Large-scale clinical validation of a remote digital health cognitive assessment in patients with stroke against established clinical scales

Dragos-Cristian Gruia[1,2], Valentina Giunchiglia[1,4], Aoife Coghlan[1,2], Sophie Brook[1], Soma Banerjee[2], Jo Kwan[2], Peter J. Hellyer[3], Adam Hampshire[1,3], Fatemeh Geranmayeh[1,2]

1. Department of Brain Sciences, Imperial College London, London, UK
2. Imperial College Healthcare NHS Trust, London, UK
3. Centre for Neuroimaging Sciences, IoPPN, King's College London, London, UK
4. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Corresponding Author: Dr. Fatemeh Geranmayeh
Email fatemeh.geranmayeh00@imperial.ac.uk

Abstract

Background: Stroke is a major contributor to global mortality and morbidity, with significant cognitive sequalae which is often underdiagnosed and undertreated. Unsupervised remote digital health technology is expected be a game changer in addressing such diagnostic challenges. However, it remains unclear whether these remote assessments match the reliability, validity, and sensitivity of standard neuropsychological tests.

Methods: We developed a novel online test (IC3) that offers deep phenotyping of cognitive sequalae after stroke, requiring minimal clinician input. Large scale normative data from over 6,000 UK-based older adults was collected, and state-of-the-art Bayesian modelling was used to determine patient-specific impairment scores in a cohort of 90 stroke survivors. The battery was evaluated for test-retest reliability, differences in performance between supervised and non-supervised settings, learning effects across multiple timepoints, psychometric validity and against standard neuropsychological tests.

Findings: IC3 had high test-retest reliability, minimal learning effects and consistent performance across supervised vs. non-supervised settings. Additionally, a novel Bayesian predictive model showcased the platform's heightened sensitivity to detecting cognitive impairments in patients with stroke, outperforming conventional neuropsychological screening tools. Furthermore, IC3 derived scores had a stronger association with patient-reported functional outcomes, highlighting its potential in guiding diagnosis and monitoring.

Interpretation: IC3 digital cognitive assessment tool is a feasible, sensitive and efficient when tested in patients with stroke. Its robustness, scalability, and predictive capabilities offer advantages over existing tools, potentially transforming both clinical diagnosis and large-scale research in stroke-related cognitive impairment.

Funding: This research is funded by the UK Medical Research Council (MR/T001402/1).

Introduction

Stroke is a leading cause of death and disability globally, with frequent cognitive sequalae affecting three-quarters of survivors.[1] The spectrum of cognitive impairments associated with stroke encompasses domain-specific deficits, such as aphasia, neglect, and memory impairment as well as domain-general deficits usually associated with co-existing small vessel disease such as executive / attentional dysfunction and reduction in processing speed.[2] Collectively, these impairments have a detrimental impact on poststroke recovery, engagement with therapeutic interventions and lower quality of life among patients.[3,4] As a result, early detection of these impairments has been recommended by key stake holders and national and international guidelines for stroke management.[5–7]

Despite a lack of universally accepted approach for identifying post-stroke cognitive deficits, there is wide consensus that stroke survivors should be screened for cognitive impairment during their hospital admission using a stroke-specific cognitive screen.[5,8] There remains considerable variability in clinical practice, with deployed assessments ranging from extensive neuropsychological test batteries tailored to a specific cognitive domain (e.g., language), to global measures of cognition using shallow non-stroke specific tools like the MOCA.[9] The choice often being influenced by personal preferences, availability, cost and time pressures, leaving little prospect for generalisability of findings across sites.

Availability of a cost-effective, reliable, scalable and comprehensive screening tool that provides a stroke-specific deep phenotyping of cognition would be game-changing for clinical diagnosis as well as enabling much-needed large-scale population-based research for studying the mechanisms of post-stroke cognitive recovery. To address this gap, we developed a novel digital adaptive tool: The Imperial Comprehensive Cognitive Assessment in Cerebrovascular Disease (IC3).[10] IC3 is a digital assessment designed to require minimal input from a clinician in detecting both domain-general and domain-specific cognitive deficits in patients after stroke. It involves a comprehensive account of cognitive domains known to be impaired following stroke, including memory, language, executive function, attention, numeracy, praxis as well as hand motor ability and clinical and neuropsychiatric questionnaires.

Here, we present and analyse extensive normative data derived from over 6,000 UK-based older adults using the IC3 tool, highlighting its ability to map cognition at a large scale, in a time- and cost-efficient manner. Leveraging the large sample size, and state-of-the-art Bayesian modelling, we created patient-specific predictive scores that captured the effects of demographic and neuropsychiatric variables as well as language proficiency, dyslexia and device on cognitive performance. IC3's validity as a remote cognitive screening tool was demonstrated through a thorough set of sub-analyses that showed its reliability and feasibility, equivalent performance between supervised and non-supervised settings, and minimal learning effects across 4 timepoints.

In patients with stroke, the IC3-derived scores mapped well onto well-established first-line clinical screening tools (MOCA), with superior sensitivity to mild cognitive impairment and with patient reported functional outcome measures (post-stroke quality of life), explaining twice as much variation than the first-line clinical screening tool MOCA. Using a data-driven approach (factor analysis) we showed that IC3 scores map intuitively into cognitive domains often affected in stroke. Overall, our results show that IC3 platform has high feasibility and validity and that it can be used to monitor cognition across diverse populations with stroke and cerebrovascular disease.

Method

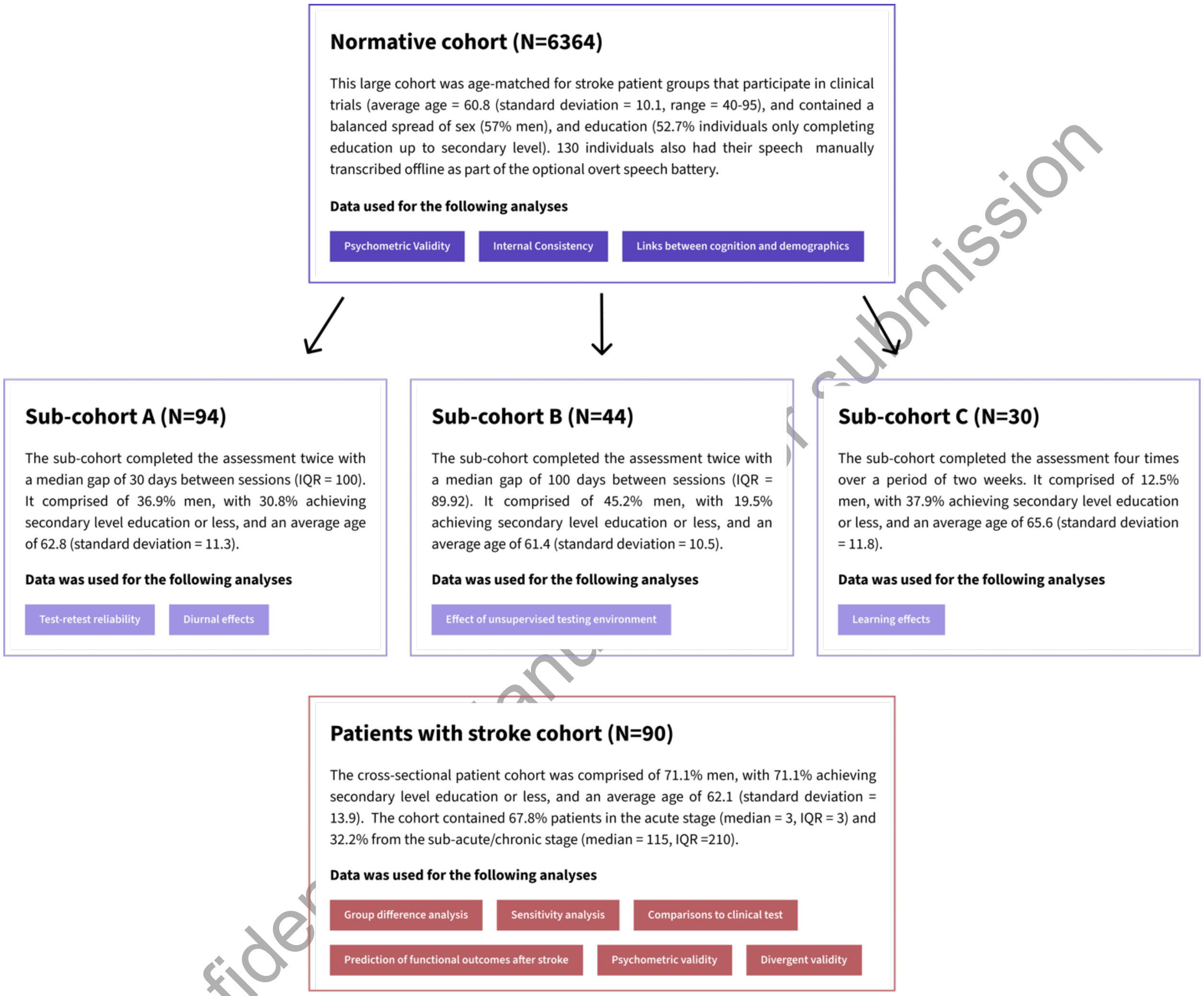
Participants

The different participant cohorts employed in this study are described below and shown in Figure 1.

Normative cohort. A study invitation was extended to 25,000 individuals, over the age of 40, residing in Great Britain, all of whom had previously participated in the Great British Intelligence Test (a nationwide initiative aimed at mapping cognition within the general population) and consented to being recontacted for other studies.[11] Ultimately, 7,095 healthy older adults provided their consent and initiated the IC3 cognitive battery, with 5,639 participants (79.5%) successfully completing all 22 tasks. The data was collected remotely online, between October and November 2022. Participation in the study was voluntary with no monetary incentive. In addition, we collected data from 138 individuals via the Imperial Clinical Research Facility participant registry for reliability and validity purposes. The combined normative cohort contained 6364 healthy older adults, following participant exclusion and pre-processing steps (discussed in detail in Supplementary Material 3.).

Normative sub-cohorts used for the reliability and validity analyses (Figure 1, Sub Cohort A-C). A smaller sample of controls who performed the assessment multiple times was collected separately to test the battery's reliability and validity. These were recruited using the Imperial Clinical Research Facility participant registry. A total of 94 participants (Sub-cohort A) completed the assessment twice, as part of the test-retest reliability analysis. Of the 94 participants, 44 (Sub-cohort B) performed the assessment both supervised (in person), and unsupervised (remote) with the order counterbalanced.

The remainder (50/94) completed both sessions unsupervised. A subset (N=30/94; Sub-Cohort C) completed the assessment remotely 4 times in total over the course of two weeks, to allow for the estimation of any potential learning effects. No monetary reward was provided, aside for travel reimbursements, wherever appropriate.



Normative cohort (N=6364)

This large cohort was age-matched for stroke patient groups that participate in clinical trials (average age = 60.8 (standard deviation = 10.1, range = 40-95), and contained a balanced spread of sex (57% men), and education (52.7% individuals only completing education up to secondary level). 130 individuals also had their speech manually transcribed offline as part of the optional overt speech battery.

Data used for the following analyses

Sub-cohort A (N=94)

The sub-cohort completed the assessment twice with a median gap of 30 days between sessions (IQR = 100). It comprised of 36.9% men, with 30.8% achieving secondary level education or less, and an average age of 62.8 (standard deviation = 11.3).

Data was used for the following analyses

Sub-cohort B (N=44)

The sub-cohort completed the assessment twice with a median gap of 100 days between sessions (IQR = 89.92). It comprised of 45.2% men, with 19.5% achieving secondary level education or less, and an average age of 61.4 (standard deviation = 10.5).

Data was used for the following analyses

Sub-cohort C (N=30)

The sub-cohort completed the assessment four times over a period of two weeks. It comprised of 12.5% men, with 37.9% achieving secondary level education or less, and an average age of 65.6 (standard deviation = 11.8).

Data was used for the following analyses

Patients with stroke cohort (N=90)

The cross-sectional patient cohort was comprised of 71.1% men, with 71.1% achieving secondary level education or less, and an average age of 62.1 (standard deviation = 13.9). The cohort contained 67.8% patients in the acute stage (median = 3, IQR = 3) and 32.2% from the sub-acute/chronic stage (median = 115, IQR =210).

Data was used for the following analyses

Group difference analysis
Sensitivity analysis
Comparisons to clinical test
Prediction of functional outcomes after stroke
Psychometric validity
Divergent validity

Figure 1. Overview of the study cohorts discussed in this study, along with the analysis conducted on each group.

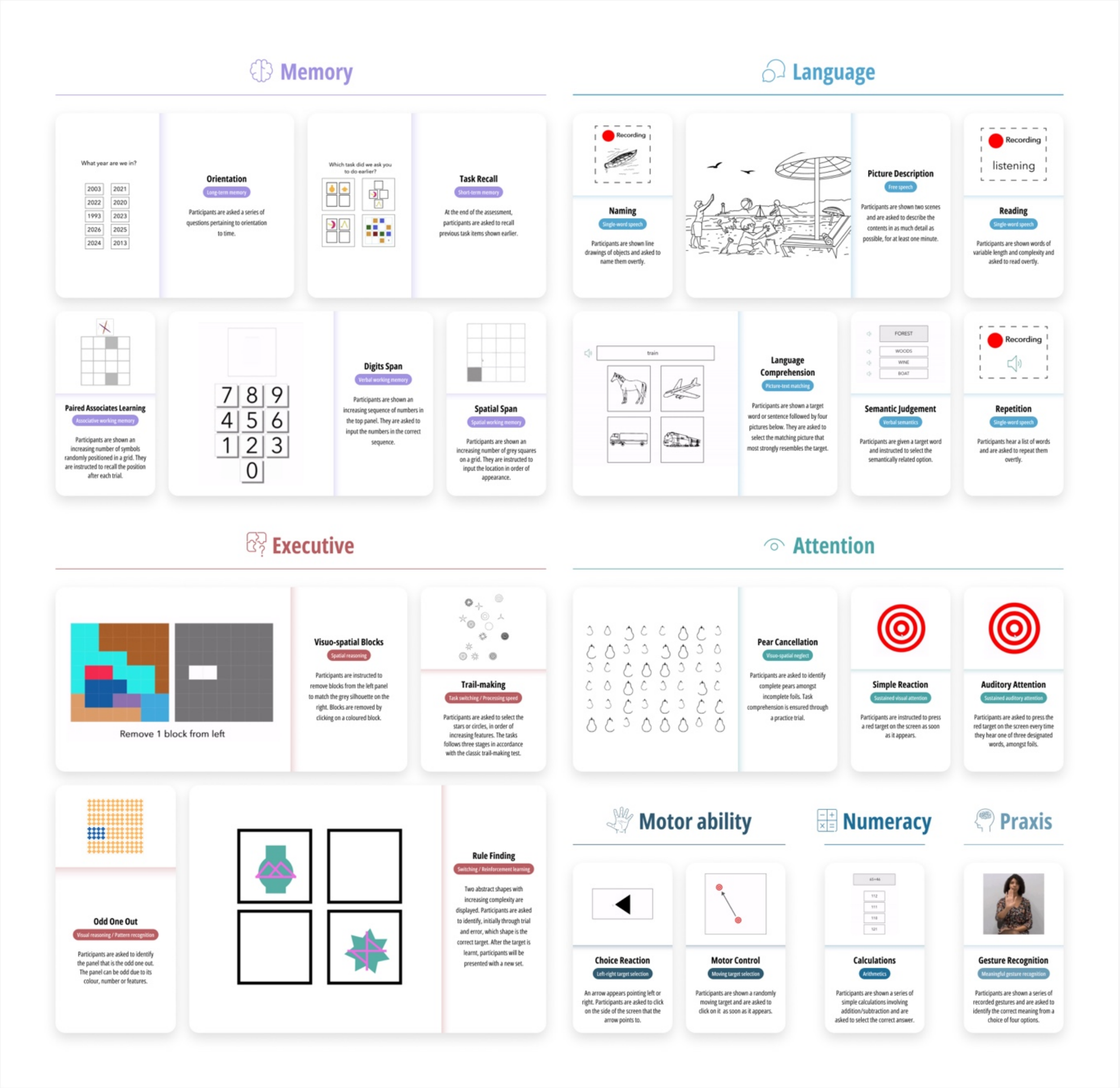
Patient with stroke cohort. Patients with radiologically confirmed stroke were recruited from Imperial College Healthcare NHS Trust. Exclusion criteria included pre-stroke diagnosis of dementia, severe visuospatial problems and mental health diagnoses, fatigue limiting engagement with the IC3 beyond 15 minutes and inability to understand task instructions. Consecutively recruited patients underwent the digital IC3 assessment (see Supplementary Material 5.1. for detailed demographic information). The scores were compared with clinical pen-and-paper cognitive screens (MOCA).

All participants gave informed consent. The data was acquired as part of a longitudinal observational clinical study approved by UK's Health Research Authority (Registered under NCT05885295; IRAS:299333; REC:21/SW/0124). Patients also underwent blood biomarkers testing and brain imaging which will not be analysed in this paper. A lesion overlap map is shown in Supplementary Material 5.2. demonstrates the lesion distribution in patients who had MRI brain imaging.

Cognitive tasks, speech-based tasks and neuropsychiatric questionnaires

A graphical overview and a detailed description of the cognitive assessments are available in Figure 2 and Supplementary Materials 1. respectively. These cover 18 short cognitive tasks with additional 4 optional speech production tasks covering a wide range of cognitive domains. The tasks are followed by several clinically-validated questionnaires (Apathy Evaluation Scale Fatigue Scale, Geriatric Depression Scale, and Instrumental Activities of Daily Living).[12–15]

IC3 is freely available via a web-browser on any modern device (smartphone, tablet, computer/laptop) via a weblink and is implemented through Cognitron (https://www.cognitron.co.uk/), a state-of-the art platform for remote neuropsychological testing rapidly being adopted by large scale population studies both in the UK and internationally.[16,17]


Memory

Orientation
Long-term memory
Participants are asked a series of questions pertaining to orientation to time.

Task Recall
Short-term memory
At the end of the assessment, participants are asked to recall previous task items shown earlier.

Paired Associates Learning
Associative working memory
Participants are shown an increasing number of symbols randomly positioned in a grid. They are instructed to recall the position after each trial.

Digits Span
Verbal working memory
Participants are shown an increasing sequence of numbers in the top panel. They are asked to input the numbers in the correct sequence.

Spatial Span
Spatial working memory
Participants are shown an increasing number of grey squares on a grid. They are instructed to input the location in order of appearance.

Language

Naming
Single-word speech
Participants are shown line drawings of objects and asked to name them overtly.

Picture Description
Free speech
Participants are shown two scenes and are asked to describe the contents in as much detail as possible, for at least one minute.

Reading
Single-word speech
Participants are shown words of variable length and complexity and asked to read overtly.

Language Comprehension
Picture-text matching
Participants are shown a target word or sentence followed by four pictures below. They are asked to select the matching picture that most strongly resembles the target.

Semantic Judgement
Verbal semantics
Participants are given a target word and instructed to select the semantically related option.

Repetition
Single-word speech
Participants hear a list of words and are asked to repeat them overtly.

Executive

Visuo-spatial Blocks
Spatial reasoning
Participants are instructed to remove blocks from the left panel to match the grey silhouette on the right. Blocks are removed by clicking on a coloured block.

Trail-making
Task switching / Processing speed
Participants are asked to select the stars or circles, in order of increasing features. The tasks follows three stages in accordance with the classic trail-making test.

Odd One Out
Visual reasoning / Pattern recognition
Participants are asked to identify the panel that is the odd one out. The panel can be odd due to its colour, number or features.

Rule Finding
Switching / Reinforcement learning
Two abstract shapes with increasing complexity are displayed. Participants are asked to identify, initially through trial and error, which shape is the correct target. After the target is learnt, participants will be presented with a new set.

Attention

Pear Cancellation
Visuo-spatial neglect
Participants are asked to identify complete pears amongst incomplete foils. Task comprehension is ensured through a practice trial.

Simple Reaction
Sustained visual attention
Participants are instructed to press a red target on the screen as soon as it appears.

Auditory Attention
Sustained auditory attention
Participants are asked to press the red target on the screen every time they hear one of three designated words, amongst foils.

Motor ability

Choice Reaction
Left-right target selection
An arrow appears pointing left or right. Participants are asked to click on the side of the screen that the arrow points to.

Motor Control
Moving target selection
Participants are shown a randomly moving target and are asked to click on it as soon as it appears.

Numeracy

Calculations
Arithmetics
Participants are shown a series of simple calculations involving addition/subtraction and are asked to select the correct answer.

Praxis

Gesture Recognition
Meaningful gesture recognition
Participants are shown a series of recorded gestures and are asked to identify the correct meaning from a choice of four options.

Figure 2. Graphical overview of 22 IC3 tasks organised by the main cognitive domains tested: memory, language, executive, attention, motor ability, numeracy and praxis. The four optional speech production tasks (naming, repetition, reading, picture description) allow speech to be recorded and manually analysed off-line.

Cognitive and speech data filtering

Given the remote nature of the cognitive testing, to ensure that the normative data was derived from fully engaged, healthy participants who understood the task instructions, we implemented three levels of data filtering (subject-level, task-level, and trial-level), by filtering out data that was invalid or of poor quality. See Supplementary Materials 3. for a detailed description. For the four optional speech production tasks data was manually analysed offline by 6 trained expert raters with high Inter-rater reliability (Intraclass correlation=0.87-0.90). See Supplementary Materials 2 for speech marking guidelines.

IC3 validation

A set of sub-analyses were used to assess IC3's reliability, equivalent performance between supervised and non-supervised settings, and learning effects (see Supplementary Materials 5.3. – 5.7.).

Bayesian modelling on the large normative sample derives patient-specific predictive scores

Constrained by relatively small normative sample sizes, existing cognitive batteries have traditionally been limited to accounting for the effects of demographic factors by stratifying the normative sample into even smaller sub-groups, often limited to one or two variables (age and occasionally education).[9,18–20] In this study, we leverage the large normative sample of 6364 individuals and Bayesian modelling (see below), to create patient-specific predictive scores with higher precision, accounting for 8 additional confounding factors (age, sex, education, language proficiency, device, depression, dyslexia, anxiety; highlighted with an asterisk in Supplementary Material 5.1.).

State-of-the-art Bayesian posterior predictions for modelling the relationship between cognition and confounding factors in controls. Bayesian regression analyses containing all 8 covariates were performed separately for all tasks, to estimate the effects of each of the covariates on individual task performance. For the optional speech production tasks (repetition, naming, reading), the smaller sample (N=130) precluded the inclusion of depression, anxiety and dyslexia as covariates. For full details of Bayesian regression models see Supplementary Materials 4.1.- 4.2.)

Patient-specific impairment thresholds. Bayesian modelling was used to create patient-specific impairment thresholds correcting for the afore-mentioned confounding variables. This was done by (i) training the Bayesian regression models on the normative sample as described above, (ii) using the derived posterior distributions to estimate patient-specific predicted performance, converted to std units, and iii) subtracting the observed patient performance (also in std units) from the predicted performance derived from step 'ii'. A resulting negative "deviation from norm" score suggests that the patient had a deficit in that specific task by a given magnitude in std units, such that a score of -1 represents an impairment of 1 std from corresponding demographically matched control group. Using these estimates, boundaries for the severity of the cognitive impairments were assigned as -1.5 (mild), -2.0 (moderate) and -2.5 (severe) std below the mean.

Results

Normative sample

Participant characteristics. The cleaned normative sample consisted of 6364 individuals. The number of participants included varied for each task (N= 4782-6290) depending on task-specific data filtering and the fact that tasks at the end of the assessment had fewer timepoints. Supplementary Material 5.1. outlines the demographics and additional factors that could affect cognitive performance.

Relationships between cognition and confounding demographic factors in the normative sample. Standardised coefficients were obtained from task-specific Bayesian regression models for the 8 confounding covariates (mean R^2:11.29%; range: 1.3-53.5). The lower range of the R^2 was driven by tasks with ceiling effects and low inter-subject variability in the controls (e.g., 3.9/4.0 and 3.8/4.00 for Orientation and Task Recall mean performance respectively).

The effects of the covariates on cognition are shown in Figure 3 where warm and cool colours represent positive and negative standardised coefficients. Cognitive performance generally worsened with age as shown by negative coefficients shown in purple. The exception is Semantic Judgement in keeping with previous literature demonstrating age-related improvement in language function.[21] Dyslexia and English as second language had a strong negative effect on performance particularly on tasks involving language and numeracy skills. Device was a strong confounding factor in task that relied on speed and motor dexterity as the main outcome measure (e.g. Simple Rection Time, Choice Reaction Time and Motor Control). This effect is understandable given faster responses on touch screen compared to mouse/trackpad-operated devices. Higher education levels were related to better cognitive performance across tasks that involved language and numeracy, with the least effect on tasks that primarily captured motor dexterity (e.g., Motor Control, Choice Reaction Task, Simple Reaction Time tasks).

Overall, the regression models provide intuitive and interpretable relationships between cognition and the 8 confounding covariates, providing a solid argument for inclusion of these co-variates in deriving patient-specific impairment thresholds.

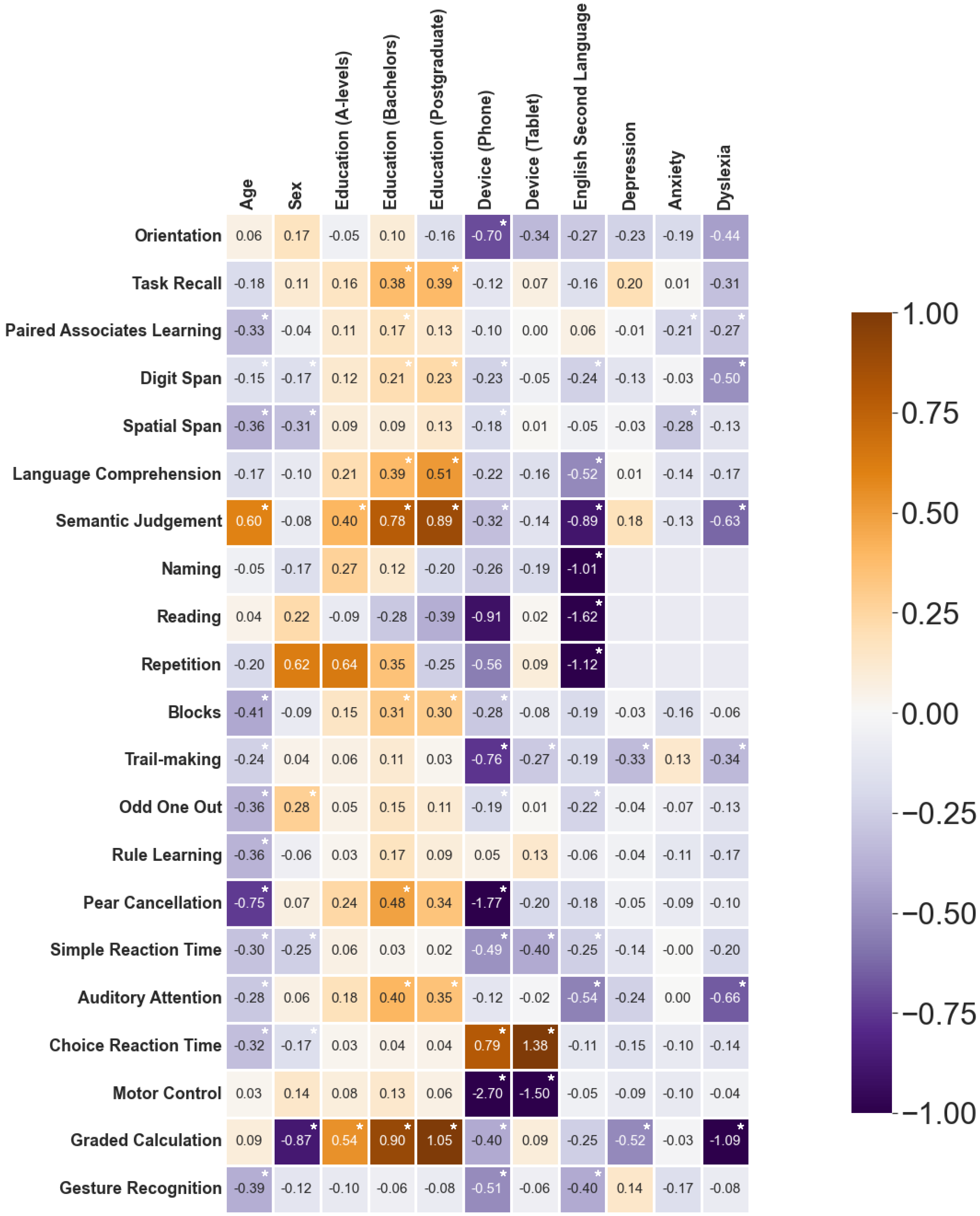


Figure 3. **Relationship between cognitive performance and 8 confounding factors in the large normative sample, quantified via standardised regression coefficients.** The regression reference categories for Education and Device were '≤GCSE-educated' and 'Desktop Computer', respectively. Negative coefficients (cool colours) indicate the covariate was associate with worse cognitive performance. Empty cells represent

coefficient not used in the modelling for the optional speech production tasks given the smaller normative sample. '*'=Uncorrected significant standardised coefficients.

Reliability of IC3

Internal Consistency

The task-level internal consistency was generally good, particularly for tasks with high variability on the primary outcome measure, with 17/21 tasks showing split-half alpha values surpassing the standard threshold for good internal consistency (α=0.70). See Supplementary Material 5.3. for more detailed methodology. Nevertheless, a small subset of tasks exhibited lower alpha values (i.e., Orientation, Task Recall, Gesture Recognition, and Calculation, marked with '†' in Table 1). A well-known limitation of the split-half reliability is its dependency on a large number of trials, with shorter tasks having inherently low reliability estimates.[22] Thus, the low alpha values may be attributed to the low number of trials (4-6) in these short screening tasks, despite the tasks being preferable in clinical settings due to their lower burden on patients. A further factor is the ceiling effect in these short tasks such that, a single error results in a significant change in the trial's relative ranking and estimated covariance between trials.

High test-retest reliability across time

There was no significant difference (FDR corrected) in performance between two sessions on any of the IC3 tasks with strong equivalence across all measures in sub-cohort A (Table 1, Supplementary Material 5.4.). There was moderate to high correlation between sessions for tasks with high variance within the group, and smaller correlation for those with low variance due to ceiling effect despite similar group means across sessions. Overall, these results demonstrate a stable performance of control group on the IC3 across time.

Furthermore, diurnal effects on performance were examined through effects of inter-session time interval and that of time of day when the assessment was performed. Our results show that these time-related factors do not modulate change in cognitive ability across the two sessions (Supplementary Material 5.5.).

Table 1. Internal consistency (full normative cohort) and test-retest reliability (sub-cohort A) analyses of the older adult controls.

		Internal Consistency		Test-retest reliability sub-study							
Domain	Task	N	Alpha	N	Mean Time 1	Mean Time 2	Group mean difference P-value	Effect size	Test of equivalence P-value	R-value	Normally Distributed
Memory	Orientation	6290	0.14†	92	3.88	3.9	0.74	0.07	<0.001	0.1†	No
	Task Recall	4782	0.04†	91	3.79	3.87	0.34	0.2	<0.001	0.31†	No
	Paired Associates Learning	5721	0.73	87	4.68	4.97	0.29	0.22	0.03	0.31	Yes
	Digits Span	5279	0.73	86	6.69	7.0	0.1	0.24	0.03	0.59	Yes
	Spatial Span	5867	0.70	90	5.08	5.22	0.46	0.13	0.01	0.45	Yes
Language	Comprehension	5025	0.76	91	18.44	18.52	0.49	0.05	0.01	0.45	No
	Semantic Judgement	4995	0.94	92	22.4	22.58	0.46	0.14	<0.001	0.62	No
	Naming	132ˆ	0.79	85	0.97	0.98	0.41	0.13	<0.001	0.46	No
	Reading	132ˆ	0.72	87	0.99	0.99	0.19	0.23	<0.001	0.12†	No
	Repetition	132ˆ	0.83	86	0.99	0.98	0.12	0.33	<0.001	-0.05†	No
Executive	Blocks	5039	0.70	54‡	6.52	7.28	0.1	0.22	0.03	0.77	Yes
	Trail-making*	5189	0.81	82	1.02	1.05	0.92	0.01	<0.001	0.08†	No
	Odd One out	5672	0.75	91	10.73	11.23	0.1	0.23	0.03	0.59	Yes
	Rule Learning	5485	0.83	85	6.75	7.24	0.1	0.27	0.04	0.42	No
Attention	Pear Cancellation	6219	0.88	85	0.98	0.99	0.1	0.41	<0.001	0.3†	No
	Simple Reaction Time*	5510	0.95	91	390.66	386.12	0.66	0.06	<0.001	0.68	Yes
	Auditory Attention	5757	0.73	91	35.24	35.44	0.35	0.15	<0.001	0.32†	No
Motor Ability	Choice Reaction Time	5181	0.88	87	54.59	53.9	0.65	0.08	0.01	0.28	Yes
	Motor Control	5208	0.88	93	28.96	28.95	0.97	0.01	<0.001	0.14†	No
Numeracy	Calculation	5681	0.33†	91	7.86	7.88	0.82	0.04	<0.001	0.09†	No
Praxis	Gesture Recognition	5280	0.05†	91	7.48	7.55	0.59	0.09	<0.001	0.36†	No

Note: ‘N’= Sample size. ‘ˆ’= Lower sample size due to tasks requiring manual transcription of overt speech. ‘Alpha’= 5000 iterations of permutation-based randomly split-half sampled trial-level data were used to calculate internal consistency. ‘†’= Unreliable alpha values due to lack of variance in controls and low number of trials. Nevertheless, test-retest reliability suggests that all these tasks are stable across time. ‘Group Difference P-value’= Obtained from a paired-samples t-test or Wilcoxon test depending on the stated distribution. ‘Effect size’= Cohen’s D. ‘Test of equivalence P-value’= P-value obtained from a parametric on non-parametric test of equivalence dependent on the stated distribution; significant p-values equate equivalence. P-values are FDR-corrected. ‘R-value’= Coefficient obtained from a Pearson or Spearman correlation dependent on the stated distribution. ‘*’= Indicates tasks where higher values suggest worse performance. ‘‡’= Lower N caused by exclusion of 32 participants whose version of Blocks differed between sessions due to improvements in task design. ‘†’ in R-value = Indicates that the correlation values may be low due to lack of variance in both sessions, despite highly similar mean values between sessions.

Equivalent cognitive performance between unsupervised and supervised testing environment

With the exception of two tasks (Choice Reaction Time and Motor Control), there was no difference between performance of supervised and supervised testing environment. (Supplementary Material 5.6.). Only two tasks that measured hand motor dexterity were performed better in a supervised setting. Given that the performance on these tasks was strongly dependent on the device type used (Figure 3), additional Bayesian regression modelling of these two tasks was conducted with device and environment as predictors which confirmed that the session effects for these tasks were driven not by the supervised/unsupervised environment but by the device used to perform the assessment. (Supplementary Material 5.6.). Thus, we conclude that there were no direct effects of unsupervised testing environment.

IC3 is robust to learning effects across four timepoints

We examined the potential effect of learning on repeated testing, in 30 controls (sub-cohort C) who completed the IC3 four times over the course of two weeks. The results showed that IC3 assessment has minimal learning effects within the expected testing time interval of cognitive tool (detailed analyses shown in Supplementary Material 5.7.).

IC3 in patients with stroke

IC3 is sensitive to group differences across all tasks

Data from 90 patients with stroke was analysed. The IC3 testing was performed during the acute post-stroke phase in 68% (3±3 days post stroke) and in the sub-acute/chronic post-stroke phase in 32% (115±210 days post stroke). See Supplementary Materials 5.1. and 5.2..

The task-level average group performance was calculated in ‘deviation from expected’ standard deviation units as described above. Healthy controls significantly outperformed patients across all tasks, showing moderate-to-large effect sizes in the majority of the tasks after accounting for the effects of eight confounding factors (19/21 tasks, $P<0.05$ FDR corrected; d>0.43 for all tasks). See Figure 4 and Supplementary Material 5.8. for detailed statistical results.

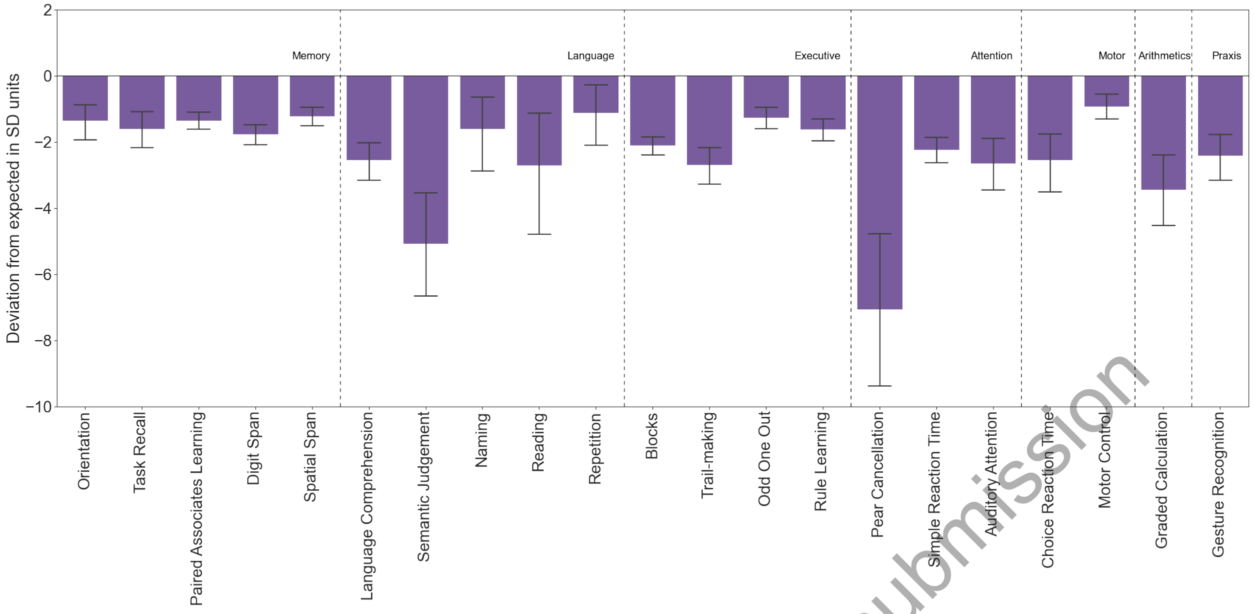


Figure 4: Patients with stroke had significantly worse performance than controls (FDR corrected). Error bar= 95% CI. 0= mean control performance.

The IC3 online cognitive assessment correlates with clinical neuropsychological scores and functional impairment after stroke

A data driven 'IC3 global composite score', was derived from factor analysis based on combined data from patients and controls (Supplementary Material 5.9.). This had a positive loading on individual tasks and accounted for 47% of the variance (Figure 5A). Six group factors, each loading on a subset of tests were also derived, intuitively mapping onto Executive Function (F1), Language/Numeracy (F2), Working Memory (F3), Attention (F4), Motor Ability (F5) and Memory Recall (F6). The factor analysis fit was robust (CFI=0.94, RMSEA=0.01) with good internal consistency (omega=0.79).

In keeping with task-level results, IC3 global score (g) was significantly lower in patients compared to controls (P<.001, Supplementary Material 5.9.). Furthermore, the IC3 global score and total MOCA scores were significantly correlated in patients (r(78)=0.58, R^2=0.33, *P*< 0.001, Fig. 5B), indicating that the IC3 performance maps intuitively onto clinically validated neuropsychological screens.

To assess the external validity of IC3, IC3 global performance was also related to functional impairment after stroke as defined by the Instrumental Activities of Daily Living (IADL) score.[15] Worse global cognitive performance on the IC3 was associated with worse functional impairment after stroke (r(78)=0.51, R^2=0.26, *P*<0.001, Fig. 5C).

Conversely, MOCA had a considerably weaker relationship with functional deficits post-stroke, explaining approximately half of the variation explained by the IC3 (r(78)=0.38, R^2=0.14, *P*< 0.001). A separate linear regression analysis was conducted to quantify this difference, with MOCA and IC3 as independent predictors, and IADL as the dependent variable. The results show that there was no longer a main effect of MOCA on functional impairment when accounting for the IC3 global score (*P*=.22), whilst the IC3 global score remained highly significant (*P*=.004). Moreover, the inclusion of MOCA only explained an additional 1% of variance, suggesting that MOCA does not bring any additional information beyond what is captured by the IC3. The interaction between MOCA scores and global IC3 was also found to be significant (*P*=.04), suggesting that the IC3 was more strongly predictive of the IADL than the MOCA.

Furthermore, we demonstrate strong divergent validity for IC3 (Supplementary Material 5.10.), as shown by an expected lack of correlation between cognitive performance in patients and variables known not to be related to cognition (i.e. admission cholesterol levels).

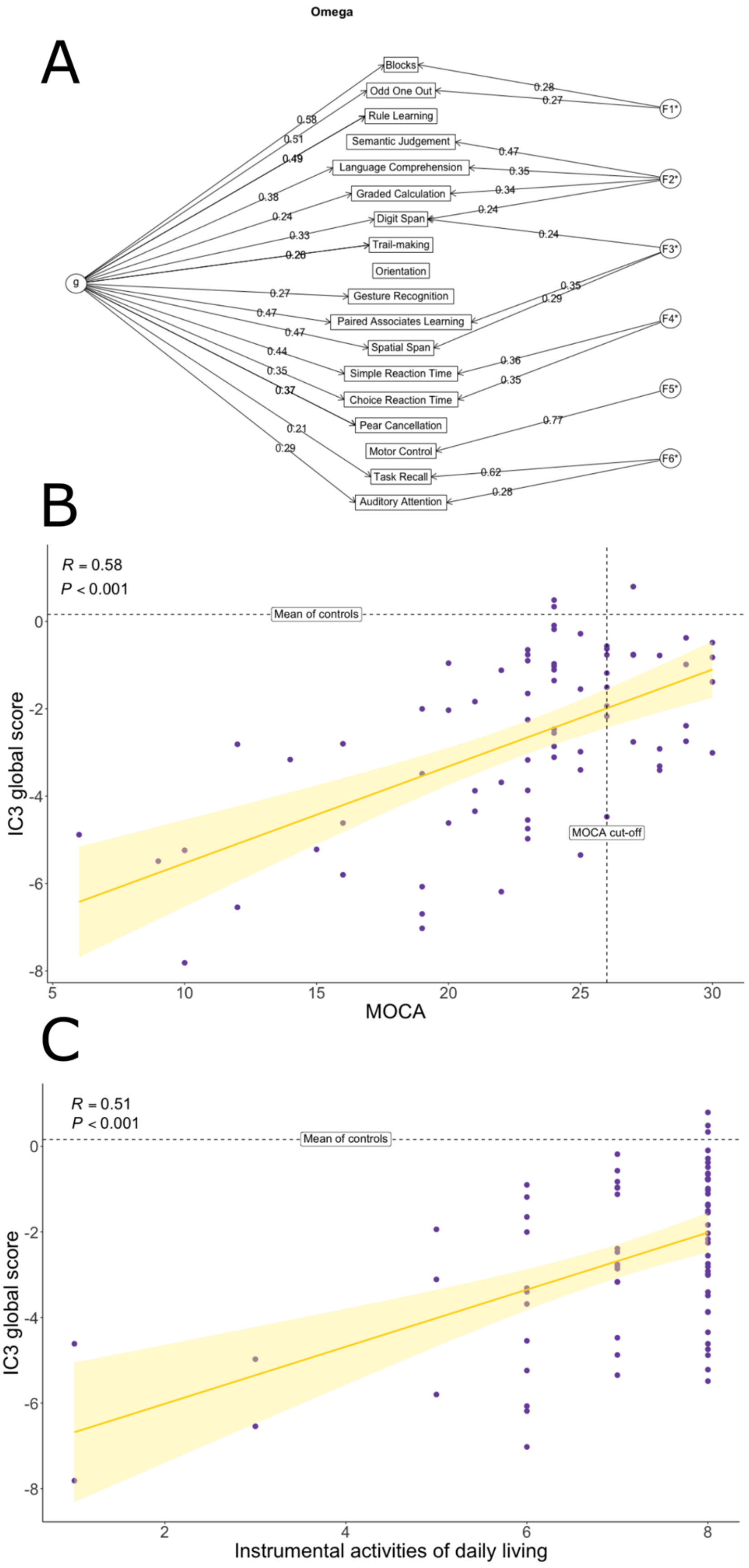

Omega
A
Blocks
Odd One Out
Rule Learning
Semantic Judgement
Language Comprehension
Graded Calculation
Digit Span
Trail-making
Orientation
Gesture Recognition
Paired Associates Learning
Spatial Span
Simple Reaction Time
Choice Reaction Time
Pear Cancellation
Motor Control
Task Recall
Auditory Attention
g
F1*
F2*
F3*
F4*
F5*
F6*
0.58
0.51
0.49
0.38
0.24
0.33
0.26
0.27
0.47
0.47
0.44
0.35
0.37
0.21
0.29
0.28
0.27
0.47
0.35
0.34
0.24
0.24
0.35
0.29
0.36
0.35
0.77
0.62
0.28
B
R = 0.58
P < 0.001
Mean of controls
MOCA cut-off
IC3 global score
0
-2
-4
-6
-8
5
10
15
20
25
30
MOCA
C
R = 0.51
P < 0.001
Mean of controls
IC3 global score
0
-2
-4
-6
-8
2
4
6
8
Instrumental activities of daily living

Figure 5A. The solution to a bifactor exploratory factor analysis on the combined dataset of patients and controls. The global cognitive measure for the IC3 is defined as the 'g' factor. The remaining 6 factors intuitively map to Executive Function (F1), Language/Numeracy (F2), Working Memory (F3), Attention (F4), Motor Ability (F5) and Memory Recall (F6). **Figure 5B**. Correlation between IC3 global score (g) and clinically validated total MOCA score in patients (N=80). Vertical dotted line: MOCA cut-off for normality. Horizontal dotted line: mean global IC3 in controls. Shaded area represents 95% Confidence Interval. **Figure 5C.** Relationship between IC3 global score (g) and quality of life metrics (IADL). Horizontal dotted line: mean global IC3 in controls. Error bar represents 95% Confidence Interval.

The IC3 online cognitive assessment is more sensitive to mild cognitive impairment compared to MOCA

IC3 showed high sensitivity at both the domain level and the task-level (as shown in high true positives in dark purple, Figure 6). Given that IC3 was specifically designed to detect mild impairment, the sensitivity of the MOCA screening tool was also assessed against the IC3 and found to be weaker (Figure 6, in yellow). See Supplementary Material 5.11. for details on how the domains were chosen.

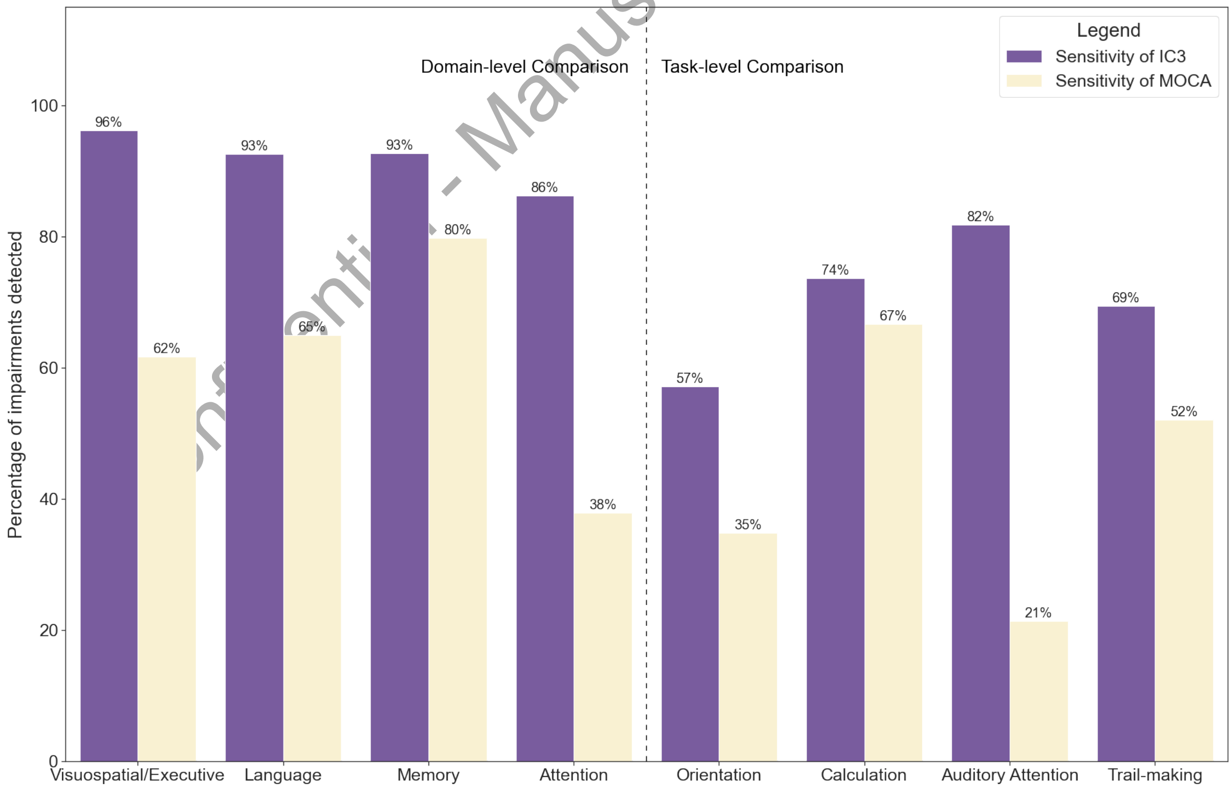


Figure 6. Sensitivity of IC3 against MOCA (purple) and vice versa (yellow) in detecting impairment at domain and task level where a clear 1:1 mapping was present.

In addition, as shown in Figure 7, IC3 was able to detect a substantial proportion of impairment in patients, even when those patients were classed as 'healthy' according to their MOCA performance (≥26/30). These impairments were detected in both the acute (dark purple) and the sub-acute/chronic stage (light purple) after stroke, highlighting IC3's ability to detect mild impairment, undetected by clinical screens, in all stages of recovery after stroke.

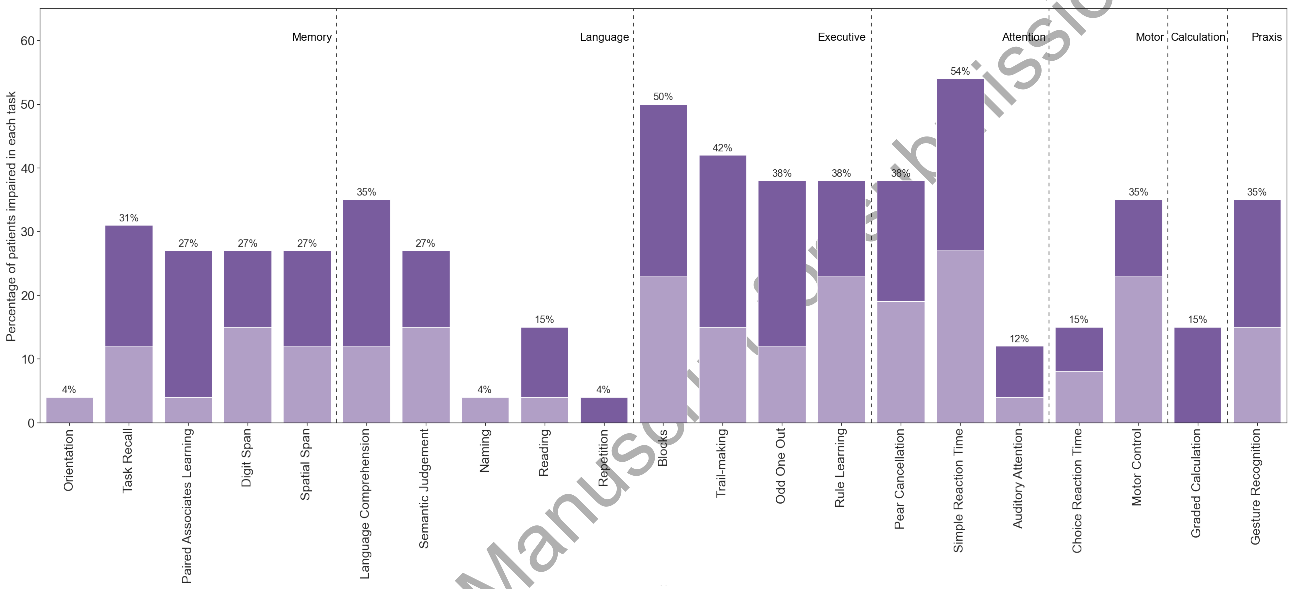


Figure 7. Percentage of patients classed as impaired on the IC3, within a sub-group of stroke patients that were deemed cognitively healthy on the MOCA (N=27). Dark purple= scores from the acute stage of the stroke. Lighter purple= sub-acute/chronic stage.

Discussion

The current study presents the IC3, a novel digital online cognitive testing tool that allows for large-scale identification and monitoring of cognitive sequalae after stroke and related vascular disorders. The robustness and reliability of the IC3 platform was extensively demonstrated and underscored through high test-retest reliability, equivalent performance between supervised and non-supervised settings, minimal learning effects across multiple timepoints and high psychometric validity. An extensive normative sample of more than 6,000 older adults from the United Kingdom, age-matched to the stroke population, was systematically gathered and leveraged to calculate patient-specific prediction and impairment scores. While accounting for a wide range of demographic and neuropsychiatric variables, IC3 was able to differentiate healthy controls and stroke survivors across all tasks on the IC3 platform, with effect sizes ranging from moderate to large (d=0.43-1.59).

Importantly, patient outcomes derived from the IC3 demonstrated strong concordance with results from clinical scores available in the patients (MOCA), with superior sensitivity to detecting mild cognitive impairments and stronger correlation with patient-reported functional impairments post-stroke (IADL). Reassuringly, the convergence validity of the IC3-derived outcomes in patients was balanced against an expected good divergent validity as shown with no discernible correlation with admission cholesterol levels. Collectively, these findings provide compelling evidence supporting the validity of remote digital testing via the IC3 platform as a valuable clinical tool for assessing and monitoring cognition following stroke and related vascular disorders.

The IC3 platform is the only cognitive tool that: 1- provides scalability and cost-effectiveness in post-stroke cognitive monitoring in both healthcare and research settings, 2- has sensitivity to mild cognitive impairments, 3- is stroke-specific, meeting the national guidelines requirements for cognitive assessment in stroke, 4- produces nuanced response metrics per individual tests (e.g., accuracy, reaction time and trial-by-trial variability), and 5- has the ability to output automated real-time patient predictive scores accounting for confounding demographic factors.[5,7,8,23]

These features afford several advantages over conventional pen-and-paper clinical testing by reducing reliance on resource-stretched healthcare professionals and promoting equity in provision of cognitive testing (e.g. physically disabled patients in whom attendance to healthcare or research setting is difficult). This, together with the gamified format, optimises engagement, and performance, making it an appealing research and screening tool for patients with stroke.

Currently available stand-alone cognitive screening tools commonly used in routine clinical care are either not stroke-specific, or not comprehensive. Commonly used tools, such as MOCA, Mini Mental State Examination, and Addenbrooke's Cognitive Examination-Revised, are tailored to detect deficits in neurodegenerative dementias and are not tailored to patients with stroke, who often have domain-specific as well as domain-general deficits.[9,24,25] Additionally, increasingly used cognitive screens designed specifically for stroke, such as the OCS, although more sensitive than MOCA, are not comprehensive enough to allow a deep cognitive phenotyping, missing the milder end of the severity spectrum.[18] The digital OCS-Plus only assesses memory and executive function and requires a trained staff to administer, thus limiting its scalability and affordability compared to the IC3.[19] Other digital platforms such as CANTAB, usually used in the setting of research into neurodegenerative/psychiatric disorders, are not stroke-specific nor able to provide patient-predictive scores.[26] The IC3 assessment tool developed and applied in this study addresses these shortcomings, building the foundation for routine detailed monitoring of cognition in healthcare setting, and for scalable large-scale population-based studies of poststroke cognitive impairment.

The evidence presented in this paper is further corroborated by recent studies from our groups, showing the feasibility of online cognitive testing for identification and monitoring of impairments in neurological disorders, such as traumatic brain injury and autoimmune limbic encephalitis.[27,28] Collectively, these studies have shown that online assessments correlate well with standard clinical evaluations and exhibit comparable, if not higher sensitivity. The current results not only align with these findings, but also provide much more extensive reliability and validity metrics, showcasing the 1) robustness of the IC3 assessment, 2) its superiority against standard clinical screens and 3) its ability to predict functional outcomes after stroke. These results provide supporting evidence for integration of such platforms in clinical care pathway, providing additional insight into clinical decisions; for instance, on when to perform, potentially costly, imaging or in-person testing during the follow-up period of the disease, and inform rehabilitation decisions.

Nevertheless, it is important to consider the current findings in the context of certain limitations. Similar to most standard pen-and-paper tests, the IC3 tasks were administered in a predetermined order. Consequently, missing data were more likely to pertain to tasks towards the latter part of the battery potentially leading to underrepresenting participants with higher level of impairments for these tasks. However, it is worth noting that this phenomenon is inherent in all cognitive assessment methodologies, and the inability to complete the assessment can itself be considered a meaningful outcome measure. In the current study, only patients who had completed at least 50% of the IC3 assessment were included in the analysis, and the last task of the IC3 normative cohort has 4782 respondents surpassing the normative sample size for most conventional assessment tools.

Furthermore, since the assessment requires sustained engagement for 10-15 minutes at a time, it may not be suitable in severely impaired patients with poorer engagement and/or reduced capacity.[29] Instead, the assessment is most appropriate for patients with mild-to-moderate impairment, who have the highest potential for regaining independence and who may benefit most from personalized treatments and rehabilitation.

Employing technology to administer unsupervised cognitive assessments on a large scale could inadvertently exclude individuals with limited access to technology. Nevertheless, the IC3 platform was designed to accommodate both supervised and unsupervised settings, and our findings indicate no discernible differences in performance between the two. Therefore, to minimise such health inequity, patients may undergo IC3 assessment in-person in the clinical setting.

Due to its cost-effectiveness and scalability, IC3 can be further developed for wide adoption as a clinical diagnostic and monitoring tool for patients with stroke and related vascular disorders. This will be tested in clinical trials within the healthcare setting. Such implementation will facilitate the detection and longitudinal monitoring of cognitive impairment after stroke at minimal cost. Given its sensitivity to mild impairment, IC3 can be used as the main cognitive outcome measure in clinical research studies in patients with stroke. We are currently adopting this approach in a longitudinal observational study of post-stroke cognition alongside blood biomarkers and brain imaging to identify mechanisms of recovery following stroke.[10] It is anticipated that these findings will inform tailored, personalized rehabilitation strategies for more effective recovery, a prospect not achievable with current assessment methodologies.

Acknowledgements. We are grateful to all the patients and participants for their time and involvement in the study. We thank Sabia Combrie for her contribution patient data collection, Joseph Coghlin for his help with control data collection, Joanna Friedland, Myrto Mavrotas, Charlotte Barot, Ziyuan Cai and Georgia Gvero for manually transcribing the speech data. We are grateful to Daniel Alexandre Romano Alves for advising the design of the IC3 graphics.

Declaration of competing interests. AH is Co-director and owner of H2CD Ltd, and owner and director of Future Cognition Ltd, which support online studies and develop custom cognitive assessment software respectively. PH reports personal fees from H2CD Ltd, outside the submitted work.

Data sharing statement. Data sharing requests from researchers within academic institutions should be directed to the corresponding author.

Funding. This research is funded by the UK Medical Research Council (MR/T001402/1). Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre and the NIHR Imperial Clinical Research Facility. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. DCG is funded by Imperial College London. SB is funded by Impact Acceleration Award (PSP415–EPSRC IAA, and PSP518–MRC IAA). AH is supported by the Biomedical Research Centre at Imperial College London. V.G. is supported by the Medical Research Council (MR/W00710X/1).

References

1. Douiri A, Rudd AG, Wolfe CD. Prevalence of poststroke cognitive impairment: South London stroke register 1995–2010. Stroke. 2013;44(1):138–45.

2. Hamilton OKL, Backhouse EV, Janssen E, Jochems ACC, Maher C, Ritakari TE, et al. Cognitive impairment in sporadic cerebral small vessel disease: A systematic review and meta-analysis. Alzheimer's & Dementia. 2021;17(4):665–85.

3. Stolwyk RJ, Mihaljcic T, Wong DK, Hernandez DR, Wolff B, Rogers JM. Post-stroke Cognition is Associated with Stroke Survivor Quality of Life and Caregiver Outcomes: A Systematic Review and Meta-analysis. Neuropsychol Rev [Internet]. 2024 Mar 11 [cited 2024 Mar 16]; Available from: https://doi.org/10.1007/s11065-024-09635-5

4. Stolwyk RJ, Mihaljcic T, Wong DK, Chapman JE, Rogers JM. Poststroke Cognitive Impairment Negatively Impacts Activity and Participation Outcomes. Stroke. 2021 Feb;52(2):748–60.

5. McMahon D, Micallef C, Quinn TJ. Review of clinical practice guidelines relating to cognitive assessment in stroke. Disability and Rehabilitation. 2022 Nov 20;44(24):7632–40.

6. Hill G, Regan S, Francis R, Mead G, Thomas S, Salman RAS, et al. Research priorities to improve stroke outcomes. The Lancet Neurology. 2022 Apr 1;21(4):312–3.

7. Party ISW. National clinical guideline for Stroke for the United Kingdom and Ireland. London: Royal College of Physicians. 2023;

8. Quinn TJ, Richard E, Teuschl Y, Gattringer T, Hafdi M, O'Brien JT, et al. European Stroke Organisation and European Academy of Neurology joint guidelines on post-stroke cognitive impairment. European Stroke Journal. 2021 Sep;6(3):I–XXXVIII.

9. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. Journal of the American Geriatrics Society. 2005;53(4):695–9.

10. Gruia DC, Trender W, Hellyer P, Banerjee S, Kwan J, Zetterberg H, et al. IC3 protocol: a longitudinal observational study of cognition after stroke using novel digital health technology. BMJ Open. 2023 Nov 1;13(11):e076653.

11. Hampshire A. Great british intelligence test protocol. 2020;

12. Marin RS, Biedrzycki RC, Firinciogullari S. Reliability and validity of the apathy evaluation scale. Psychiatry Research. 1991 Aug 1;38(2):143–62.

13. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale: application to patients with multiple sclerosis and systemic lupus erythematosus. Archives of neurology. 1989;46(10):1121–3.

14. Herrmann N, Mittmann N, Silver IL, Shulman KI, Busto UA, Shear NH, et al. A validation study of The Geriatric Depression Scale short form. International Journal of Geriatric Psychiatry. 1996;11(5):457–60.

15. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. The gerontologist. 1969;9(3_Part_1):179–86.

16. Hampshire A, Azor A, Atchison C, Trender W, Hellyer PJ, Giunchiglia V, et al. Cognition and Memory after Covid-19 in a Large Community Sample. New England Journal of Medicine. 2024 Feb 29;390(9):806–18.

17. Atchison CJ, Davies B, Cooper E, Lound A, Whitaker M, Hampshire A, et al. Long-term health impacts of COVID-19 among 242,712 adults in England. Nat Commun. 2023 Oct 24;14(1):6588.

18. Demeyere N, Riddoch MJ, Slavkova ED, Bickerton WL, Humphreys GW. The Oxford Cognitive Screen (OCS): Validation of a stroke-specific short cognitive screening tool. Psychological Assessment. 2015 Sep;27(3):883–94.

19. Demeyere N, Haupt M, Webb SS, Strobel L, Milosevich ET, Moore MJ, et al. Introducing the tablet-based Oxford Cognitive Screen-Plus (OCS-Plus) as an assessment tool for subtle cognitive impairments. Sci Rep. 2021 Apr 12;11(1):8000.

20. Bickerton WL, Demeyere N, Francis D, Kumar V, Remoundou M, Balani A, et al. The BCoS cognitive profile screen: Utility and predictive value for stroke. Neuropsychology. 2015 Jul;29(4):638–48.

21. Hartshorne JK, Germine LT. When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. Psychol Sci. 2015 Apr 1;26(4):433–43.

22. Pronk T, Molenaar D, Wiers RW, Murre J. Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. Psychon Bull Rev. 2022 Feb 1;29(1):44–54.

23. Giunchiglia V, Gruia D, Lerede A, Trender W, Hellyer P, Hampshire A. Iterative decomposition of visuomotor, device and cognitive variance in large scale online

cognitive test dat [Internet]. 2023 [cited 2024 Mar 19]. Available from: https://europepmc.org/article/PPR/PPR665566

24. Kurlowicz L, Wallace M. The mini-mental state examination (MMSE). Vol. 25, Journal of gerontological nursing. SLACK Incorporated Thorofare, NJ; 1999. p. 8–9.

25. Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR. The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. International Journal of Geriatric Psychiatry. 2006;21(11):1078–85.

26. Smith PJ, Need AC, Cirulli ET, Chiba-Falek O, Attix DK. A comparison of the Cambridge Automated Neuropsychological Test Battery (CANTAB) with "traditional" neuropsychological testing instruments. Journal of Clinical and Experimental Neuropsychology. 2013 Mar 1;35(3):319–28.

27. Del Giovane M, Trender WR, Bălăeţ M, Mallas EJ, Jolly AE, Bourke NJ, et al. Computerised cognitive assessment in patients with traumatic brain injury: an observational study of feasibility and sensitivity relative to established clinical scales. EClinicalMedicine. 2023;59.

28. Shibata K, Attaallah B, Tai XY, Trender W, Hellyer PJ, Hampshire A, et al. Remote digital cognitive assessment reveals cognitive deficits related to hippocampal atrophy in autoimmune limbic encephalitis: a cross-sectional validation study. eClinicalMedicine [Internet]. 2024 Feb 2 [cited 2024 Mar 16];0(0). Available from: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(24)00016-6/fulltext

29. Bill O, Zufferey P, Faouzi M, Michel P. Severe stroke: patient profile and predictors of favorable outcome. Journal of Thrombosis and Haemostasis. 2013 Jan 1;11(1):92–9.