & Rockwood, K. Variations in case definition affect prevalence but not outcomes of mild cognitive impairment. *Neurology* **61** (9), 1179–1184. https://doi.org/10.1212/01.WNL.0000089238.07771.C (2003).
56. Fisk, J. D. & Rockwood, K. Outcomes of incident mild cognitive impairment in relation to case definition. *J. Neurol. Neurosurg. Psychiatry*. **76** (8), 1175–1177. https://doi.org/10.1136/jnnp.2004.053751 (2005).
57. Petersen, R. C. Mild cognitive impairment as a diagnostic entity. *J. Intern. Med.* **256** (3), 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x (2004).
58. Troyer, A. K. et al. Item and associative memory in amnestic mild cognitive impairment: Performance on standardized memory tests. *Neuropsychology* **22** (1), 10. https://doi.org/10.1037/0894-4105.22.1.10 (2008).
59. Amieva, H. et al. Prodromal alzheimer's disease: successive emergence of the clinical symptoms. *Ann. Neurol. Offi. J. Am. Neurol. Assoc. Child. Neurol. Soc.* **64** (5), 492–498. https://doi.org/10.1002/ana.21509 (2008).
60. Wilson, R. S., Leurgans, S. E., Boyle, P. A. & Bennett, D. A. Cognitive decline in prodromal alzheimer disease and mild cognitive impairment. *Arch. Neurol.* **68** (3), 351–356. https://doi.org/10.1001/archneurol.2011.31 (2011).
61. Chaytor, N. & Schmitter-Edgecombe, M. Working memory and aging: A cross-sectional and longitudinal analysis using a self-ordered pointing task. *J. Int. Neuropsychol. Soc.* **10** (4), 489–503. https://doi.org/10.1017/S1355617704104013 (2004).
62. Daigneault, S. & Braun, C. M. Working memory and the self-ordered pointing task: Further evidence of early prefrontal decline in normal aging. *J. Clin. Exp. Neuropsychol.* **15** (6), 881–895. https://doi.org/10.1080/01688639308402605 (1993).
63. Shimamura, A. P. & Jurica, P. J. Memory interference effects and aging: Findings from a test of frontal lobe function. *Neuropsychology* **8** (3), 408 (1994).
64. West, R., Ergis, A. M., Winocur, G. & Saint-Cyr, J. The contribution of impaired working memory monitoring to performance of the self-ordered pointing task in normal aging and parkinson's disease. *Neuropsychology* **12** (4), 546. https://doi.org/10.1037/0894-4105.12.4.546 (1998).
65. Martin, R. C. et al. Loss of calculation abilities in patients with mild and moderate alzheimer disease. *Arch. Neurol.* **60** (11), 1585–1589. https://doi.org/10.1001/archneur.60.11.1585 (2003).
66. Dodd, C., Athauda, R. & Adam, M. Designing user interfaces for the elderly: A systematic literature review. https://aisel.aisnet.org/acis2017/61. (2017).
67. Scarpina, F. & Tagini, S. The stroop color and word test. *Front. Psychol.* **8**, 557. https://doi.org/10.3389/fpsyg.2017.00557 (2017).
68. Wolach, A. H., McHale, M. A. & Tarlea, A. Numerical stroop effect. *Percept. Motor Skills*. **98** (1), 67–77. https://doi.org/10.2466/pms.98.1.67-77 (2004).
69. Wechsler, D. *Wechsler Adult Intelligence Scale–Third Edition (WAIS–III)* (The Psychological Corporation, San Antonio, 1997).
70. Petrides, M. & Milner, B. Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*. **20** (3), 249–262. https://doi.org/10.1016/0028-3932(82)90100-2 (1982).

71. Geva, S., Cooper, J. M., Gadian, D. G., Mishkin, M. & Vargha-Khadem, F. Impairment on a self-ordered working memory task in patients with early-acquired hippocampal atrophy. *Dev. Cogn. Neurosci.* **20**, 12–22. https://doi.org/10.1016/j.dcn.2016.06.001 (2016).
72. The Ministry of Health and Welfare in South Korea. Study to develop Korean dementia screening test. The Ministry of Health and Welfare in South Korea. http://uci.or.kr/G500:1310377-02012057888653 (2020).
73. Katsaouni, N., Tashkandi, A., Wiese, L. & Schulz, M. H. Machine learning-based disease prediction from genotype data. *Biol. Chem.* **402** (8), 871–885. https://doi.org/10.1515/hsz-2021-0109 (2021).
74. Lenka, S., Bisoy, S., Priyadarshini, R. & Sain, M. Empirical analysis of ensemble learning for imbalanced credit scoring datasets: A systematic review. *Wirel. Commun. Mob. Comput.* **2022**, 6584352. https://doi.org/10.1155/2022/6584352 (2022).
75. Berron, D. et al. A remote digital memory composite to detect cognitive impairment in memory clinic samples in unsupervised settings using mobile devices. *NPJ Digit. Med.* **7** (1), 79. https://doi.org/10.1038/s41746-024-00999-9 (2024).
76. Young, S. R. et al. An initial exploratory examination of the sensitivity and specificity of MyCog mobile using the Mini-Cog as a proxy for suspected cognitive impairment. *J. Prim. Care Community Health.* **15**, 21501319241297771. https://doi.org/10.1177/21501319241297771 (2024).
77. Oyama, A. et al. Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. *Sci. Rep.* **9** (1), 12932. https://doi.org/10.1038/s41598-019-49275-x (2019).
78. Yamada, Y. et al. Combining multimodal behavioral data of gait, speech, and drawing for classification of Alzheimer's disease and mild cognitive impairment. *J. Alzheimers Dis.* **84** (1), 315–327. https://doi.org/10.3233/JAD-210684 (2021).
79. Jonell, P. et al. Multimodal capture of patient behaviour for improved detection of early dementia: Clinical feasibility and preliminary results. *Front. Comput. Sci.* **3**, 642633. https://doi.org/10.3389/fcomp.2021.642633 (2021).
80. Fraser, K. C. et al. Predicting MCI status from multimodal Language data using cascaded classifiers. *Front. Aging Neurosci.* **11**, 205. https://doi.org/10.3389/fnagi.2019.00205 (2019).
81. Little, A., Hemsley, D., Bergmann, K., Volans, J. & Levy, R. Comparison of the sensitivity of three instruments for the detection of cognitive decline in elderly living at home. *Br. J. Psychiatry*. **150** (6), 808–814. https://doi.org/10.1192/bjp.150.6.808 (1987).
82. Katz, M. J. et al. T-MoCA: A valid phone screen for cognitive impairment in diverse community samples. *Alzheimers Dement. (Amst)*. **13** (1), e12144. https://doi.org/10.1002/dad2.12144 (2021).
83. Ye, S. et al. A computerized cognitive test battery for detection of dementia and mild cognitive impairment: instrument validation study. *JMIR Aging*. **5** (2), e36825. https://doi.org/10.2196/36825 (2022).
84. Dorociak, K. E. et al. Utility and usability of brief, remote, and self-administered cognitive assessment for older adults with intact cognition and mild cognitive impairment: A novel approach to increase access during the COVID-19 pandemic. *Alzheimer's Dement.* **17**, e055848. https://doi.org/10.1002/alz.055848 (2021).
85. Price, E. et al. Validation of a smartphone-based approach to in situ cognitive fatigue assessment. *JMIR mHealth uHealth*. **5** (8), e6333. https://doi.org/10.2196/mhealth.6333 (2017).
86. Shuman-Paretsky, M. J., Belser-Ehrlich, J. & Holtzer, R. Psychometric properties of the brief fatigue inventory in community-dwelling older adults. *Arch. Phys. Med. Rehabil.* **95** (8), 1533–1539. https://doi.org/10.1016/j.apmr.2014.03.026 (2014).
87. Yuan, J. et al. Association between the digital clock drawing test and neuropsychological test performance: Large community-based prospective cohort (Framingham heart Study). *JMIR Aging*. **4** (2), e27407. https://doi.org/10.2196/27407 (2021).
88. Brooke, J. S. U. S. A quick and dirty usability scale. In *Usability Evaluation in Industry* (eds Jordan, P. W., Thomas, B., Weerdmeester, B. A. & McClelland, A. L.) 189–194 (Taylor & Francis, London, 1996).
89. Bangor, A., Kortum, P. & Miller, J. Determining what individual SUS scores mean: Adding an adjective rating scale. *J. Usability Stud.* **4** (3), 114–123 (2009).

Acknowledgments

The authors extend their sincere gratitude to the anonymous reviewers who dedicate their time and expertise to evaluate this manuscript. The funding agency was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Author contributions

Conceptualization, G. H. Kim, B. R. Kim, J. H. Lee and J. W. Kim; Methodology, H. J. Ko, B. R. Kim, J. H. Lee G. H. Kim and J. W. Kim; Software, D. H. Kim, U. J. Kwon and S. K. Lim and J. S. Kim; Investigation H. J. Ko, J. S. Kim, W. Kim and B. H. Yun; Formal Analysis, J. S. Kim and J. H. Lee; Data Curation, H. J. Ko, J. S. Kim, W. Kim and B. H. Yun; Writing—Original Draft, H. J. Ko and J. S. Kim; Writing—Review & Editing, H. J. Ko, J. S. Kim, S. Y. Park, G. H. Kim and J. H. Lee; Visualization, J. S. Kim and J. H. Lee.; Funding Acquisition, G. H. Kim, J. H. Lee and J. W. Kim; Supervision, J. H. Lee and G. H. Kim.

Funding

This study received funding from (i) the Gangwon Information & Culture Industry Promotion Agency(GICA) and National IT Industry Promotion Agency(NIPA), grand funded by the Korea government(MSIT) (Approval and commercialization of digital medical devices for diagnosis and treatment of cognitive impairment based on artificial intelligence), (ii) the same project, funded separately by the National IT Industry Promotion Agency (NIPA) through MSIT (Grant number: PJTH09062502101800010001000100) and (iii) Korea Institute of Police Technology (KIPoT; Police Lab 2.0 program) grant funded by MSIT (RS-2023-00281194; Development for a remotely-operated testimony system for the children based on AI and Cloud technology).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-14130-9.

Correspondence and requests for materials should be addressed to J.H.L. or G.H.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025