


PENSIEVE-AI a brief cognitive test to detect cognitive impairment across diverse literacy

Received: 30 September 2024

Accepted: 14 March 2025

Published online: 23 March 2025

 Check for updates

Tau Ming Liew^{1,2,3,4}✉, Jessica Yi Hui Foo⁵, Howard Yang⁵, Sze Yan Tay⁶, Way Inn Koay⁶, King Fan Yip⁷, Simon Kang Seng Ting⁸, Kaavya Narasimhalu⁸, Weishan Li⁸, Congyuan Tan⁵, Danlin Luo⁹, Rebecca Chong⁹, Rachel Shong⁵, Christopher Sia¹⁰, Gerald Choon-Huat Koh⁴ & Julian Thumboo^{2,11}

Undiagnosed cognitive impairment is a pervasive global issue, often due to subtle nature of early symptoms, necessitating the use of brief cognitive tests for early detection. However, most brief tests are not scalable (requiring trained professionals), and are not designed for lower literacy groups (e.g. in underserved communities). Here, we developed PENSIEVE-AITM, a drawing-based digital test that is less dependent on literacy, and can be self-administered in <5 min. In a prospective study involving 1758 community-dwelling individuals aged 65 and older from Singapore (education range = 0–23 years), our deep-learning model showed excellent performance in detecting clinically-adjudicated mild cognitive impairment and dementia (AUC = 93%), comparable to traditional neuropsychological assessments (AUC = 94%, $P_{\text{comparison}} = 1.000$). Results were consistent even across education subgroups. Being less dependent on literacy, PENSIEVE-AI holds promise for broader deployment in literacy-diverse populations similar to Singapore (e.g. some Asian and lower- and middle-income countries), potentially improving early detection and intervention of cognitive impairment.

Undiagnosed cognitive impairment (CI) is a global challenge¹, with 60–90% of individuals with CI never receiving a formal diagnosis^{2,3}. Individuals with undiagnosed CI miss out on timely clinical care⁴ (e.g. cognitive enhancers, behavioral management, and caregiver support)^{5–8}, which can affect their well-being^{9,10} and increase their risk of premature nursing home placement^{11–13}. They may also not receive adequate support to manage and coordinate the care of their chronic diseases^{14,15}, resulting in suboptimal disease management, inappropriate healthcare utilization, and higher healthcare costs^{16,17}.

Recently, the importance of early diagnosis has been further underscored by growing literature on early interventions for CI^{18,19}, such as risk factor modification²⁰ and anti-amyloid monoclonal antibodies^{21,22}.

Early symptoms of CI are often subtle. Without objective cognitive tests, these symptoms are easily mistaken for normal ageing^{1,5,23,24}. To address this inherent challenge, various international bodies^{23–25} have advocated the use of brief cognitive tests to facilitate case-finding among high-risk individuals in the community²⁵. Although many brief cognitive tests exist in the literature (e.g. Montreal Cognitive

¹Department of Psychiatry, Singapore General Hospital, Outram Road, Singapore 169608, Singapore. ²SingHealth Duke-NUS Medicine Academic Clinical Programme, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore. ³Health Services and Systems Research, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore. ⁴Saw Swee Hock School of Public Health, National University of Singapore, 12 Science Drive 2, #10-01, Singapore 117549, Singapore. ⁵Government Technology Agency of Singapore, 10 Pasir Panjang Road, #10-01, Singapore 117438, Singapore. ⁶Department of Psychology, Singapore General Hospital, Singapore 169608, Singapore. ⁷Department of Geriatric Medicine, Singapore General Hospital, Outram Road, Singapore 169608, Singapore. ⁸Department of Neurology, National Neuroscience Institute, Singapore General Hospital, Singapore 169608, Singapore. ⁹Caregiving and Community Mental Health Division, Agency for Integrated Care, 5 Maxwell Road, #10-00, Singapore 069110, Singapore. ¹⁰Home Team Science & Technology Agency, 1 Stars Avenue, #12-01, Singapore 138507, Singapore. ¹¹Health Services Research Unit, Singapore General Hospital, Outram Road, Singapore 169608, Singapore. ✉e-mail: liew.tau.ming@singhealth.com.sg

Assessment²⁶, Mini-Mental State Examination²⁷, Mini-Cog²⁸, Memory Impairment Screen²⁹, Brief Cognitive Assessment Tool³⁰), most are labor-intensive and require trained professionals^{1,23,24,31}, which limit their scalability in community settings. Equally important, most tests were developed in populations with high literacy (e.g. White populations³²), and are predicated on the assumption that respondents are able to read and write in a language³³. This may limit the usefulness of cognitive tests in underserved communities with lower literacy (e.g. in some non-White communities, and in lower- and middle-income countries [LMICs]), which often have the largest number of individuals with undiagnosed CI^{32,34}. This has also led to call by the 2024 Lancet Commission on dementia care³² to address the unmet need for brief cognitive tests that are suited for individuals with lower literacy.

Digital cognitive tests hold promise as scalable tools for detecting CI in community settings, by leveraging artificial intelligence (AI) to automate the administration and scoring of brief cognitive assessments³⁵, thereby reducing dependence on trained professionals in case-finding efforts. However, despite their potential, digital cognitive tests is still a relatively nascent field³⁵. Few digital tests have undergone rigorous validation for the detection of CI in community settings³⁶, especially in populations with lower literacy³⁷. To address the unmet need for scalable case-finding tools that are suited for lower literacy groups, we have purpose-built an AI-based digital cognitive test (denoted as PENSIEVE-AITM) which has the following features:

- Designed to be self-administered (using touch-screen tablets and pre-recorded audio instructions), thus reducing dependence on trained professionals.
- Takes <5 min to complete (comprising only four drawing tasks), making it well-suited as a brief case-finding tool in community-settings.
- Relies on drawing tasks alone, thus reducing dependence on respondents' ability to read or write in a language³³, and potentially allowing broader implementation in communities with varying literacy (such as in Singapore and other Asian populations). Arguably, drawing tasks can still be affected by literacy level^{38,39}; but they are among the earliest skills that individuals develop before learning to read or write in a language, with the ability to draw shown to pre-date language development even in human civilizations³³.

Using a large, community-representative sample from Singapore, this study aimed to:

- (1) Train an image-based deep-learning model to detect mild cognitive impairment and dementia (MCI/dementia) using the four drawing tasks in PENSIEVE-AI.
- (2) Examine the effects of key demographic features (e.g. education, test language) in improving model performance, given prior literature on the potential influence of these features on drawing tasks^{38,39}.
- (3) Compare the performance of the deep-learning model to several commonly used assessment tools in detecting MCI/dementia, across participants with lower and higher literacy.

Of note, as a city-state in South-East Asia, Singapore offers a unique testbed to develop the new digital tool. Its 6-million-strong population serves as a microcosm of Asia, representing an amalgamation of Asian culture and comprising multiple Asian ethnicities, including Chinese, Malay, Indian, and other ethnic groups⁴⁰. This diversity provides a robust testing ground for assessing the new tool's performance across varied cultural and linguistic backgrounds. Additionally, the current cohort of older individuals in Singapore witnessed the country's transformation from a traditional, lower-income, Asian society to a more westernized, higher-income country⁴¹. Consequently, this cohort of older Singaporeans encompasses a wide range

of educational backgrounds, from minimal formal education to tertiary education. By validating the digital tool in such a heterogeneous population, we sought to demonstrate its potential for broader implementation in similar multiethnic and literacy-diverse settings beyond Singapore, such as in populations across East and South Asia, and potentially in some LMICs.

Results

A total of 1758 participants were included (Table 1), with 239 (13.6%) having clinically-adjudicated MCI/dementia. Given the nature of community recruitment, most cases were in early stages of CI (CDR global ≤ 1). Participants had a median age of 72 years and a median education of 10 years. Most participants could self-administer PENSIEVE-AI in <5 min (i.e. 69.1% self-administered; and 77.0% completed in under five minutes), with a median completion time of 3.7 min. However, participants with MCI/dementia were more likely to need some supervision to navigate the digital interface, and took longer to complete PENSIEVE-AI (4.6–6.7 min).

Study samples were split into approximately 40% for Training sample and 20% for Validation sample (rounded to whole numbers), with the remaining set aside as Test sample. The sample split was done using the random approach, stratified by the clinical diagnosis (i.e. normal cognition, MCI and Dementia) to ensure balanced representation of clinical diagnosis across the split samples. Following the random split, the participant characteristics were largely comparable across the three split samples, as seen in Table 2. Training sample was used to train deep-learning models to distinguish MCI/dementia from normal cognition, and Validation sample was used to fine-tune model hyperparameters. Meanwhile, Test sample (i.e. single hold-out test set) evaluated actual performance of trained models in distinguishing MCI/dementia from normal cognition, and was used to select the best-performing model and the optimal cutoffs.

Table 3 presents the results of trained models in Test sample ($n = 658$). VGG-16 performed better than SwinTransformer among image-based models (Table 3A); CLIP performed better than CNN-GRU among alternative models (Table 3B). Drawing activities (e.g. replaying audio instructions, repeated drawing attempts, long pauses between drawing strokes) further improved performance of image-based models, with VGG-16 + Drawing activities achieving the best-performing model (area under receiver-operating-characteristic curve, AUC = 93.2%; area under precision-recall curve, PR-AUC = 70.8%). Using this best model, we further examined effects of basic demographics (i.e. age, sex, education, and test language) (Table 3C); of which, only education improved model performance further (i.e. similar AUC of 93.1%, with further improvement of PR-AUC to 74.1%), and hence VGG-16 + Drawing activities + Education was selected as the final model (bold-faced in Table 3). Based on this final selected model, we conducted ablation studies to understand relative contributions of the four drawing tasks in detecting MCI/dementia (Table 3D) – Complex figure recall alone had the greatest utility in detecting MCI/dementia (AUC = 89.8%); adding Complex figure copy improved AUC to 91.8%, and further addition of Clock drawing improved AUC to 92.1%.

Table 4 compares the performance of PENSIEVE-AI and other commonly used assessment tools in Test sample ($n = 658$). PENSIEVE-AI had comparable performance to NTB (Neuropsychological Test Battery) and MoCA (Montreal Cognitive Assessment) for detecting MCI/dementia (AUC = 93.1–95.3%), even across the lower education subgroup (AUC 90.0–95.0%) and the higher education subgroup (AUC = 95.0–98.2%). In contrast, iAD8 (the Eight-item Informant Interview to Differentiate Aging and Dementia) had significantly lower AUC for MCI/dementia, particularly among participants ≤ 10 years of education (AUC = 73.2%, $p < 0.001$ when compared to PENSIEVE-AI). For the detection of dementia, all four tools (i.e. PENSIEVE-AI, NTB, MoCA, iAD8) have comparable AUCs of >90%. AUC results remained

Table 1 | Characteristics of the study participants (n = 1758)

Variable	Overall sample (n = 1758)	Normal cognition (n = 1519)	MCI (n = 195)	Dementia (n = 44)	P value ^a
Age, median (IQR) [range]	72 (68, 76) [65, 101]	71 (68, 75) [65, 93]	74 (69, 80) [65, 91]	80 (76, 82) [66, 101]	2.22 × 10 ⁻¹⁹
Years of education, median (IQR) [range]	10 (9, 12) [0, 23]	10 (10, 13) [0, 23]	10 (6, 12) [0, 21.5]	10 (2, 10) [0, 17]	5.73 × 10 ⁻¹¹
Male sex, n (%)	640 (36.4)	532 (35.0)	95 (48.7)	13 (29.5)	5.77 × 10 ⁻⁴
Ethnicity, n (%)					0.346
Chinese	1644 (93.5)	1427 (93.9)	178 (91.3)	39 (88.6)	
Malay/Indian	90 (5.1)	73 (4.8)	13 (6.7)	4 (9.1)	
Eurasian/Others	24 (1.4)	19 (1.3)	4 (2.1)	1 (2.3)	
MoCA total score, median (IQR)	26 (24, 28)	27 (25, 28)	21 (17, 24)	14 (10, 18)	2.87 × 10 ⁻⁸⁵
NTB Global Z-score, median (IQR)	-0.2 (-0.6, 0.1)	-0.1 (-0.4, 0.2)	-1.0 (-1.3, -0.7)	-1.6 (-2.0, -1.2)	3.70 × 10 ⁻¹⁰⁶
Global CDR, n (%)					3.25 × 10 ⁻¹¹⁵
0	1488 (84.6)	1475 (97.1)	13 (6.7)	0 (0.0)	
0.5	241 (13.7)	44 (2.9)	182 (93.3)	15 (34.1)	
1	21 (1.2)	0 (0.0)	0 (0.0)	21 (47.7)	
2	7 (0.4)	0 (0.0)	0 (0.0)	7 (15.9)	
3	1 (0.1)	0 (0.0)	0 (0.0)	1 (2.3)	
iAD8, median (IQR)	0 (0, 1)	0 (0, 1)	2 (0, 4)	6 (4, 8)	2.03 × 10 ⁻⁴²
Test language selected in PENSIEVE-AI, n (%)					1.56 × 10 ⁻⁵
English	1119 (63.7)	999 (65.8)	100 (51.3)	20 (45.5)	
Mandarin Chinese	639 (36.3)	520 (34.2)	95 (48.7)	24 (54.5)	
Mode of administration for PENSIEVE-AI, n (%)					3.28 × 10 ⁻²⁴
Self-administered (i.e. no supervision needed)	1214 (69.1)	1112 (73.2)	95 (48.7)	7 (15.9)	
Minimal supervision by research coordinator	544 (30.9)	407 (26.8)	100 (51.3)	37 (84.1)	
Completion of PENSIEVE-AI in <5 min, n (%)	1354 (77.0)	1234 (81.2)	109 (55.9)	11 (25.0)	2.53 × 10 ⁻²⁹
Time to complete PENSIEVE-AI (minutes), median (IQR)	3.7 (3.1, 4.8)	3.6 (3.0, 4.5)	4.6 (3.5, 6.1)	6.7 (5.1, 8.8)	9.78 × 10 ⁻²⁶

MCI mild cognitive impairment, IQR interquartile range, MoCA Montreal Cognitive Assessment, NTB neuropsychological test battery, CDR Clinical Dementia Rating, iAD8 the Eight-item Informant Interview to Differentiate Aging and Dementia (informant version).

^aTest of difference across diagnoses: Chi-square test for categorical variables, and Kruskal-Wallis test for continuous variables.

largely similar in the two sensitivity analyses (Table 4), when prevalence of MCI/dementia was increased to reflect average prevalence in most communities (i.e. 20%^{42–47} and 35%^{43,44,46,47} respectively). Additionally, several post-hoc analyses were conducted to examine potential AI biases in PENSIEVE-AI’s performance across various demographic subgroups. As seen in Supplementary Table 1, PENSIEVE-AI maintained similar AUC in detecting MCI/dementia even across the subgroups of age, sex, ethnicity, test language, and mode of administration.

Test statistics of PENSIEVE-AI are plotted in Fig. 1a. Adopting two-cutoff approach, the lower cutoff (probability ≥ 13%) had 85.7% sensitivity and 97.5% negative predictive value, and was used to rule out MCI/dementia (for individuals with probability scores below the cutoff); while the upper cutoff (probability ≥ 45%) had 98.8% specificity and 85.1% positive predictive value, and identified those who were likely to have MCI/dementia (i.e. to rule in MCI/dementia). These two cutoffs provide an intermediate range between them (greyed area in Fig. 1a), identifying those who may be at higher risk and potentially require further monitoring or assessment. The optimal cutoffs varied slightly with changing prevalence of MCI/dementia, as seen in Figs. 1b and 1c.

Effectively, the two cutoffs identified 3 risk categories for cognitive impairment: (1) Less likely to have cognitive impairment; (2) Higher risk of cognitive impairment; and (3) Likely to have cognitive impairment. These 3 categories, along with their cross-tabulation with the final diagnoses, are presented in Table 5. In the first category (i.e. Less likely to have cognitive impairment), 92–98% of individuals had normal cognition. In the second category (i.e. Higher risk of cognitive

impairment), 20–40% of individuals were diagnosed with MCI. In the third category (i.e. Likely to have cognitive impairment), 85–88% of the individuals had MCI/dementia, with a large proportion having dementia (26–36%). Distinctions between these 3 risk categories are also visible in Fig. 2. The first category (white region with probability scores below the lower cutoff) identified those with normal cognition; the third category (dark grey region with probability scores above the upper cutoff) identified almost all individuals with dementia; while the second category (light grey region between the lower and upper cutoffs) mostly captured those with MCI. Detailed results on test statistics are available in Supplementary Tables 2–4.

Discussion

Summary of findings

Brief cognitive tests are crucial for detecting subtle, early symptoms of CI. However, most require trained professionals, limiting their scalability; many were developed from high literacy populations, limiting their usefulness in lower literacy subgroups. In this study, we developed a purpose-built AI tool for early detection of CI, based on 4 drawing tasks that can be self-administered by most participants in <5 min and do not rely on the ability to read or write in a language. The new PENSIEVE-AI was trained and validated using clinically-adjudicated diagnoses in a large, prospectively-recruited community sample. Among trained deep-learning models, VGG-16 demonstrated the highest performance; adding Drawing activities (e.g. pauses between drawing strokes) significantly improved performance, adding Education marginally improved performance, and adding Test language did

Table 2 | Comparison of participant characteristics across training, validation and test samples

Variable	Training sample (<i>n</i> = 700) ^a	Validation sample (<i>n</i> = 400) ^a	Test sample (<i>n</i> = 658) ^a	<i>P</i> value ^b
Age, median (IQR)	72 (69, 76)	71 (68, 76)	71 (68, 75)	0.174
Years of education, median (IQR)	10 (8, 12)	10 (8, 12)	10 (10, 13)	0.012
Male sex, <i>n</i> (%)	257 (36.7)	147 (36.8)	236 (35.9)	0.936
Ethnicity, <i>n</i> (%)				0.470
Chinese	657 (93.9)	374 (93.5)	613 (93.2)	
Malay/Indian	32 (4.6)	21 (5.2)	37 (5.6)	
Eurasian/Others	11 (1.6)	5 (1.2)	8 (1.2)	
MoCA total score, median (IQR)	26 (24, 28)	26 (24, 28)	26 (24, 28)	0.576
NTB Global Z-score, median (IQR)	-0.2 (-0.6, 0.1)	-0.2 (-0.6, 0.1)	-0.2 (-0.6, 0.1)	0.801
Global CDR, <i>n</i> (%)				0.899
0	589 (84.1)	337 (84.2)	562 (85.4)	
0.5	98 (14.0)	55 (13.8)	88 (13.4)	
1	9 (1.3)	7 (1.8)	5 (0.8)	
2	3 (0.4)	1 (0.2)	3 (0.5)	
3	1 (0.1)	0 (0.0)	0 (0.0)	
Diagnosis, <i>n</i> (%)				0.932
Normal cognition	599 (85.6)	346 (86.5)	574 (87.2)	
Mild cognitive impairment	83 (11.9)	44 (11.0)	68 (10.3)	
Dementia	18 (2.6)	10 (2.5)	16 (2.4)	

IQR interquartile range, MoCA Montreal Cognitive Assessment, NTB neuropsychological test battery; CDR, Clinical Dementia Rating.

^aThe study samples were randomly split into approximately 40% for Training sample and 20% for Validation sample (rounded to whole numbers), with the remaining set aside as Test sample.

^bTest of difference across split samples: Chi-square test for categorical variables, and Kruskal-Wallis test for continuous variables.

not improve performance. The best-performing model (VGG-16 + Drawing activities + Education) demonstrated excellent performance in detecting MCI/dementia, comparable to detailed neuropsychological testing and MoCA. Results remained consistent across education subgroups, and when prevalence of MCI/dementia was readjusted to reflect average prevalence in most communities (i.e. ~15–25% for MCI^{42–44} and ~5–10% for dementia)^{45–47}.

Interpretation of findings

Our findings highlight a key strength of PENSIEVE-AI in particular, and digital cognitive tests in general. Despite PENSIEVE-AI's brevity (comprising only 4 test items and taking <5 min to complete), it achieves comparable AUC to detailed neuropsychological testing (which often requires at least 1–2 h to complete). This success is possibly attributed to the capture of additional data on test processes (i.e. drawing activities)³⁵, which provides valuable information to offset the reduced number of test items. As seen in our findings, this test process data was highly informative in guiding diagnosis. Plausibly, nuanced behaviors during cognitive testing better reflect subtle cognitive changes than final test scores, especially at early stages of CI. In a way, this process data mimics conventional practice of qualitative observations during detailed neuropsychological testing, providing complementary information to final test scores. Such process data would otherwise not be feasibly captured in pen-and-paper versions of brief cognitive tests, due to the labor intensity of recording such qualitative observations in routine clinical practice.

Although many digital cognitive tests have been developed in the literature³⁷, most are pilot studies with smaller samples and primarily correlated with another neuropsychological test (i.e. without evaluation over actual clinical diagnoses)³⁷. The Brain Health Assessment (BHA) is among the few that showed promising results³⁶. BHA shares some similarities with PENSIEVE-AI – both were trained on gold standard clinical diagnosis in large community samples; BHA also involves 4 tasks capturing various cognitive domains and reported similar AUC (up to 91.9%) for detecting MCI/dementia³⁶. However, BHA differs in

design from PENSIEVE-AI, requiring trained professionals to administer the 10 min test, whereas PENSIEVE-AI is designed to be primarily self-administered in <5 min. BHA was developed from White populations with high literacy (average education of 16–17 years in the development sample³⁶, with recent pilot validations in non-White populations)^{36,48,49}, in contrast to PENSIEVE-AI's development within a multiethnic Asian population with lower literacy (average education of 10 years).

In extant literature, few digital cognitive tests rely solely on drawing tasks, with clock drawing being most widely adopted⁵⁰. Consistent with the literature, our findings indicate that clock drawing alone is insufficient to detect early CI^{51,52}, and must be combined with at least one other test that evaluates another cognitive domain^{52–54}. Additionally, our findings further show that memory tasks are crucial for detecting early CI, possibly to capture early memory decline related to the most common aetiology (i.e. Alzheimer's disease). The final model – incorporating a memory task and 3 other drawing tasks – achieved an AUC of >93%, which is among the highest reported to date for drawing-based digital cognitive tests. Consistent with the literature, our findings also suggest some influence of educational attainment on drawing-based tasks^{38,39}, whereby the inclusion of education as a covariate further improved model performance (Table 3C). At the same time, the findings also affirm our initial hypothesis that drawing tasks may possibly be less affected by literacy – the education covariate only marginally improved model performance; and after including the education covariate, the final model demonstrated comparable performance to detailed neuropsychological testing even among individuals with lower literacy (Table 4).

Implications of findings

PENSIEVE-AI offers a scalable solution for case-finding of CI in the community. To address the global challenge of undiagnosed CI^{1–3}, the International Association of Gerontology and Geriatrics has advocated for annual evaluation of cognitive function among older age-groups (e.g. all individuals ≥70 years)²⁵. Yet few viable options are available to

Table 3 | Comparison of the performance of trained models for distinguishing MCI/dementia from normal cognition in the Test sample (n = 658)

Models ^a	AUC, % (95% CI) ^b	PR-AUC, % (95% CI) ^b
(A) Image-based models		
VGG-16	88.2 (84.5–91.9)	59.2 (49.1–69.3)
VGG-16 + Drawing activities^c	93.2 (91.0–95.5)	70.8 (62.7–78.9)
SwinTransformer	83.7 (78.9–88.5)	52.1 (42.6–61.7)
SwinTransformer + Drawing activity	85.2 (80.4–90.1)	62.0 (52.5–71.6)
(B) Alternative models (i.e. sequential model and zero-shot Vision Language Model)		
CNN-GRU	82.8 (77.8–87.8)	50.8 (40.7–60.9)
CNN-GRU + Drawing activities	84.9 (80.2–89.5)	56.5 (46.8–66.2)
CLIP	86.6 (82.4–90.7)	57.7 (48.1–67.3)
CLIP + Drawing activities	88.4 (84.5–92.4)	66.1 (57.9–74.3)
(C) Best model ^c + Basic demographics		
VGG-16 + Drawing activities + Age	92.4 (89.7–95.1)	70.8 (62.7–78.9)
VGG-16 + Drawing activities + Sex	91.6 (88.6–94.6)	69.5 (61.5–77.5)
VGG-16 + Drawing activities + Education^d	93.1 (90.6–95.6)	74.1 (66.4–81.8)
VGG-16 + Drawing activities + Test language	92.8 (90.3–95.3)	69.5 (61.1–77.9)
VGG-16 + Drawing activities + Age + Sex + Education	91.6 (88.7–94.4)	69.4 (61.0–77.7)
VGG-16 + Drawing activities + Age + Sex + Education + Test language	91.8 (88.9–94.8)	69.9 (61.7–78.1)
Drawing activities + Age + Sex + Education + Test language ^e	79.1 (74.0–84.2)	46.6 (37.1–56.0)
Age + Sex + Education + Test language ^e	74.5 (68.6–80.4)	37.3 (28.1–46.5)
(D) Subset of drawing tasks based on the final selected model ^d		
Complex figure copy	78.5 (73.7–83.2)	36.0 (27.6–44.4)
Simple figure copy	65.8 (59.5–72.1)	23.5 (16.7–30.3)
Clock drawing	76.2 (70.2–82.2)	42.4 (32.6–52.2)
Complex figure recall	89.8 (86.1–93.5)	63.5 (53.6–73.3)
Complex figure copy + Simple figure copy	78.0 (73.0–83.0)	33.0 (25.4–40.6)
Complex figure copy + Clock drawing	77.0 (71.2–82.8)	41.2 (31.7–50.7)
Complex figure copy + Complex figure recall	91.8 (88.8–94.8)	69.6 (61.4–77.9)
Simple figure copy + Clock drawing	75.3 (69.9–80.7)	39.2 (29.9–48.5)
Complex figure copy + Simple figure copy + Complex figure recall	90.8 (87.3–94.2)	68.0 (59.4–76.5)
Complex figure copy + Clock drawing + Complex figure recall	92.1 (89.1–95.0)	73.3 (65.4–81.1)

AUC area under the receiver operating characteristics curve, PR-AUC area under the precision-recall curve, CNN-GRU Convolutional neural network–Gated Recurrent Unit, CLIP Contrastive Language-Image Pretraining

^aFurther description on the models and model architecture is available in Supplementary Method 5.

^bThe 95% CI of AUC were computed using a non-parametric approach proposed by DeLong et al.⁷¹. The 95% CI of PR-AUC were computed using 1000 bootstrap resampling.

^cThis bolded row indicate the best-performing model (VGG-16 + Drawing activities) among the initially trained models in (A) and (B).

^dThis bolded row indicate the final selected model (VGG16 + Drawing activities + Education). This model was then used in the ablation studies in (D) to understand the relative contributions of the four drawing tasks in detecting MCI/dementia.

^eThese baseline models were provided mainly for comparison purposes, that is, by omitting VGG-16 and Drawing activities to examine the impact on the overall performance. Of note, the baseline models provided further evidence on the contributions of VGG-16 and Drawing activities to the overall performance; with AUC dropping to 79.1% when VGG-16 was omitted, and with AUC further dropping to 74.5% when Drawing activities were omitted.

date for large-scale deployment. Unlike most brief cognitive tests, PENSIEVE-AI does not require trained professionals to administer, making it well-suited as a scalable tool for case-finding of CI in large populations. Given the brevity of PENSIEVE-AI, it can be easily embedded within routinely-conducted comprehensive geriatric assessments in the community, or used as a follow-up assessment tool in conjunction with subjective questionnaires^{55,56} (i.e. to provide more conclusive evidence of CI among individuals screened positive through subjective questionnaires)⁵⁷. Considering the finding that PENSIEVE-AI can be self-administered by a majority of participants, it may also be deployed as standalone kiosks in community settings with high volumes of higher-risk older persons (e.g. primary care clinics), to allow individuals with cognitive concerns to complete brief cognitive evaluations. This approach can be especially cost-saving, as it does not require professional staff to man the community kiosks and at most

only needs lay volunteers to be on standby to supervise those with difficulty navigating the digital interface.

At the population level, PENSIEVE-AI can serve as an efficient risk-stratification tool. As shown in Fig. 1, cutoffs for PENSIEVE-AI can be adjusted depending on prevalence of MCI/dementia in different populations, to identify individuals with varying risks of CI. Low-risk individuals (<10% probability of MCI/dementia) may possibly be reassured and advised to repeat the test after a longer time horizon (e.g. 3–5 years). Intermediate-risk individuals (~25–40% probability of MCI/dementia) can be advised to consult a physician if concerned about cognition, or to repeat the test in 1 year for closer monitoring. High-risk individuals (>85% probability of MCI/dementia) will benefit from direct referral to memory clinics for further assessment and management. Notably, the high-risk group largely captures most of the individuals with dementia (Fig. 2); thus, this category can also be the

Table 4 | Performance of PENSIEVE-AI for detecting cognitive impairment in the Test sample ($n = 658$), and a comparison with the performance of other commonly used assessment tools

Assessment Tools	All education subgroups		≤ 10 years of education ^a		> 10 years of education ^a	
	AUC, % (95% CI)	<i>P</i> value ^b	AUC, % (95% CI)	<i>P</i> value ^b	AUC, % (95% CI)	<i>P</i> value ^b
Detection of MCI/dementia						
PENSIEVE-AI	93.1 (90.6–95.6)	Ref	90.0 (86.3–93.8)	Ref	98.2 (96.6–99.8)	Ref
NTB ^c	94.0 (91.6–96.4)	1.000	93.2 (90.2–96.2)	0.456	96.2 (91.4–100)	1.000
MoCA ^c	95.3 (93.2–97.4)	0.520	95.0 (92.4–97.6)	0.057	95.0 (90.9–99.1)	0.372
iAD8	75.2 (69.0–81.3)	1.48×10^{-7}	73.2 (66.3–80.1)	4.96×10^{-5}	83.5 (70.0–97.0)	0.094
Detection of dementia						
PENSIEVE-AI	96.6 (94.6–98.5)	Ref	95.5 (93.1–97.9)	Ref	96.5 (89.4–100)	Ref
NTB ^c	98.5 (97.5–99.6)	0.147	97.9 (96.3–99.4)	0.134	99.7 (99.0–100)	1.000
MoCA ^c	95.5 (90.1–100)	1.000	93.8 (85.1–100)	1.000	98.9 (97.7–100)	1.000
iAD8	92.0 (83.2–100)	0.884	90.8 (80.8–100)	1.000	98.9 (97.2–100)	1.000
Sensitivity analysis 1 (Prevalence of MCI/dementia=20%)^d						
Detection of MCI/dementia						
PENSIEVE-AI	93.4 (90.4–96.3)	Ref	90.8 (86.4–95.2)	Ref	97.9 (95.8–100)	Ref
NTB ^c	95.1 (92.5–97.8)	0.867	93.9 (90.3–97.5)	0.580	98.2 (95.4–100)	1.000
MoCA ^c	95.9 (93.6–98.2)	0.501	95.4 (92.2–98.5)	0.198	96.7 (92.8–100)	1.000
iAD8	75.5 (68.0–82.9)	1.78×10^{-5}	73.9 (65.3–82.4)	0.001	85.6 (72.2–99.0)	0.197
Detection of dementia						
PENSIEVE-AI	95.5 (92.9–98.1)	Ref	94.3 (90.7–97.8)	Ref	94.6 (83.8–100)	Ref
NTB ^c	98.0 (96.5–99.5)	0.158	97.8 (95.8–99.8)	0.114	99.3 (97.9–100)	1.000
MoCA ^c	94.8 (89.1–100)	1.000	93.2 (84.4–100)	1.000	97.7 (95.2–100)	1.000
iAD8	91.5 (82.7–100)	1.000	91.0 (81.2–100)	1.000	98.4 (96.0–100)	1.000
Sensitivity analysis 2 (Prevalence of MCI/dementia=35%)^e						
Detection of MCI/dementia						
PENSIEVE-AI	93.1 (89.5–96.7)	Ref	90.1 (84.7–95.6)	Ref	99.1 (97.3–100)	Ref
NTB ^c	93.9 (90.1–97.6)	1.000	93.9 (89.6–98.3)	0.807	94.2 (82.6–100)	1.000
MoCA ^c	95.9 (93.1–98.6)	0.630	95.2 (91.5–98.8)	0.289	96.4 (91.0–100)	1.000
iAD8	76.0 (68.0–83.9)	3.44×10^{-4}	74.2 (65.0–83.4)	0.012	85.3 (65.6–100)	0.528
Detection of dementia						
PENSIEVE-AI	92.3 (87.8–96.8)	Ref	89.9 (83.8–95.9)	Ref	93.4 (79.7–100)	Ref
NTB ^c	96.5 (93.8–99.2)	0.180	94.9 (90.9–98.9)	0.284	100 (100–100)	1.000
MoCA ^c	91.8 (84.7–98.9)	1.000	89.8 (79.5–100)	1.000	99.5 (98.2–100)	1.000
iAD8	90.6 (81.3–100)	1.000	88.0 (77.1–99.0)	1.000	99.5 (98.2–100)	1.000

AUC area under the receiver operating characteristics curve, 95% CI 95% confidence interval, MCI mild cognitive impairment, Ref reference, NTB neuropsychological test battery, MoCA Montreal Cognitive Assessment, iAD8 the Eight-item Informant Interview to Differentiate Aging and Dementia (informant version).

^aEducation subgroups were stratified based on median split.

^b*P* values were based on comparison of AUC between the PENSIEVE-AI and the respective assessment tools, using a non-parametric approach proposed by DeLong et al.⁷¹ $P < 0.05$ indicates significant difference in AUC between the PENSIEVE-AI and the respective assessment tools. *P* values were Bonferroni-adjusted to minimize the risk of Type 1 error in the context of multiple testing.

^cThe AUCs of NTB and MoCA were likely overestimated in this study, as their data informed the diagnostic process. Therefore, readers should exercise caution when comparing these results to those of PENSIEVE-AI, viewing them as general indicators rather than reflections of actual, real-world performance.

^dPrevalence of MCI/dementia was readjusted to 20% in the Test sample, based on prior meta-analytic findings that community prevalence was ~15% for MCI and ~5% for dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 256 participants with normal cognition (80%), 48 participants with MCI (15%), and 16 participants with dementia (5%).

^ePrevalence of MCI/dementia was readjusted to 35% in the Test sample, based on prior meta-lytic findings that community prevalence could be as high as ~25% for MCI and ~10% for dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 104 participants with normal cognition (65%), 40 participants with MCI (25%), and 16 participants with dementia (10%).

primary focus in communities more interested in detecting dementia than MCI. The risk-stratification approach described here is also summarized in Table 6.

Future directions

Moving forward, there are several avenues to further expand the applicability and utility of PENSIEVE-AI. An immediate direction is to translate the tool into other local languages and dialects in Singapore (e.g. Malay, Tamil, Cantonese, Hokkien, and Teochew), to enhance accessibility and inclusivity within Singapore's multiethnic population.

This effort is readily attainable, as PENSIEVE-AI is largely language-neutral (i.e. the test input is based on drawing data alone) and requires only the translation of test instructions. This approach is also supported by findings from the current study, which show that test language had minimal impact on PENSIEVE-AI's overall performance (as seen in Table 3C and Supplementary Table 1). On a related note, PENSIEVE-AI may also hold potential for broader implementation in other literacy-diverse populations similar to Singapore (e.g. those across East and South Asia and some LMICs), given current findings that it is less affected by literacy (as shown in Table 3C and Table 4).

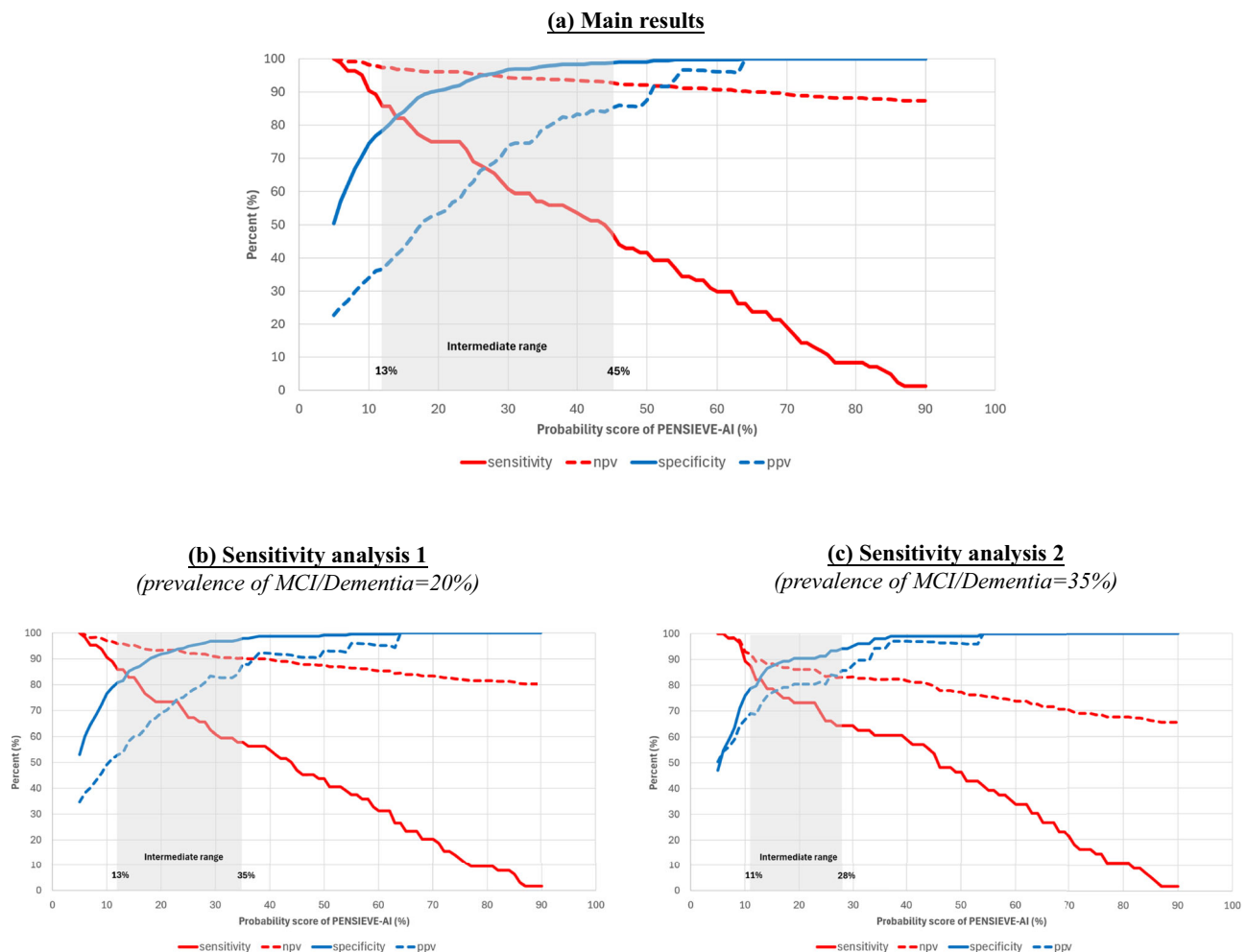


Fig. 1 | Plot of sensitivity, specificity, NPV and PPV based on probability scores of PENSIEVE-AI in the Test sample ($n = 658$). **a** Main results based on all the Test sample ($n = 658$). Adopting two-cutoff approach, the lower cutoff identifies sensitivity and negative predictive value (NPV) (red lines) which are $>85\%$ each, and is used to rule out mild cognitive impairment and dementia (MCI/dementia) when probability scores fall below this threshold. The upper cutoff identifies specificity and positive predictive value (PPV) (blue lines) which are $>85\%$ each, and is used to rule in MCI/dementia when probability scores exceed this threshold. The grey area (demarcated by the lower and upper cutoffs) represents the intermediate range, identifying those who may be at higher risk and may require further monitoring or assessment. **b** Results based on Sensitivity analysis 1, whereby the prevalence of MCI/dementia was readjusted to 20% in the Test sample, based on prior meta-analytic findings that community prevalence was $\sim 15\%$ for MCI and $\sim 5\%$ for

dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 256 participants with normal cognition (80%), 48 participants with MCI (15%), and 16 participants with dementia (5%). **c** Results based on Sensitivity analysis 2, whereby the prevalence of MCI/dementia was readjusted to 35% in the Test sample, based on prior meta-analytic findings that community prevalence could be as high as $\sim 25\%$ for MCI and $\sim 10\%$ for dementia. In the Test sample a subset of participants with MCI and dementia were randomly selected to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 104 participants with normal cognition (65%), 40 participants with MCI (25%), and 16 participants with dementia (10%). Source data are provided as a Source Data file.

However, this potential will need to be verified through future validation in populations beyond Singapore, considering prior literature on the potential cultural impact on drawing tasks similar to those used in PENSIEVE-AI^{38,58}. Lastly, although current findings demonstrated the usefulness of PENSIEVE-AI for detecting the presence of MCI/dementia cross-sectionally, efforts are also ongoing to evaluate its utility in generating a global cognitive score, alongside further longitudinal evaluations of the psychometrics of this score with respect to test-retest reliability and validity in tracking cognitive decline over time.

Limitations

Several limitations are notable. First, PENSIEVE-AI is less useful for individuals with severe visual impairment or hand movement difficulties, as it requires the ability to see on-screen figures and draw with a stylus. Second, while being language-neutral is a strength of

PENSIEVE-AI, it can also pose a limitation. The drawing-based tasks may plausibly capture less information on the language domain, potentially reducing PENSIEVE-AI's sensitivity in detecting language dysfunction. This limitation is particularly relevant in young-onset CI, where language dysfunction can be more prevalent as the initial presentation (due to higher proportions of non-Alzheimer's diseases in young-onset CI, e.g. frontotemporal lobar degeneration). Third, while digital cognitive tests have inherent strengths and appeal³⁵, they also present new barriers, particularly for individuals with lower literacy and in LMICs. For example, individuals with lower literacy may be unfamiliar with using technology, and some LMICs may have limited access to touch-screen tablets and technology infrastructure. To mitigate these limitations, we conducted extensive user design iterations in this study to tailor PENSIEVE-AI to the needs of older individuals with less digital literacy (Supplementary Method 1). We also ensured that PENSIEVE-AI is compatible

Table 5 | Cross-tabulation between the output from PENSIEVE-AI and the final diagnosis in Test sample (n = 658)

Output from PENSIEVE-AI ^a	Final diagnosis		
	Normal cognition	MCI	Dementia
Less likely to have CI, n (%) ^b	461 (97.5)	12 (2.5)	0 (0.0)
Higher risk of CI, n (%) ^b	106 (76.8)	28 (20.3)	4 (2.9)
Likely to have CI, n (%) ^b	7 (14.9)	28 (59.6)	12 (25.5)
Sensitivity analysis 1 (Prevalence of MCI/dementia = 20%)^c			
Less likely to have CI, n (%) ^d	209 (95.9)	9 (4.1)	0 (0.0)
Higher risk of CI, n (%) ^d	42 (70.0)	17 (28.3)	1 (1.7)
Likely to have CI, n (%) ^d	5 (11.9)	22 (52.4)	15 (35.7)
Sensitivity analysis 2 (Prevalence of MCI/dementia = 35%)^e			
Less likely to have CI, n (%) ^f	82 (92.1)	7 (7.9)	0 (0.0)
Higher risk of CI, n (%) ^f	16 (55.2)	12 (41.4)	1 (3.4)
Likely to have CI, n (%) ^f	6 (14.3)	21 (50.0)	15 (35.7)

CI cognitive impairment, MCI mild cognitive impairment.

^aTwo-cutoff approach was adopted for PENSIEVE-AI. The lower cutoff has high sensitivity and negative predictive value (>85% respectively), and is used to rule out MCI/dementia (for individuals with probability scores below the cutoff). The upper cutoff has high specificity and positive predictive value (>85% respectively), and identifies those who are likely to have MCI/dementia. These two cutoffs provide an intermediate range between them, identifying those who may be at higher risk and require further monitoring or assessment.

^bProbability cutoff for the main results (i.e. prevalence of MCI/dementia=13.6%): <13% (Less likely to have CI), 13–44% (Higher risk of CI), ≥45% (Likely to have CI).

^cPrevalence of MCI/dementia was readjusted to 20% in the Test sample, based on prior meta-analytic findings that community prevalence was ~15% for MCI and ~5% for dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 256 participants with normal cognition (80%), 48 participants with MCI (15%), and 16 participants with dementia (5%).

^dProbability cutoff for the first sensitivity analysis (i.e. prevalence of MCI/dementia=20%): <13% (Less likely to have CI), 13–34% (Higher risk of CI), ≥35% (Likely to have CI).

^ePrevalence of MCI/dementia was readjusted to 35% in the Test sample, based on prior meta-analytic findings that community prevalence could be as high as ~25% for MCI and ~10% for dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 104 participants with normal cognition (65%), 40 participants with MCI (25%), and 16 participants with dementia (10%).

^fProbability cutoff for the second sensitivity analysis (i.e. prevalence of MCI/dementia=35%): <11% (Less likely to have CI), 11–27% (Higher risk of CI), ≥28% (Likely to have CI).

with generic, low-specification touch-screen tablets, and requires only intermittent internet connection to generate results from cloud-hosted deep-learning models. Additionally, we designed PENSIEVE-AI as an assessor- or center-based tool (i.e. not installed on older individuals’ personal devices), so that only a limited number of tablets are needed for large-scale assessments in the community. Fourth, residual AI biases may still exist despite our best efforts to minimize the biases (e.g. through extensive recruitment efforts to obtain community-representative samples, ensuring the project team comprised a diversity of age, sex, ethnicity and professional discipline [i.e. geriatric psychiatrist, geriatrician, neurologist, psychologist], and conducting post-hoc analyses to ensure no systematic biases across demographic subgroups). As an example of residual AI biases, individuals who chose to participate in this study may differ from those who opted not to, potentially reducing the community-representativeness of recruited samples. We mitigated this limitation by employing diverse sources of community recruitment, as well as by emphasizing ‘Detect dementia early’ in our recruitment publicity (rather than a conventional invitation to participate in research, which tends to attract a distinct group of individuals) (Supplementary Method 2). Fifth, while clinicians who determined the diagnoses in this study were blinded to the drawing data from PENSIEVE-AI, they were not blinded to participants’ demographic information (e.g. age, sex and education) because such details are often essential for accurate diagnosis (e.g. information on previous levels of cognitive abilities is critical when making clinical judgment on the presence of “significant cognitive decline”)⁵⁹. While access to this demographic information might introduce potential bias in the results of PENSIEVE-AI, the risk is arguably low. As demonstrated in Table 3C, baseline models (using demographic information alone) contributed minimally to PENSIEVE-AI’s overall performance, in contrast to the substantial contributions of the drawing tasks. Sixth, PENSIEVE-AI is not intended to replace comprehensive clinical and neuropsychological assessments, as it does not provide a definitive diagnosis nor granular information on specific cognitive deficits^{18,60,61}.

Conclusions

Using a large community sample, we developed an AI-based, drawing-based cognitive test that can be self-administered in <5 min by most participants. Despite its brevity and ease of use, PENSIEVE-AI demonstrated excellent performance in detecting MCI/dementia, comparable to detailed neuropsychological testing. It can be a valuable tool in situations where detailed neuropsychological testing is not feasible, such as being embedded within community assessments or deployed as community kiosks to identify individuals requiring further intervention. As PENSIEVE-AI is less affected by language or literacy, it holds the potential for broader implementation in other literacy-diverse settings similar to Singapore, such as in populations across East and South Asia and some LMICs.

Methods

Ethical approval

This study complies with all relevant ethical regulations. The research protocol was reviewed and approved by SingHealth Centralized IRB (reference: 2021/2590). Informed consent was obtained from all participants, or their legally authorized next-of-kin (for participants without mental capacity to consent)⁶². Participants who completed the research assessments received Singapore Dollar \$80 as compensation for their time, inconvenience and transportation costs.

Study procedures

This was a nationally-funded study in Singapore to develop an AI tool for early detection of CI (Project PENSIEVE). From March 2022 to August 2024, we prospectively recruited community-dwelling older persons based on the following criteria: (1) At higher-risk of CI (i.e. aged ≥65 years²⁵, and having at least one of the three chronic diseases: diabetes mellitus, hypertension, or hyperlipidemia); (2) Able to follow simple instructions in English or Mandarin Chinese; (3) Did not having severe visual impairment that could affect the ability to complete drawing tasks (note: to ensure generalizability, participants were included as long as they could see pictures on a piece of paper held before them); and (4) Had an informant who knew the participant well

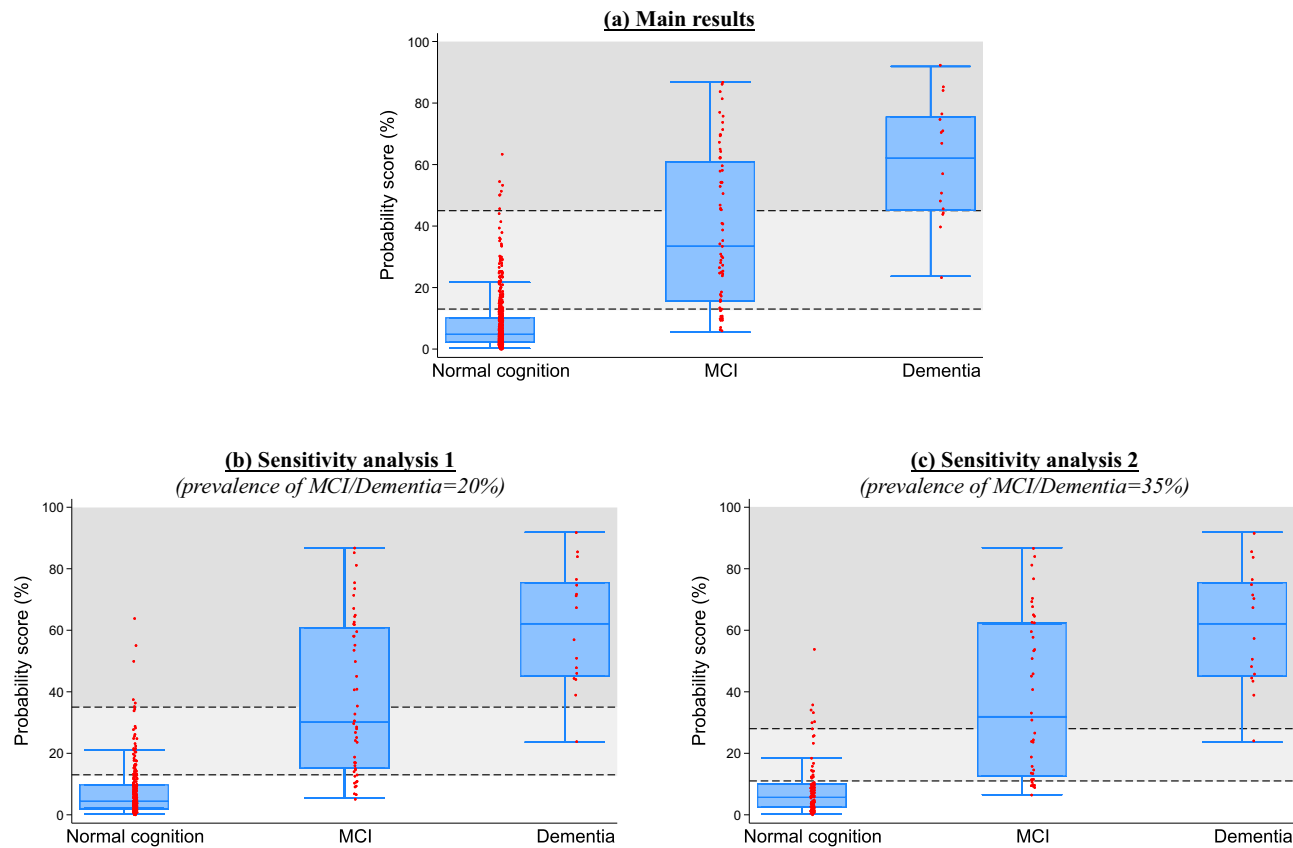


Fig. 2 | Box plots showing the distribution of PENSIEVE-AI's probability scores in the Test sample ($n = 658$). **a** Main results based on all the Test sample ($n = 658$). The box plot's center line, box limits, and whiskers denote the median, lower and upper quartiles, and 1.5 \times interquartile range, respectively. The red dots represent the individual datapoints. The two horizontal dashed lines represent the two optimal cutoffs for PENSIEVE-AI. The lower cutoff has high sensitivity and negative predictive value (>85% each), and is used to rule out mild cognitive impairment and dementia (MCI/dementia) when probability scores fall below this threshold (as shown by the white region). The upper cutoff has high specificity and positive predictive value (>85% each), and identifies individuals likely to have MCI/dementia (when probability scores exceed this threshold, as shown by the dark grey region). The light grey region (demarcated by the lower and upper cutoffs) represents the intermediate range, identifying individuals who may be at higher risk and may require further monitoring or assessment. **b** Results based on Sensitivity analysis 1, whereby the prevalence of MCI/dementia was readjusted to 20% in the Test sample,

based on prior meta-analytic findings that community prevalence was ~15% for MCI and ~5% for dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected in the Test sample to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 256 participants with normal cognition (80%), 48 participants with MCI (15%), and 16 participants with dementia (5%). **c** Results based on Sensitivity analysis 2, whereby the prevalence of MCI/dementia was readjusted to 35% in the Test sample, based on prior meta-analytic findings that community prevalence could be as high as ~25% for MCI and ~10% for dementia. In the Test sample, a subset of participants with MCI and dementia were randomly selected in the Test sample to readjust the prevalence in the dataset (see Methods section for further details). The resulting dataset comprised 104 participants with normal cognition (65%), 40 participants with MCI (25%), and 16 participants with dementia (10%). Source data are provided as a Source Data file.

Table 6 | Potential clinical implications based on output from PENSIEVE-AI

Output from PENSIEVE-AI	Risk communication	Potential implications
Less likely to have CI	<10% chance to have CI	• Repeat the test after a longer time interval (e.g. in 3–5 years)
Higher risk of CI	~25–40% chance to have CI	• Consult a physician if concerned about cognition • Repeat the test in 1 year
Likely to have CI ^a	>85% chance to have CI	• Referral to memory clinic for further assessment and management

CI cognitive impairment.

^aThis category captures most of the individuals with dementia, and can be the main focus in communities that are primarily interested in detecting dementia.

(e.g. family member or friend). Recruitment sources included 14 community roadshows by study team, clients of community partners, home visits by community volunteers, media publicity (radio, online articles, and posters), and word-of-mouth referrals from participants who had completed research assessments. To ensure that the recruited samples are representative of the community, the study's publicity materials placed much emphasis on the key message of 'Detect dementia early' (along with direct referrals to memory clinics in the

event of significant findings), rather than the conventional invitation to participate in research (which may inadvertently attract a distinct group of individuals). Samples of these publicity materials (e.g. study banner, poster, brochure) are presented in Supplementary Method 2. The recruited participants received comprehensive assessments, which included semi-structured interviews with participants and their informants, detailed neuropsychological testing, and observational notes of participants' behavior during assessments. Details on the

comprehensive assessments are available in Supplementary Methods 3, 4. Diagnoses of MCI and dementia were made via consensus conference (by 3 dementia specialists). Dementia was diagnosed using the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition) criteria⁵⁹. MCI was diagnosed using the modified Petersen criteria⁶³. Normal cognition was diagnosed when participants were found not to have dementia or MCI.

Measures

The new digital cognitive test (henceforth denoted as PENSIEVE-AI™) comprises 4 drawing tasks, namely: (1) complex figure copy; (2) simple figure copy; (3) clock drawing; and (4) complex figure recall (i.e. recall of complex figure from the first task). Respondents were provided a 12.4-inch touch-screen tablet and a stylus, and asked to follow on-screen voice instructions to complete the 4 drawing tasks on the tablet (of note, the same drawing prompts were used for every assessment, to ensure consistency in administration across different assessments). Throughout the 4 tasks, drawing activities (e.g. drawing motions, replaying audio instructions, repeated drawing attempts) were also captured within the tablet and included as input data for model training. The 4 tasks were designed to cover the cognitive domains of Visuospatial abilities (tasks 1 and 2)⁶⁴, Attention and Executive function (task 3)⁶⁴, Memory (task 4)⁶⁴, and Language (ability to follow through audio instructions). Details on user design of PENSIEVE-AI are available in Supplementary Method 1. Of note, PENSIEVE-AI was completed by participants before the start of comprehensive assessments, and the dementia specialists who were determining the diagnosis in consensus conference were blinded to the drawings and drawing activities from PENSIEVE-AI (but not blinded to participants' demographic information such as age, sex and education).

Three alternative assessment tools were included in the analyses as comparators to PENSIEVE-AI. These tools represent three common types of assessments in cognitive evaluations: an informant questionnaire (iAD8; the Eight-item Informant Interview to Differentiate Aging and Dementia)⁵⁵, a brief cognitive test (MoCA; Montreal Cognitive Assessment)²⁶ and detailed neuropsychological testing (NTB; Neuropsychological Test Battery)⁶⁵. They are briefly described in the next paragraph, with further details available in Supplementary Method 4. It is important to note that the dementia specialists in this study were blinded to iAD8 results, but not blinded to those of MoCA or NTB. Given that the data from MoCA and NTB were used to inform the diagnostic process, the performance of MoCA and NTB were likely overestimated in this study (i.e. actual performance of MoCA and NTB would be lower than reported). Accordingly, readers should exercise caution when comparing these results to those of PENSIEVE-AI, interpreting them as general indicators rather than reflections of actual, real-world performance.

iAD8⁵⁵ is a brief questionnaire that requires informants to rate changes in participants' cognition and function in the past few years (through yes/no responses). Its 8 items can be completed in ~3–5 min, with higher scores indicating greater cognitive problems. MoCA²⁶ comprises 12 items that test participants in various cognitive domains. It can be completed in ~15–20 min, with higher scores reflecting better cognitive function. The NTB⁶⁵ takes ~60 min to complete, and includes seven neuropsychological tests measuring the key cognitive domains of Visuospatial abilities (Benson Complex Figure Copy), Working memory (Craft Story 21 Immediate Recall), Delayed memory (Craft Story 21 Delayed Recall and Benson Complex Figure Recall), Language (Verbal Fluency–Animal), Attention/Processing speed (Trail Making Test–Part A), and Executive function (Trail Making Test–Part B).

Statistics & reproducibility

In Training and Validation samples, we experimented with image-based models (i.e. VGG-16⁶⁶ and Swin Transformer)⁶⁷, sequential models (i.e. CNN-GRU)⁶⁸, and zero-shot vision-language models (i.e.

CLIP)⁶⁹. While the drawings and the drawing activities were the main input data for model training, we also explored the inclusion of basic demographic features (e.g. age, sex, educational attainment, and test language) to assess their potential effects in improving model performance. Models were trained using focal loss⁷⁰ due to the unbalanced dataset. Focal loss is a technique that helps the model to pay more attention to harder-to-classify examples that it often misclassifies, rather than those it already classifies correctly. It achieves this by dynamically adjusting the contribution of each example to the overall training process. Based on the predicted probability of each example, it reduces the loss contribution from well-classified examples, thereby allowing the model to focus more on harder, misclassified examples. This method is particularly useful in situations where there are far more examples of one class (i.e. normal cognition) compared to another (i.e. MCI/dementia), which can otherwise overwhelm the model. By focusing on the harder, less frequent examples, the model thus improves its ability to identify those rarer examples of MCI/dementia. Further details on model training are presented in Supplementary Method 5.

In Test sample, predicted probabilities of the trained models were compared using area under receiver-operating-characteristic curve (AUC) – supplemented by area under precision-recall curve (PR-AUC) – to select the best-performing model for PENSIEVE-AI in distinguishing MCI/dementia from normal cognition. Thereafter, AUC of PENSIEVE-AI was compared to the AUCs of 3 other commonly-used assessment tools (i.e. iAD8, MoCA, NTB) using a non-parametric approach proposed by DeLong et al.^{71–74}, with analyses stratified by education subgroups (i.e. ≤10 years of education and >10 years of education, based on median-split). A two-cutoff approach^{75–79} was adopted for PENSIEVE-AI. First cutoff has high sensitivity and negative predictive value (>85% respectively), and is used to rule out MCI/dementia (i.e. when probability scores fall below first cutoff). Second cutoff has high specificity and positive predictive value (>85% respectively), and identifies those who are likely to have MCI/dementia. This two-cutoff approach has been recommended in recent literature⁷⁹, as it enhances test performance^{75–78}, reduces effects of prevalence on test performance⁷⁶, and prioritizes healthcare resources for those more likely to benefit⁷⁵.

As secondary analysis, the performance of PENSIEVE-AI was evaluated for distinguishing dementia from non-dementia. Additionally, two sensitivity analyses were conducted in Test sample to evaluate robustness of results when prevalence of MCI/dementia was readjusted to reflect average prevalence in most communities:

- (1) Prevalence of MCI/dementia was artificially readjusted to 20%, based on prior meta-analytic findings that community prevalence was ~15% for MCI^{42–44} and ~5% for dementia^{45–47}. Readjustment of prevalence was done by randomly selecting only a subset of participants with MCI and normal cognition – for each participant with dementia, 3 participants with MCI and 16 participants with normal cognition were randomly selected (i.e. so that the final dataset corresponded to 5% prevalence for dementia and 15% prevalence for MCI).
- (2) Prevalence of MCI/dementia was artificially readjusted to 35%, based on prior meta-analytic findings that community prevalence could be as high as ~25% for MCI^{43,44} and ~10% for dementia^{46,47}. Readjustment of prevalence was done by randomly selecting only a subset of participants with MCI and normal cognition – for each participant with dementia, 2.5 participants with MCI and 6.5 participants with normal cognition were randomly selected (i.e. so that the final dataset corresponded to 10% prevalence for dementia and 25% prevalence for MCI).

Statistical analyses were conducted in Stata (version 18). No statistical method was used to predetermine sample size. During the

initial study planning, we estimated that at least 1000 samples would be required (with each sample providing 4 drawing data), guided by a well-known classification challenge in recent literature (Tiny ImageNet)⁸⁰. No data were excluded from the analyses.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study contains sensitive information about the study participants and they did not provide consent for public data sharing. The current approval by SingHealth Centralized IRB (reference: 2021/2590) does not include data sharing. A minimal dataset could be shared by request from a qualified academic investigator for the sole purpose of replicating the present study, provided the data transfer is in agreement with Singapore's legislation on the general data protection regulation and a data sharing agreement has been signed. Contact person: corresponding author. Source data are provided with this paper.

Code availability

All Stata codes required to conduct the analyses reported here are attached as a file in the source data.

References

- World Health Organization. *Dementia: a public health priority*. (World Health Organization, 2012).
- Lang, L. et al. Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. *BMJ Open* **7**, e011146 (2017).
- Liu, Y., Jun, H., Becker, A., Wallick, C. & Mattke, S. Detection Rates of Mild Cognitive Impairment in Primary Care for the United States Medicare Population. *J. Prev. Alzheimers Dis.* **11**, 7–12 (2024).
- Liew, T. M. A 4-Item Case-Finding Tool to Detect Dementia in Older Persons. *J. Am. Med. Dir. Assoc.* **20**, 1529–1534.e1526 (2019).
- Burns, A. & Iliffe, S. Dementia. *BMJ (Clin. Res. ed.)* **338**, b75 (2009).
- Ying, J., Yap, P., Gandhi, M. & Liew, T. M. Iterating a framework for the prevention of caregiver depression in dementia: a multi-method approach. *Int Psychogeriatr.* **30**, 1119–1130 (2018).
- Liew, T. M. & Lee, C. S. Reappraising the Efficacy and Acceptability of Multicomponent Interventions for Caregiver Depression in Dementia: The Utility of Network Meta-Analysis. *Gerontologist* **59**, e380–e392 (2019).
- Liew, T. M. Neuropsychiatric symptoms in early stage of Alzheimer's and non-Alzheimer's dementia, and the risk of progression to severe dementia. *Age Ageing*, <https://doi.org/10.1093/ageing/afab044> (2021).
- Thyrian, J. R. et al. Effectiveness and Safety of Dementia Care Management in Primary Care: A Randomized Clinical Trial. *JAMA psychiatry* **74**, 996–1004 (2017).
- Vickrey, B. G. et al. The effect of a disease management intervention on quality and outcomes of dementia care: a randomized, controlled trial. *Ann. Intern. Med.* **145**, 713–726 (2006).
- Cepoiu-Martin, M., Tam-Tham, H., Patten, S., Maxwell, C. J. & Hogan, D. B. Predictors of long-term care placement in persons with dementia: a systematic review and meta-analysis. *Int J. Geriatr. Psychiatry* **31**, 1151–1171 (2016).
- Jennings, L. A. et al. Health Care Utilization and Cost Outcomes of a Comprehensive Dementia Care Program for Medicare Beneficiaries. *JAMA Intern. Med.* **179**, 161–166 (2019).
- Spijker, A. et al. Effectiveness of nonpharmacological interventions in delaying the institutionalization of patients with dementia: a meta-analysis. *J. Am. Geriatrics Soc.* **56**, 1116–1128 (2008).
- Bott, N. T. et al. Systems Delivery Innovation for Alzheimer Disease. *Am. J. Geriatr. Psychiatry* **27**, 149–161 (2019).
- Elliott, R. A., Goeman, D., Beanland, C. & Koch, S. Ability of older people with dementia or cognitive impairment to manage medicine regimens: a narrative review. *Curr. Clin. Pharm.* **10**, 213–221 (2015).
- Persson, S. et al. Healthcare costs of dementia diseases before, during and after diagnosis: Longitudinal analysis of 17 years of Swedish register data. *Alzheimers Dement.* **18**, 2560–2569 (2022).
- Chay, J., Koh, W. P., Tan, K. B. & Finkelstein, E. A. Healthcare burden of cognitive impairment: Evidence from a Singapore Chinese health study. *Ann. Acad. Med. Singap.* **53**, 233–240 (2024).
- Liew, T. M. Developing a Brief Neuropsychological Battery for Early Diagnosis of Cognitive Impairment. *J. Am. Med. Dir. Assoc.* **20**, 1054 e1011–1054.e1020 (2019).
- Liew, T. M. Distinct trajectories of subjective cognitive decline before diagnosis of neurocognitive disorders: Longitudinal modelling over 18 years. *J. Prev. Alzheimers Dis.* 100123, <https://doi.org/10.1016/j.tjpad.2025.100123> (2025).
- Ngandu, T. et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. *Lancet* **385**, 2255–2263 (2015).
- van Dyck, C. H. et al. Lecanemab in Early Alzheimer's Disease. *N. Engl. J. Med* **388**, 9–21 (2023).
- Sims, J. R. et al. Donanemab in Early Symptomatic Alzheimer Disease: The TRAILBLAZER-ALZ 2 Randomized Clinical Trial. *Jama* **330**, 512–527 (2023).
- Prince, M., Bryce, R. & Ferri, C. *World Alzheimer report 2011: the benefits of early diagnosis and intervention*. (Alzheimer's Disease International, 2011).
- Gerontological Society of America. A 4-Step Process to Detecting Cognitive Impairment and Earlier Diagnosis of Dementia: Approaches and Tools for Primary Care Providers. <https://www.geron.org/images/gsa/kaer/gsa-kaer-toolkit.pdf> (2017).
- Morley, J. E. et al. Brain health: the importance of recognizing cognitive impairment: an IAGG consensus conference. *J. Am. Med. Dir. Assoc.* **16**, 731–739 (2015).
- Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
- Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198 (1975).
- Borson, S., Scanlan, J., Brush, M., Vitaliano, P. & Dokmak, A. The Mini-Cog: a cognitive 'vital signs' measure for dementia screening in multi-lingual elderly. *Int J. Geriatr. Psychiatry* **15**, 1021–1027 (2000).
- Buschke, H. et al. Screening for dementia with the memory impairment screen. *Neurology* **52**, 231–238 (1999).
- Mansbach, W. E., MacDougall, E. E. & Rosenzweig, A. S. The Brief Cognitive Assessment Tool (BCAT): a new test emphasizing contextual memory, executive functions, attentional capacity, and the prediction of instrumental activities of daily living. *J. Clin. Exp. Neuropsychol.* **34**, 183–194 (2012).
- Chong, S. A., Abidin, E., Vaingankar, J., Ng, L. L. & Subramaniam, M. Diagnosis of dementia by medical practitioners: a national study among older adults in Singapore. *Aging Ment. health* **20**, 1271–1276 (2016).
- Livingston, G. et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. *The Lancet*, [https://doi.org/10.1016/S0140-6736\(24\)01296-0](https://doi.org/10.1016/S0140-6736(24)01296-0).
- Ardila, A. et al. Illiteracy: the neuropsychology of cognition without reading. *Arch. Clin. Neuropsychol.* **25**, 689–712 (2010).
- Kalaria, R. et al. The 2022 symposium on dementia and brain aging in low- and middle-income countries: Highlights on research,

- diagnosis, care, and impact. *Alzheimers Dement* **20**, 4290–4314 (2024).
35. Staffaroni, A. M., Tsoy, E., Taylor, J., Boxer, A. L. & Possin, K. L. Digital Cognitive Assessments for Dementia: Digital assessments may enhance the efficiency of evaluations in neurology and other clinics. *Pr. Neurol. (Fort Wash. Pa)* **2020**, 24–45 (2020).
 36. Tsoy, E. et al. BHA-CS: A novel cognitive composite for Alzheimer's disease and related disorders. *Alzheimers Dement (Amst.)* **12**, e12042 (2020).
 37. Cubillos, C. & Rienzo, A. Digital Cognitive Assessment Tests for Older Adults: Systematic Literature Review. *JMIR Ment. Health* **10**, e47487 (2023).
 38. Rosselli, M. & Ardila, A. The impact of culture and education on non-verbal neuropsychological measurements: a critical review. *Brain Cogn.* **52**, 326–333 (2003).
 39. Maestri, G. et al. Cultural influence on clock drawing test: A systematic review. *J. Int Neuropsychol. Soc.* **29**, 704–714 (2023).
 40. Department of Statistics Singapore. *Population Trends 2024*, <https://www.singstat.gov.sg/-/media/files/publications/population/population2024.ashx> (2024).
 41. World Bank Group. *The World Bank in Singapore*, <https://www.worldbank.org/en/country/singapore/overview> (2024).
 42. Bai, W. et al. Worldwide prevalence of mild cognitive impairment among community dwellers aged 50 years and older: a meta-analysis and systematic review of epidemiology studies. *Age Ageing* **51**, <https://doi.org/10.1093/ageing/afac173> (2022).
 43. Hu, C. et al. The prevalence and progression of mild cognitive impairment among clinic and community populations: a systematic review and meta-analysis. *Int Psychogeriatr.* **29**, 1595–1608 (2017).
 44. Song, W. X. et al. Evidence from a meta-analysis and systematic review reveals the global prevalence of mild cognitive impairment. *Front Aging Neurosci.* **15**, 1227112 (2023).
 45. Prince, M. et al. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement* **9**, 63–75.e62 (2013).
 46. Fiest, K. M. et al. The Prevalence and Incidence of Dementia: a Systematic Review and Meta-analysis. *Can. J. Neurol. Sci.* **43**, S3–S50 (2016).
 47. Cao, Q. et al. The Prevalence of Dementia: A Systematic Review and Meta-Analysis. *J. Alzheimers Dis.* **73**, 1157–1166 (2020).
 48. Haddad, R. et al. TabCAT Brain Health Assessment: Preliminary validation in a multicultural Israeli population. *Alzheimers Dement* **20**, <https://doi.org/10.1002/alz.091664> (2024).
 49. Ogbuagu, C. et al. Feasibility and Determinants of Performance for a Tablet-Based Cognitive Assessment Tool in Rural and Urban Southeast Nigeria. *J. Alzheimers Dis.* **101**, 175–182 (2024).
 50. Chan, J. Y. C. et al. Evaluation of Digital Drawing Tests and Paper-and-Pencil Drawing Tests for the Screening of Mild Cognitive Impairment and Dementia: A Systematic Review and Meta-analysis of Diagnostic Studies. *Neuropsychol. Rev.* **32**, 566–576 (2022).
 51. Youn, Y. C. et al. Use of the Clock Drawing Test and the Rey-Osterrieth Complex Figure Test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimers Res Ther.* **13**, 85 (2021).
 52. Amini, S. et al. An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test. *J. Alzheimers Dis.* **83**, 581–589 (2021).
 53. Souillard-Mandar, W. et al. DCTclock: Clinically-Interpretable and Automated Artificial Intelligence Analysis of Drawing Behavior for Capturing Cognition. *Front Digit Health* **3**, 750661 (2021).
 54. Jannati, A. et al. Digital Clock and Recall is superior to the Mini-Mental State Examination for the detection of mild cognitive impairment and mild dementia. *Alzheimers Res Ther.* **16**, 2 (2024).
 55. Galvin, J. E. et al. The AD8: a brief informant interview to detect dementia. *Neurology* **65**, 559–564 (2005).
 56. Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H. Jr., Chance, J. M. & Filos, S. Measurement of functional activities in older adults in the community. *J. Gerontol.* **37**, 323–329 (1982).
 57. Liew, T. M. Active case finding of dementia in ambulatory care settings: a comparison of three strategies. *Eur. J. Neurol.* **27**, 1867–1878 (2020).
 58. Gonthier, C. Cross-cultural differences in visuo-spatial processing and the culture-fairness of visuo-spatial intelligence tests: an integrative review and a model for matrices tasks. *Cogn. Res Princ. Implic.* **7**, 11 (2022).
 59. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. (Amer Psychiatric Pub Incorporated, 2013).
 60. Jacova, C., Kertesz, A., Blair, M., Fisk, J. D. & Feldman, H. H. Neuropsychological testing and assessment for dementia. *Alzheimers Dement* **3**, 299–317 (2007).
 61. Ang, L. C., Yap, P., Tay, S. Y., Koay, W. I. & Liew, T. M. Examining the Validity and Utility of Montreal Cognitive Assessment Domain Scores for Early Neurocognitive Disorders. *J. Am. Med Dir. Assoc.* **24**, 314–320.e312 (2023).
 62. Attorney-General's Chambers of Singapore. *Mental Capacity Act 2008*, <https://sso.agc.gov.sg/Act/MCA2008> (2024).
 63. Petersen, R. C. & Morris, J. C. Mild cognitive impairment as a clinical entity and treatment target. *Arch. Neurol.* **62**, 1160–1163 (2005).
 64. Salimi, S. et al. Can visuospatial measures improve the diagnosis of Alzheimer's disease? *Alzheimer's. Dement (Amst., Neth.)* **10**, 66–74 (2018).
 65. Weintraub, S. et al. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). *Alzheimer Dis. Assoc. Disord.* **32**, 10–17 (2018).
 66. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, <https://doi.org/10.1109/CVPR.2016.90> (2016).
 67. Liu, Z. et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002, <https://doi.org/10.1109/iccv48922.2021.00986> (2021).
 68. Sarvadevabhatla, R. K., Kundu, J. & R, V. B. Enabling My Robot To Play Pictionary: Recurrent Neural Networks For Sketch Recognition. *Proceedings of the 24th ACM international conference on Multimedia*, 247–251, <https://doi.org/10.1145/2964284.2967220> (2016).
 69. Radford, A. et al. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763 <https://doi.org/10.48550/arXiv.2103.00020> (2021).
 70. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).
 71. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
 72. Yu, J., Yap, P. & Liew, T. M. The optimal short version of the Zarit Burden Interview for dementia caregivers: diagnostic utility and externally validated cutoffs. *Aging Ment. Health* **23**, 706–710 (2019).
 73. Liew, T. M. & Yap, P. A 3-Item Screening Scale for Caregiver Burden in Dementia Caregiving: Scale Development and Score Mapping to the 22-Item Zarit Burden Interview. *J. Am. Med Dir. Assoc.* **20**, 629–633.e612 (2019).
 74. Liew, T. M. The Optimal Short Version of Montreal Cognitive Assessment in Diagnosing Mild Cognitive Impairment and Dementia. *J. Am. Med Dir. Assoc.* **20**, 1055.e1051–1055.e1058 (2019).
 75. Dautzenberg, G., Lijmer, J. G. & Beekman, A. T. F. The Montreal Cognitive Assessment (MoCA) with a double threshold: improving

- the MoCA for triaging patients in need of a neuropsychological assessment. *Int Psychogeriatr.* **34**, 571–583 (2022).
76. Landsheer, J. A. Impact of the Prevalence of Cognitive Impairment on the Accuracy of the Montreal Cognitive Assessment: The Advantage of Using two MoCA Thresholds to Identify Error-prone Test Scores. *Alzheimer Dis. Associated Disord.* **34**, 248–253 (2020).
 77. Thomann, A. E., Berres, M., Goettel, N., Steiner, L. A. & Monsch, A. U. Enhanced diagnostic accuracy for neurocognitive disorders: a revised cut-off approach for the Montreal Cognitive Assessment. *Alzheimer's. Res Ther.* **12**, 39 (2020).
 78. Swartz, R. H. et al. Validating a Pragmatic Approach to Cognitive Screening in Stroke Prevention Clinics Using the Montreal Cognitive Assessment. *Stroke* **47**, 807–813 (2016).
 79. Jack, C. R. Jr et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimers Dement*, <https://doi.org/10.1002/alz.13859> (2024).
 80. Meta A. I. *Tiny ImageNet*, <https://paperswithcode.com/dataset/tiny-imagenet> (2023).

Acknowledgements

This research is funded by the Singapore Prime Minister Office's Smart Nation and Digital Government Office (grant number: I_20092346). Separately, T.M.L. is supported by the Singapore Ministry of Health's National Medical Research Council (grant numbers: HCSAINV23jul-0001, NMRC/CG2/005e/2022-SGH, MOH-SEEDFD22apr-0001). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. We thank the following persons for their assistance in the conduct of this study: Yanling Tan, Alcantara Leicester Shawn, Kai Xin Choo, Megan Cheng Mun Choy, Xin Tong Tan, Lydia Jia En Cheong, Jane Mee Chin Liew, Xiao Hui Ng, Spencer Peng Ming Yuen, Alcey Li Chang Ang, Xiu Ping Chue, A/Prof. Chian Min Loo, and A/Prof. Charles Thuan Heng Chuah. We also thank the following community organizations for their support in participant recruitment: Thyne Hua Kwan Moral Charities (AMK 645, Bukit Merah View, Beo Crescent), Kreta Ayer Senior Activity Centres, NTUC Health Active Ageing Centre (Lengkok Bahru), Silver Generation Office (Outreach arm of Agency for Integrated Care), People's Association, Precious Active Ageing Centre, Montfort Care, Yong-en Care Centre, Presbyterian Community Services, Lions Befrienders Service Association Singapore. Last but not least, we also express our sincere gratitude to the study participants and informants for their support to make this research possible.

Author contributions

Each author has made substantial contributions to this article in terms of conceptualization (T.M.L., C.S.), study design (T.M.L.), funding acquisition (T.M.L., C.S., J.T., G.C.H.K.), participant recruitment (T.M.L., D.L., R.C., C.T.), training of research coordinators (T.M.L., S.Y.T., W.I.K.), data acquisition (T.M.L.), data audits (T.M.L.), consensus diagnosis (T.M.L., K.F.Y., S.K.S.T., K.N., W.L., S.Y.T., W.I.K.), initial user design prototype (T.M.L.), refinement of user design and software (R.S.), development of deep learning models (J.Y.H.F., H.Y.), statistical analysis (T.M.L.), data interpretation (T.M.L., J.Y.H.F., H.Y.), writing the initial manuscript draft

(T.M.L., J.Y.H.F.), and overall supervision as the lead principal investigator (T.M.L.). All authors have reviewed the drafts and participated substantively in revisions. All authors have approved the submitted version. All authors have agreed to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated, resolved, and the resolution documented in the literature.

Competing interests

Singapore Health Services (SingHealth) and Government Technology Agency of Singapore (GovTech) have submitted a provisional patent to the Intellectual Property Office of Singapore pertaining to the methods and compositions of PENSIEVE-AI (provisional application No. 10202500088 R), with T.M.L. and J.Y.H.F. listed as the co-inventors. T.M.L. has provided consultation to Lundbeck. K.N. has provided consultation to Takeda. The remaining authors declare no conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-58201-x>.

Correspondence and requests for materials should be addressed to Tau Ming Liew.

Peer review information *Nature Communications* thanks Chi Udeh-Momoh, who co-reviewed with Tamlyn Watermeyer, and Payam Barnaghi, who co-reviewed with Antigone Fogel, for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://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