nature medicine

Article

https://doi.org/10.1038/s41591-024-03475-9

Smartwatch- and smartphone-based remote assessment of brain health and detection of mild cognitive impairment

Received: 26 March 2024

Accepted: 17 December 2024

Published online: 4 March 2025

Check for updates

Paul Monroe Butler[1,2,3] ✉, Jenny Yang[1], Roland Brown[2], Matt Hobbs[1,2], Andrew Becker[2], Joaquin Penalver-Andres[2], Philippe Syz[1], Sofia Muller[1], Gautier Cosne[2], Adrien Juraver[2], Han Hee Song[1], Paramita Saha-Chaudhuri[2], Daniel Roggen[2], Alf Scotland[2], Natalia Silveira[1], Gizem Demircioglu[2], Audrey Gabelle[2], Richard Hughes[2], Michael G. Erkkinen[3,4], Jessica B. Langbaum[4,5], Jennifer H. Lingler[4,6], Pamela Price[4,7], Yakeel T. Quiroz[4,8], Sharon J. Sha[4,9], Marty Sliwinski[4,10], Anton P. Porsteinsson[4,11], Rhoda Au[4,12], Matt T. Bianchi[1], Hanson Lenyoun[1], Hung Pham[1], Mithun Patel[1] & Shibeshih Belachew[2]

Consumer-grade mobile devices are used by billions worldwide. Their ubiquity provides opportunities to robustly capture everyday cognition. 'Intuition' was a remote observational study that enrolled 23,004 US adults, collecting 24 months of longitudinal multimodal data via their iPhones and Apple Watches using a custom research application that captured routine device use, self-reported health information and cognitive assessments. The study objectives were to classify mild cognitive impairment (MCI), characterize cognitive trajectories and develop tools to detect and track cognitive health at scale. The study addresses sources of bias in current cognitive health research, including limited representativeness (for example, racial/ethnic, geographic) and accuracy of cognitive measurement tools. We describe study design and provide baseline cohort characteristics. Next, we present foundational proof-of-concept MCI classification modeling results using interactive cognitive assessment data. Initial findings support the reliability and validity of remote MCI detection and the usefulness of such data in describing at-risk cognitive health trajectories in demographically diverse aging populations. ClinicalTrials.gov identifier: NCT05058950.

Brain aging in the information age is changing how physicians think about measuring patients, and how patients think about measuring their own health[1–5]. Smartphones, wearables and mobile computing platforms are integrated into daily life[6–10]. Everyday technologies potentially capture important information about behavior and cognition that, when properly contextualized, could offset current barriers to accurately measure and track meaningful change in cognition. To develop digital biomarkers of cognition and brain health, we need to understand how cognitive phenotypes might translate into digital phenotypes in the form of physiologic and behavioral signatures detected by multimodal wearable sensing technologies with high penetration in society[5,11–14].

Smart devices are utilized by billions worldwide. In the United States, 85% of adults own smartphones, including an estimated 94% of individuals age 50–64 years and 61% over 65 years[15,16]. Swift societal adoption of digital technologies over recent decades has

A full list of affiliations appears at the end of the paper. ✉e-mail: pmbutler@bwh.harvard.edu

enabled remote research and decentralized clinical trials (DCTs), which may deploy electronic-consent (e-consent), patient- and/or provider-reported outcomes via use of mobile applications, digital questionnaires, unsupervised clinical assessments, tele-medicine visits, wearable devices or other health technologies[17,18]. DCTs have the potential to improve accessibility, facilitate recruitment, support retention, and promote equity and diversity in clinical trials by empowering individuals from under-represented and underserved groups to participate in important scientific discoveries[19,20]. With these strategic advantages, DCTs can help address public health challenges in brain health. With a growing global population that is aging, screening for and detecting cognitive health decline represents an unmet need requiring public health attention[17].

Fifty-five million individuals worldwide suffer from dementia, with Alzheimer's disease (AD) and related dementias being the leading causes, and with numbers expected to triple by 2050 (refs. 21–23). Preclinical dementia can be heralded by subjective cognitive complaints/impairment/decline (SCC/SCI/SCD, herein referred to as SCC) from individuals and/or informants who are worried about an emerging decline in cognition and behavior. Aging individuals with SCC are at heightened risk for conversion to prodromal dementia—a stage clinically diagnosed as MCI[24–26]. MCI is a clinical syndrome with heterogeneous causes and is defined by patient and/or informant subjective concern for cognitive decline and by clinicians measuring objective cognitive deficits in the setting of preserved ability to perform activities of daily living[27–29]. MCI represents an at-risk group for AD and related dementias, and is often accompanied by modifiable medical and psychiatric comorbidities[30–32]. Large-scale detection of SCC, MCI and their underlying causes is a public health imperative[33–35]. Early detection of cognitive decline empowers individuals to enact lifestyle modifications and initiate pharmacologic and nonpharmacologic approaches to slow cognitive decline[36,37].

A core feature of MCI is objective, measurable cognitive deficits. Traditional neuropsychological assessments can be biased based on cultural and demographic factors[38]. In the setting of the diversity of the US population, where barriers exist that limit access to and engagement in healthcare and research, patient-facing technologies offer new opportunities to evaluate cognitive performance and real-world behaviors that might signal changes in brain health[39]. Compared with traditional testing in controlled clinical settings, digital assessment in everyday environments can provide more frequent, contemporaneous and ecologically relevant information about cognitive and functional status. Two emerging technologies may be considered to advance this field: passive and interactive[38,39]. Passive technologies are meant to continuously track behaviors without specific user input, whereas interactive approaches elicit active user engagement to gauge cognitive and behavioral performance at various intervals[40,41]. The aims of real-world technologies to monitor cognition are to supplement traditional in-clinic neuropsychological measurements, equip patients with tools to self-track their cognitive health and to enable population-based brain health screening[38].

Passive tracking with wearables has enabled digital measures of changes in sleep, motor function and behavior that might precede the earliest stages of cognitive decline as seen in clinically defined MCI[42–44]. Research by Thompson et al.[45] supports the feasibility, reliability and validity of remote digital cognitive assessments with comparisons with in-clinic digital assessment[45]. Berron et al.[46] and Nicosia et al.[47] implemented various remote digital assessments in clinical and control populations providing further support of the validity of this unsupervised interactive measurement approach[46–48]. In this context, the Intuition Brain Health study was designed to capture the largest comprehensive US-population-based cohorts with multimodal passive and interactive digital signals from aging individuals on a continuum of susceptibility to cognitive decline. The study design described allows for risk characterization and development of prognostic taxonomies via a holistic passive- and interactive-feature generation framework.

Results

Enrollment and recruitment strategy

Intuition enrollment began 20 September 2021 and, after 18 months, on 13 March 2023 achieved its overall planned enrollment goal with 23,004 US adult residents (μ (mean) = 58.0 years, σ = 15.6; range, 21–86). The flow of enrollment activity is depicted in Fig. 1. Study data sources are described in Extended Data Table 1. Baseline enrollment status was defined by those consenting participants who provided adequate demographic and health history information necessary for cohort assignment and who attempted an initial 30-min cognitive assessment battery (Cambridge Neuropsychological Test Automated Battery from Cambridge Cognition (CANTAB); Methods). There were 126,640 Study App downloads, and 60,324 individuals subsequently performed prescreening, e-consent and email confirmation. From this group, 50.7% (n = 30,613) were eligible to continue and, for those disqualified, the leading reasons for disqualification were failure to provide core demographic information required for cohort assignment, failure to complete the identity verification step or due to cohorts reaching full enrollment (Supplementary Section 1). From the eligible group, 97.5% (n = 29,858) proceeded through onboarding and orientation to study activities; 77.1% (n = 23,004) of onboarded participants achieved baseline enrollment status by completing a baseline CANTAB battery. Cohorts were defined by age and risk for cognitive decline, including controls aged 21–59 years in early and middle (EM) adulthood (Controls-EM), controls aged 60–86 years in late adulthood (L) with low- and high-risk for cognitive decline (Controls-L Low, Controls-L High), participants aged 50–86 years with SCC, individuals to self-report receiving a diagnosis of MCI aged 21–49 years in early and middle adulthood (MCI-EM) and 50–86 years (MCI), and those individuals with clinically confirmed (CC) MCI (MCI-CC) aged 50–86 years referred from clinical sites and/or medical record review (Methods). For further discussion on the enrollment flow, see Supplementary Section 1.

We used an adaptive, multichannel recruitment strategy to promote timely enrollment with the highest yield approaches coming from targeted email campaigns that contributed an estimated 33% of enrollment, and word-of-mouth referrals that contributed approximately 32%. Other notable strategic areas included App Store and study website traffic, which recruited 13%, paid advertisements in social media and internet search engines (12% yield), traditional clinical site-based referrals (7%) and from community and brain health advocacy events (3%). Word-of-mouth referrals were notably more important for recruitment in under-represented populations based on race/ethnicity, which accounted for 56% of Asian/Asian American, 39% of Black/African American, and 37% of Latino/Hispanic participants.

As enrollment proceeded, the cohorts that reached prespecified target sizes and capped were the Controls-EM ($n \geq 6{,}000$) at 7 months after study start, the SCC cohort ($n \geq 2{,}000$), the MCI-EM ($n \geq 1{,}000$) at 9 months, and Controls-L Low and High-Risk at 15-months ($n \geq 12{,}000$). Enrollment for the MCI-CC cohort began with a limited pilot group to gather learnings to streamline the process for additional sites. Additional clinical sites and new workflows were then initiated (Supplementary Fig. 3). The COVID-19 pandemic impacted clinical trial research globally with notable reductions in research activities[49,50]. Despite these limitations, with slower than expected site-based activity, we recruited 414 MCI-CC participants and 79.0% (n = 327) completed baseline enrollment.

From the baseline enrollees, 98.6% (N = 22,676) proceeded to order the study-provisioned Apple Watch, and 83.5% (N = 18,934) paired the Apple Watch to an iPhone and sent study data from both devices. Demographics of this Apple-Watch-activated subpopulation are provided in Supplementary Table 4 and differed nominally (<1.0%) in demographic characteristics from the baseline enrolled.

Baseline study population characteristics

A total of 23,004 participants completed baseline enrollment. Demographics, medical history and risk factor profile are described in Table 1

https://doi.org/10.1038/s41591-024-03475-9

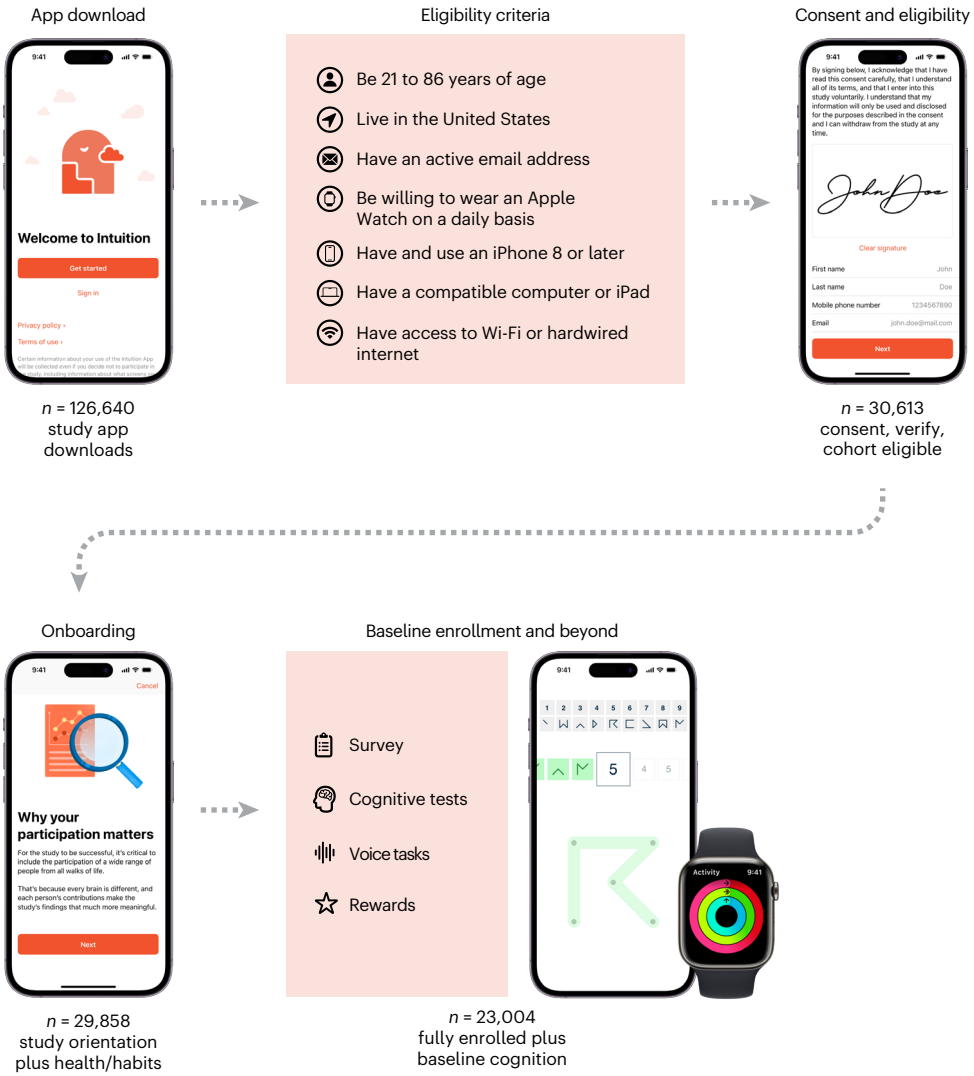


Fig. 1 | Intuition study enrollment flow. The steps from Study App download to complete baseline enrollment are shown stepwise from left to right across two rows, including screening and enrollment totals by stage along the bottom of the diagram. Subjects were required to already own an iPhone to be used in the study and an Apple Watch was provisioned once participants completed baseline enrollment including a 30-min cognitive assessment battery.

and Supplementary Table 3. Baseline enrollees were from all 50 US States and resembled the relative geographic diversity of state-based US population densities (Extended Data Fig. 1 and Supplementary Table 2). The study population was 64.4% female and 31.5% reported racial and ethnic diversity, including 9.9% Asian; 8.1% Black/African American; 7.7% Hispanic/Latino; and 5.8% as multi-racial, Native American or Pacific Islander. In terms of education and income, 34.1% reported less than a bachelor's degree, and 22.1% with annual household incomes below US$50,000. At a cohort level, there was greater proportional enrollment of individuals with educational attainment less than a bachelor's degree in the SCC and MCI-EM and pooled MCI age 50–86 years (that is, MCI and MCI-CC), with 44.6.%, 56.7% and 41.2.%, respectively.

The prevalences of modifiable and nonmodifiable dementia risk factors are reported in Table 1. In comparing pooled MCI with pooled aging Controls-L (Low- and High-Risk) cardiovascular and metabolic illnesses were similar in prevalence, but participants with pooled MCI had higher rates of traumatic brain injury (TBI, 30.9 versus 16.7%), mental health disorders (32.7 versus 18.1%), active smoking (4.5 versus 2.2%), and were more likely to be male (44.8 versus 34.6%). Controls-EM compared with Controls-L (Low-/High-risk) reported higher prevalence of mental health disorders (23.2 versus 15.4/21.0%). Controls-L (Low/High) compared with Controls-EM reported higher prevalence of age-associated cardiovascular and metabolic-related illness (for example, hypertension, 25.7/71.3 versus 16.8%) and hearing impairment (14.8/27.9 versus 3.3%). First-degree family history of dementia was more prevalent in Controls-L (Low/High) versus Controls-EM (20.2/58.3 versus 8.8%).

Although age-associated cardiometabolic illnesses (for example, cardiac disease) were more prevalent in pooled Controls-L and pooled MCI compared with Controls-EM, the likelihood of co-occurring risk factors differed. In Extended Data Fig. 2, we provide visualizations that capture pairwise odds ratios between individual risk factors in Controls-EM, pooled Controls-L and in participants with pooled MCI. The results show intensified associations between selected risk factors that are both expected and unexpected in relation to known epidemiologic patterns. For example, in Controls-EM and pooled Controls-L, expected associations between cardiometabolic risk factors were strong for obesity, type II diabetes, hypertension and hyperlipidemia. Although these expected associations were also present in the pooled MCI group, elevated pairwise odds ratios cross-linked modifiable and nonmodifiable risk factors, such as the association of substance use to heart disease and its sequelae. Whereas risk factors were prevalent in all groups, participants with pooled MCI versus Controls-L were more likely to report three or more comorbid conditions.

https://doi.org/10.1038/s41591-024-03475-9

Table 1 | Participant baseline demographics and self-reported health history

	Total	Control EM	Control-L Low	Control-L High	SCC	MCI-EM	MCI	MCI-CC
Demographics, *n*	23,004*	6,329	6,427	6,078	2,154	1,133	556	327
Age, years (σ)	58.0 (15.6)	39.6 (11.3)	67.5 (5.4)	68.5 (5.5)	64.2 (7.5)	35.2 (8.3)	65.2 (8.2)	66.8 (8.7)
Sex, female (%)	14,655 (64.4%)	4,135 (65.4%)	4,394 (69.5%)	3,607 (60.8%)	1,311 (61.1%)	722 (63.8%)	306 (55.0%)	180 (55.0%)
White	17,063 (75.0%)	3,934 (62.2%)	5,310 (84.0%)	5,082 (85.7%)	1,544 (71.9%)	513 (45.4%)	395 (71.0%)	285 (87.2%)
Asian	2,242 (9.9%)	1,100 (17.4%)	449 (7.1%)	224 (3.8%)	224 (10.4%)	185 (16.4%)	51 (9.2%)	9 (2.8%)
Black/African American	1,845 (8.1%)	668 (10.6%)	279 (4.4%)	380 (6.4%)	181 (8.4%)	262 (23.2%)	55 (9.9%)	20 (6.1%)
Hispanic or Latino	1,744 (7.7%)	738 (11.7%)	255 (4.0%)	283 (4.8%)	204 (9.5%)	206 (18.2%)	45 (8.1%)	13 (4.0%)
Multiracial or other	1,326 (5.8%)	537 (8.5%)	223 (3.5%)	215 (3.6%)	165 (7.7%)	132 (11.7%)	45 (8.1%)	9 (2.8%)
Marital status: married	14,375 (63.2%)	3,448 (54.5%)	4,466 (70.6%)	4,061 (68.5%)	1,393 (64.9%)	450 (39.8%)	331 (59.5%)	226 (69.1%)
Education: <Bachelor's	7,748 (34.1%)	2,079 (32.9%)	1,731 (27.4%)	1,984 (33.5%)	957 (44.6%)	641 (56.7%)	229 (41.2%)	127 (38.8%)
Annual income: <US$50,000 (σ)	5,032 (22.1)	1,620 (25.6%)	945 (14.9%)	1,092 (18.4%)	580 (27.0%)	511 (45.2%)	204 (36.7%)	80 (24.5%)
Medical history and risk factors								
Dyslipidemia (%)	9,351 (41.1%)	993 (15.7%)	2,373 (37.5%)	4,114 (69.4%)	1,150 (53.6%)	249 (22.0%)	295 (53.1%)	177 (54.1%)
Hypertension	8,641 (38.0%)	1,064 (16.8%)	1,623 (25.7%)	4,226 (71.3%)	1,045(48.7%)	298 (26.3%)	238 (42.8%)	147 (45.0%)
Obesity, BMI ≥ 30	6,971 (30.7%)	1,945 (30.7%)	810 (12.8%)	2,807 (47.3%)	762 (35.5%)	406 (35.9)	148 (26.6%)	93 (28.4%)
Family history, dementia	6,562 (28.9%)	555 (8.8%)	1,280 (20.2%)	3,454 (58.3%)	812 (37.8%)	140 12.4%)	207 (37.2%)	114 (34.9%)
Traumatic brain injury	4,181 (18.4%)	1,099 (17.4%)	822 (13.0%)	1,219 (20.6%)	449 (20.9%)	319 (28.2%)	176 (31.7%)	97 (29.7%)
Hearing impairment	3,579 (15.7%)	212 (3.4%)	936 (14.8%)	1,654 (27.9%)	473 (22.0%)	75 (6.6%)	134 (24.1%)	95 (29.1%)
Type 2 diabetes	2,389 (10.5%)	274 (4.3%)	96 (1.5%)	1,426 (24.1%)	381 (17.7%)	73 (6.5%)	94 (16.9%)	45 (13.8%)
Cardiac disease	1,992 (8.8%)	101 (1.6%)	215 (3.4%)	1,219 (20.6%)	295 (13.7%)	38 (3.4%)	76 (13.7)	48 (14.7%)
Tobacco, active use	862 (3.8%)	323 (5.1%)	41 (0.6%)	225 (3.8%)	114 (5.3%)	119 (10.5%)	30 (5.4%)	10 (31%)
Alcohol, heavy use	636 (2.8%)	66 (1.0%)	120 (1.9%)	310 (5.2%)	98 (4.6%)	8 (0.7%)	22 (4.0%)	12 (3.7%)
Mental health history								
Mental illness, any history	4,903 (21.6%)	1,471 (23.2%)	972 (15.4%)	1,248 (21.0%)	597 (27.8%)	326 (28.8%)	167 (30.0%)	122 (37.3%)
Depression, PHQ-2 (σ)	0.79 (1.18)	1.00 (1.25)	0.39 (0.82)	0.45 (0.87)	1.33 (1.33)	2.09 (1.52)	1.51(1.49)	1.09 (1.42)
Anxiety, GAD-2	0.85 (1.21)	1.17 (1.36)	0.45 (0.77)	0.47 (0.81)	1.31 (1.33)	2.17 (1.72)	1.44 (1.48)	1.10 (1.39)

Health history data comes from Study App custom self-report questions about medical and social history. BMI, body mass index.

Risk factor and illness profiles differed between MCI-EM and pooled MCI. In terms of cognitive and neuropsychiatric symptom burden, MCI-EM registered the highest SCC scores compared with all other cohorts based on mean cognitive function instrument (CFI) 14-item (CFI-14), everyday cognition scale (E-Cog) 12-item (E-Cog-12), Patient Health Questionnaire 2-item (PHQ-2) and Generalized Anxiety Disorder 2-item (GAD-2) scores (Table 1). Example differences in CFI-14-item-level responses included MCI-EM versus pooled MCI reporting 'difficulty concentrating' (63.4 versus 46.7%) and 'mood changes' (50.8 versus 32.0%) as the initial neurocognitive symptoms, whereas participants with pooled MCI reported 'word-finding difficulty' with higher frequency compared with MCI-EM (56.7 versus 39.2%).

In terms of the diagnostic journey, 57.0% of pooled MCI were diagnosed by a neurologist or neuropsychologist compared with MCI-EM, where 53.5% reported primary care physicians or psychiatrists making the diagnosis. MCI were more likely than MCI-EM to undergo extensive cognitive testing (62.7 versus 33.8%) and neuroimaging (50.5 versus 32.6%).

Adherence

Device data and cognitive assessments. Cumulative study adherence (Fig. 2 and Supplementary Table 7) was evaluated in all participants who completed baseline CANTAB assessment, activated an Apple Watch and who were enrolled for at least 12 months ($n = 17,583$). This represented 92.9% of $n = 18,934$ participants who activated an Apple Watch. Passive device use adherence was defined by those consenting participants who were (1) actively using their iPhone and sharing device data (for example, Sensor Kit/Health Kit) through the custom Intuition Study App, and (2) providing at least 4 h of daily Apple Watch wear data. CANTAB assessment adherence was defined by completion of a monthly battery. Quarterly assessment adherence was defined by those participants completing at least seven 'burst' sessions (~2 min) across the 14-day window. Adherence patterns by race/ethnicity are reported in Supplementary Fig. 4.

Tele-research adherence. Tele-research sessions were offered to $N = 1,943$ participants (Supplementary Section 2e), including all self-reported MCI ($N = 1,689$), new-onset MCI ($N = 116$) reported on biannual Study App surveys, a sampling of Control participants with three successive low CANTAB performances ($N = 41$) and a sample of MCI-CC ($N = 97$) referred from clinical sites. Overall tele-research adherence was 52.2%, with $N = 1,015$ interviews performed that included tele-Montreal Cognitive Assessment (MoCA). Adherence by group was 50.0% ($N = 844$) for self-reported MCI at baseline, 63.8% ($N = 71$ of 116 eligible) were for new-onset MCI reported on biannual Study App surveys, 63.4% ($N = 26$) for low CANTAB and 73.2% ($N = 71$) for MCI-CC. MCI-CC were sampled as an internal validity check on the process for MCI-label-adjudication. The expert tele-research clinicians were blinded to clinical status and in 93.0% (66 of 71) of cases there was concordance between the tele-research MCI-label-adjudication and clinical site-based diagnosis. In the remaining 7% ($N = 5$) of cases the label was adjudicated as 'possible MCI.' Across all tele-research sessions,

https://doi.org/10.1038/s41591-024-03475-9

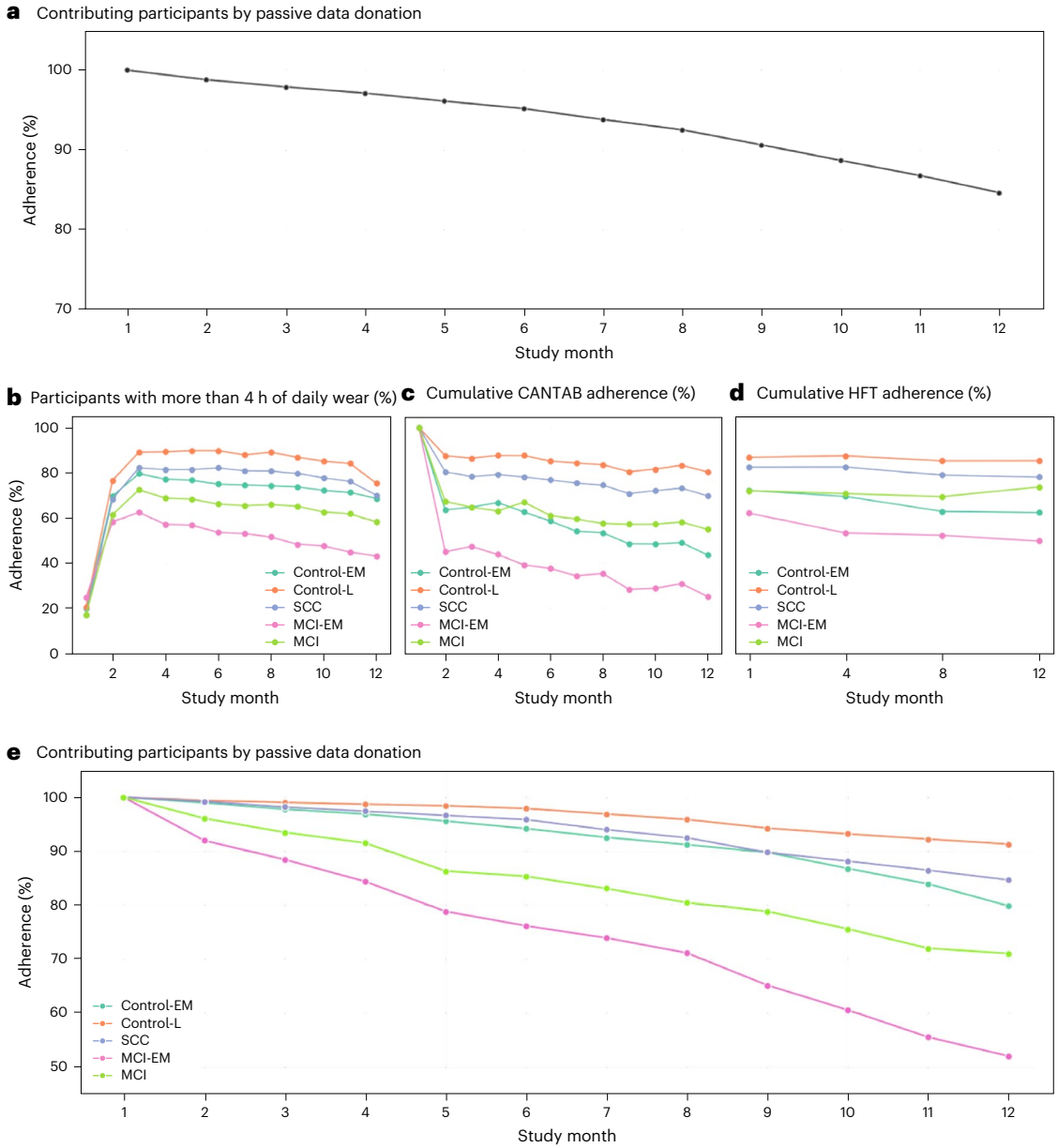


Fig. 2 | Twelve-month study adherence for device use and cognitive assessments. a, Percentage of contributing participants in the study, defined by those providing iPhone passive data sharing. **b**–**e**, Percent adherence by cohort is plotted for Apple Watch (**b**), monthly CANTAB assessment (**c**), quarterly HFT assessment (**d**) and overall passive data adherence by group (**e**). MCI, pooled MCI self-report and clinically confirmed cases. Contributing participants were defined by participants with screen unlock data from the iPhone over a 12-month period. Sample sizes were $n = 17{,}583$ (**a**), $n = 5{,}552$ Control-EM, $n = 9{,}245$ Control-L (low/high risk), $n = 1{,}544$ SCC, $n = 935$ MCI-EM and $n = 307$ MCI (**b**–**e**). Participants considered adherent to the HFT was calculated as those who, over the course of the 2-week burst period, had 7 days with at least one assessment completed. iPhone device use was defined by sharing Sensor Kit data evidence by App-Usage data sharing. Apple Watch adherence was defined by passive device data for at least 4 h per day. Numerical values for adherence by group are listed in Supplementary Table 7.

$N = 22$ (2.2%) of encounters were identified as definitely invalid either on the tele-MoCA or interview due to various circumstances, such as audiovisual and technical problems or language issues.

Baseline cognition

Feasibility. In terms of the feasibility to perform unsupervised monthly CANTAB and quarterly burst cognitive assessments, participants reported on whether the session was 'distracted' for any reason. For CANTAB, there were $N = 281{,}879$ total monthly sessions completed, and 18.9% ($n = 53{,}363$) marked as distracted at the end of the session. By age and cognitive status, Control participants under 50 years reported 23.7% of CANTAB sessions as distracted compared with 18.1% of Control participants over 50 years, 17.8% of SCC, 15.2% of MCI-EM and 14.4% of pooled MCI. For high-frequency burst Cambridge Cognition (Cam-Cog) quarterly assessments: there were 3,209,054 total burst sessions completed, and 16.3% ($n = 524{,}618$) marked as distracted. By age and cognitive status, Control participants under 50 years reported 21.1% of Cam-Cog burst sessions as distracted compared with 15.8% of Control participants over 50 years, 15.0% of SCC, 13.4% of MCI-EM and 11.8% of pooled MCI.

Clinical validity and reliability. Across all cohorts, we evaluated CANTAB test–retest reliability in a sample population that completed at least ten monthly sessions. Depending on the CANTAB outcome of

https://doi.org/10.1038/s41591-024-03475-9

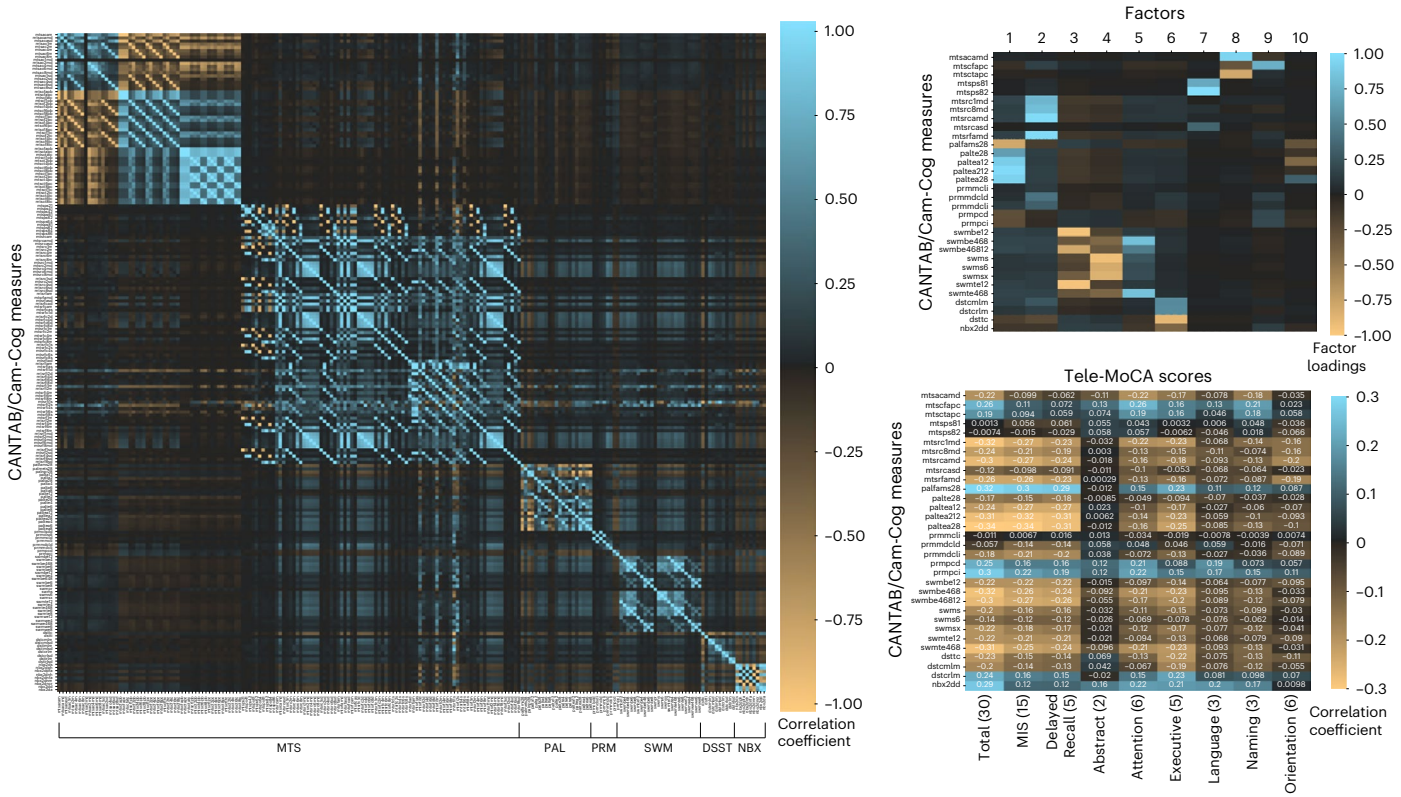


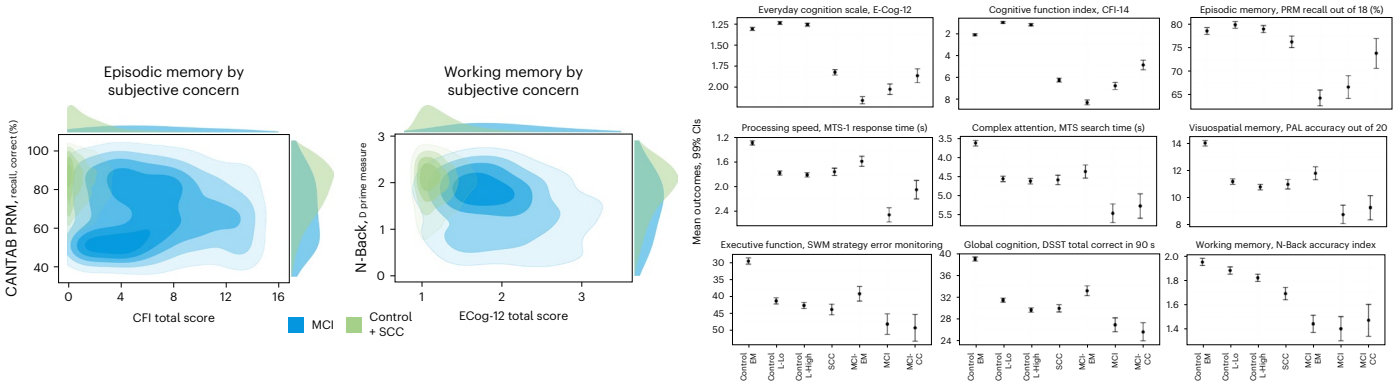
Fig. 3 | Construct validity of Cam-Cog assessments. Left, a 207-factor Pearson correlation matrix based on monthly CANTAB and quarterly high-frequency Cam-Cog variables in $N = 21,574$ baseline participants to complete assessments. Top right, exploratory factor analysis of the key cognitive outcomes for those factors with Kaiser criteria eigenvalues greater than 1.0. Bottom right, CANTAB/Cam-Cog correlations with tele-MoCA scores in participants who underwent tele-research assessment. NBX = N(2)-Back; Tele-MoCA was administered to $N = 1,015$ participants with concern for cognitive decline with v.8.1, v.8.2 or v.8.3 in the setting of a clinical research interview to evaluate cognitive health. Tele-MoCA scores are reported including the total global score out of 30 points and MoCA-defined subscores as listed. Memory impairment score (MIS) was calculated using established MoCA guidelines and a 15-point score was derived from spontaneous recall, cued recall and multiple-choice cued recall.

interest, intraclass correlation coefficients and Pearson correlations were moderate-to-high and ranged from 0.50 to 0.80 (Supplementary Fig. 6). For key CANTAB outcomes, all Pearson correlations increased across ten monthly sessions from 0.64 to 0.72 or paired associates learning (PAL) total errors adjusted, from 0.60 to 0.66 for spatial working memory (SWM) between errors, from 0.44 to 0.61 for pattern recognition memory (PRM) percent correct on delayed recall and from 0.74 to 0.80 on match-to-sample (MTS) median reaction times to correct trials. To understand the validity of performance results obtained from unsupervised cognitive assessments, we evaluated cognitive outcomes by age, education and cognitive status to identify expected patterns consistent with that reported in the neuropsychology literature from controlled testing environments[51]. Extended Data Fig. 3 plots key task-specific cognitive outcomes by age in the pooled Controls population. Expected age-associated trends were observed across eight selected outcomes (six CANTAB/two Cam-Cog) that reflected different cognitive domains such as for visuospatial memory (Extended Data Fig. 3a), executive function (Extended Data Fig. 3b), learning and immediate recall (Extended Data Fig. 3c), episodic memory (Extended Data Fig. 3d), processing speed (Extended Data Fig. 3e), complex attention (Extended Data Fig. 3f), working memory (Extended Data Fig. 3h) and global cognition (Extended Data Fig. 3g). Baseline cognitive performance was also stratified by educational attainment with expected performance distributions (Supplementary Fig. 3).

Construct validity. To evaluate the construct validity of the objective cognitive measurement approach using baseline CANTAB and Cam-Cog burst assessments, we performed a Pearson correlation matrix analysis of all outcome variables ($N = 207$ variables) for those completing CANTAB plus one session of high-frequency Cam-Cog burst testing ($N = 21,574$). Results are shown in Fig. 3. In addition, we performed an exploratory factor analysis on a focused set of key CANTAB/Cam-Cog variables that were selected based on the precedent in the literature and expert clinical judgment. Using Kaiser criterion, we identified ten main factors, which were assessed for correlation with measures from the tele-MoCA in a subset of participants ($N = 881$) to complete research tele-visits triggered by self-report of cognitive impairment (Fig. 3 and Supplementary Fig. 7). The Pearson correlation matrix demonstrated stronger within- versus between-measure relationships, suggesting the CANTAB/Cam-Cog tests were probing different cognitive constructs. Correlations between exemplar CANTAB/Cam-Cog outcomes with tele-MoCA total scores demonstrated mild-to-moderate strength correlations ($r > 0.3$ s) with PAL total errors adjusted, SWM total errors, MTS median response times and immediate learning on the PRM. Tele-MoCA memory-impairment score and delayed recall subscores correlated most with PAL total errors adjusted across all stage trials ($r = 0.31–0.34$). Attention subscores on the tele-MoCA correlated with MTS accuracy on first attempt ($r = 0.26$), immediate learning on the PRM ($r = 0.22$) and the N-Back discrimination index ($r = 0.22$). Executive function subscores on the tele-MoCA correlated with PAL first attempt memory scores ($r = 0.23$), digit symbol substitution test total correct ($r = 0.23$) and the N-Back discrimination index ($r = 0.21$).

Group validity. Known-groups validity for unsupervised baseline subjective and objective cognitive outcomes was assessed by cohorts and listed in Fig. 4 (top). Directionally valid statistical differences in baseline

https://doi.org/10.1038/s41591-024-03475-9



	Total	Control EM	Control L-low risk	Control L-high risk	SCC	MCI-EM	MCI	MCI-CC
Subjective concerns raised with clinicians (%)	2,666 (11.7)	371 (5.9)	182 (2.9)	270 (4.6)	272 (12.6)	914 (80.8)	453 (81.5)	204 (63.6)
E-Cog-12 (σ)	1.39 (0.47)	1.31 (0.41)	1.24 (0.25)	1.26 (0.26)	1.82 (0.49)	2.15 (0.71)	2.02 (0.68)	1.86 (0.72)
CFI-14	2.47 (3.00)	2.17 (2.73)	1.04 (1.05)	1.25 (1.10)	6.30 (2.24)	8.35 (3.41)	6.83 (3.93)	4.92 (3.81)
Episodic memory, PRM recalled out of 18 (%)	77.7 (15.5)	78.5 (16.5)	79.8 (13.7)	78.9 (13.6)	76.2 (15.7)	64.4 (17.8)	66.7 (17.1)	73.8 (14.5)
Processing speed, MTS response time in seconds	1.66 (0.77)	1.29 (0.56)	1.78 (0.58)	1.81 (0.56)	1.76 (0.66)	1.59 (1.09)	2.46 (2.50)	2.05 (0.76)
Complex attention, MTS search time in seconds	4.32 (1.59)	3.62 (1.19)	4.55 (1.36)	4.61 (1.35)	4.58 (1.55)	4.36 (2.11)	5.45 (4.16)	5.26 (2.41)
Visuospatial memory, PAL initial accuracy, out of 20	11.8 (4.5)	14.0 (4.0)	11.2 (4.1)	10.8 (4.1)	11.0 (4.4)	11.8 (5.1)	8.8 (5.8)	9.3 (4.2)
Executive function, SWM strategy error monitoring	38.7 (19.9)	29.6 (19.7)	41.3 (18.8)	42.6 (18.5)	43.8 (18.5)	39.2 (20.1)	48.1 (17.6)	49.2 (19.6)
Global cognition, DSST total correct in 90-sec	32.8 (8.9)	38.9 (9.6)	31.4 (7.0)	29.6 (6.7)	29.9 (7.4)	33.1 (9.1)	26.9 (8.4)	25.6 (7.4)
Working memory, N-back accuracy index	1.84 (0.64)	1.95 (0.68)	1.88 (0.57)	1.82 (0.57)	1.69 (0.64)	1.44 (0.78)	1.40 (0.74)	1.47 (0.72)

Baseline cognition	Control-L low versus high	SCC versus MCI	SCC versus MCI-CC	SCC versus control-L	MCI versus MCI-CC	MCI versus control-L	MCI-CC versus Control-L	MCI-EM versus control-EM
E-Cog-12 (*P* value)	4.39 (0.0001)	7.87 (0.0001)	1.28 (0.20)	82.09 (0.0001)	3.30 (0.001)	63.01 (0.0001)	40.01 (0.0001)	55.63 (0.0001)
CFI-14 (*P* value)	10.93 (0.0001)	4.17 (0.0001)	9.29 (0.0001)	168.63 (0.0001)	7.05 (0.0001)	98.99 (0.0001)	55.23 (0.0001)	67.37 (0.0001)
Episodic memory, *PRM* recalled out of 18 (%) (*P* value)	3.69 (0.0002)	12.48 (0.0001)	2.60 (0.009)	9.79 (0.0001)	6.29 (0.0001)	21.14 (0.0001)	7.29 (0.0001)	26.17 (0.0001)
Processing speed, MTS1 response time in seconds (*P* value)	2.94 (0.003)	11.54 (0.0001)	7.25 (0.0001)	2.20 (0.028)	2.89 (0.004)	20.36 (0.0001)	8.06 (0.0001)	13.92 (0.0001)
Complex attention, MTSa search time in seconds (*P* value)	2.48 (0.013)	7.83 (0.0001)	6.79 (0.0001)	0.30 (0.76)	0.76 (0.45)	12.68 (0.0001)	8.69 (0.0001)	16.75 (0.0001)
Visuospatial memory, PAL initial accuracy, out of 20 (*P* value)	5.46 (0.0001)	9.80 (0.0001)	6.55 (0.0001)	1.03 (0.30)	1.36 (0.17)	12.13 (0.0001)	7.40 (0.0001)	16.29 (0.0001)
Executive function, SWM strategy error monitoring (*P* value)	3.90 (0.0001)	4.94 (0.0001)	4.88 (0.0001)	4.36 (0.0001)	0.86 (0.39)	7.67(0.0001)	6.96 (0.0001)	15.06 (0.0001)
Global cognition, DSST total correct in 90 s (*P* value)	14.68 (0.0001)	8.28 (0.0001)	9.79 (0.0001)	3.73 (0.0002)	2.32 (0.02)	12.08(0.0001)	12.83 (0.0001)	18.88 (0.0001)
Working memory, N-back accuracy index (*P* value)	5.89 (0.0001)	9.21 (0.0001)	5.69 (0.0001)	11.81 (0.0001)	1.37 (0.17)	17.96 (0.0001)	11.81 (0.0001)	22.71 (0.0001)

Fig. 4 | Baseline subjective and objective cognition by cohort. Top left, density plots of MCI versus Controls plus SCC when considering the two dimensions of subjective cognitive concerns and objective cognitive performance. Bottom left, table listing baseline cognition based on self-report and key CANTAB/Cam-Cog outcomes. Top right, group means are plotted with 99% CIs by outcome and with statistical comparisons of baseline cognition by cohort, including two-sided pairwise *t*-test comparisons of baseline subjective and objective cognition and associated *P* values. MTS1, MTS 1-box reaction time; MTSa, MTS 8-box search time. Bottom right, two-sided paired *t*-test comparisons by cohorts of interest. Analysis of variance test for each cognitive outcome listed by cohort was significant with *P* values $< 1 \times 10^{-20}$. Left, all participants aged 50 years and above across cohorts ($N = 17{,}042$) and all study participants ($N = 23{,}004$). Right, all participants aged 50 years and above across cohorts ($N = 17{,}042$).

cognition were observed consistently between specific study cohorts on all selected cognitive outcomes (Fig. 4 (bottom)). For pooled MCI versus age-match SCC and Controls, we plotted two-dimensional score-based separation with subjective and objective cognition scores (Fig. 4; upper panel). Taken together, combined population-level measurement distributions of subjective cognitive concerns (CFI-14 or E-Cog-12) and objective cognitive performance (CANTAB/Cam-Cog) better separated pooled MCI versus age-matched Controls + SCC compared with considering either dimension in isolation.

Initial proof-of-concept MCI classification model. Core demographics, baseline subjective (CFI/E-Cog) and objective (CANTAB) cognition variables were evaluated for the development of a logistic regression MCI classifier model. We focused on those MCI-CC and MCI cohort participants where the clinical MCI label was validated by site-based referral, medical record review or adjudicated by tele-research visit. Figure 5 lists the demographics and health history of MCI aged 50–86 years ($N = 556$) and Control + SCC group for modeling ($N = 16{,}234$). Model details, receiver operating characteristic (ROC) curve, confusion matrix and classification accuracy calculated by the area under the curve (AUROC) (mean AUROC = 0.85 ± 0.04, 95% confidence interval (CI) plus crossvalidation) are depicted in Fig. 5. A rank ordered list of the top 40 predictors by model coefficients is reported in Supplementary Table 8, which included a mix of subjective ($N = 18$) and objective ($N = 19$) baseline cognitive predictors and demographic factors ($N = 3$). Model performance metrics were: 80.2% sensitivity (95% CI, 72.8–87.6%), 78.7% specificity (95% CI, 74.3–83.1%), 79.1% accuracy (95% CI, 75.3–82.8%), 55.6% positive predictive value (95% CI, 47.9–63.3%) and 92.3% negative predictive value (95% CI, 89.1–95.4%), given a threshold for a use case to balance sensitivity and specificity, which included a true positive rate of 0.80 and false positive rate of 0.21. In considering US aging population prevalence rates of MCI, then adjusting MCI prevalence rates from 25% (for example, 3:1 majority-to-minority class sampling) to 22% increases the negative predictive value to 93.3% and reduces the positive predictive value to 51.8% (ref. 52). To supplement the initial proof-of-concept MCI classification model and to better understand the classification accuracy of the core components of the model, we trained logistical regression MCI classification models on core demographic data only, baseline CANTAB cognitive performance and subjective cognitive survey scores (for example, CFI/E-Cog) and display reports in Extended Data Figure 9.

Examples of interactive and passive study data. With passive digital sensing and interactive high-frequency cognitive assessments, the Intuition study aims to classify mild cognitive impairment and at-risk trajectories and characterize potential digital phenotypes of cognition. Sample visualizations are provided in Extended Data Fig. 5 with longitudinal multimodal cognitive data from demographically matched study participants with and without cognitive impairment. Candidate features are selected for heuristic purposes from typical daily device use, such as 'taps per minute' from iPhone typing. From this emerging rich multimodal dataset, the goal will be to evaluate for mathematical relationships between measurements of interactive cognitive assessments and from those candidate features derived from passive sensing with the study devices (for further examples, see Extended Data Fig. 5).

https://doi.org/10.1038/s41591-024-03475-9

	Controls + SCC	MCI*
Demographics, *n*	16,234	556
Age, years (*σ*)	65.7 (6.90)	67.1 (8.35)
Sex, female (%)	10,662 (65.7%)	309 (55.6%)
White	12,476 (76.9%)	432 (77.7%)
Asian	1,043 (6.4%)	25 (4.5%)
Black/African American	999 (6.2%)	35 (6.3%)
Hispanic or Latino	874 (5.4%)	32 (5.8%)
Multi-racial or other	842 (5.2%)	32 (5.8%)
Marital status: Married	11,093 (68.3%)	375 (67.4%)
Education: <Bachelor's	5,242 (32.3%)	184 (43.1%)
Annual income: <US$50,000 (*σ*)	2,908 (17.9%)	139 (25.0%)
Dyslipidemia (%)	8,233 (50.7%)	340 (61.2%)
Hypertension	7,494 (46.2%)	273 (49.1%)
Obesity, BMI ≥ 30	4,771 (29.4%)	154 (27.7%)
Family history, dementia	5,946 (36.6%)	243 (43.7%)
Traumatic brain injury	2,131 (13.1%)	122 (21.9%)
Hearing impairment	2,030 (12.5%)	137 (24.6%)
Type 2 diabetes	2,071 (12.8%)	99 (17.8%)
Cardiac disease	2,103 (13.0%)	138 (24.8%)
Tobacco, active use	634 (3.9%)	29 (5.2%)
Alcohol, heavy use	560 (3.5%)	26 (4.7%)
Mental illness, any history	1,557 (9.6%)	156 (28.1%)
Depression, PHQ-2 (*σ*)	0.55 (0.98)	1.23 (1.48)
Anxiety, GAD-2	0.58 (0.94)	1.18 (1.38)

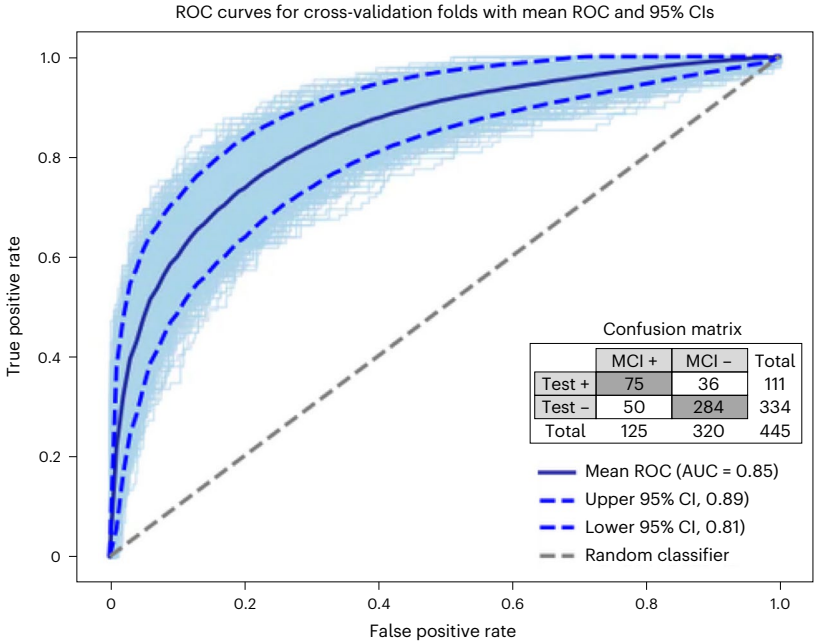


Fig. 5 | Initial MCI classification model results using baseline cognition. MCI* is the combined MCI group comprised of MCI-CC and self-reported MCI confirmed by tele-health versus Controls aged 50–86 years with and without subjective cognitive complaints. Left, baseline characteristics; right, logistic regression MCI classifier accuracy results with ROC curve. All controls aged 50 years and above were included alongside participants with CFI-defined SCC. MCI cases were clinically confirmed cases and self-reported MCI as confirmed by a tele-research visit, including a tele-MoCA to confirm impairment. The MCI classifier is a logistic regression model with ridge penalization (L2 regularization). The model incorporates all baseline CANTAB outcomes (objective cognitive performance measures), along with two subjective cognition surveys: CFI and E-Cog. The model also uses core demographic variables including age, sex and education level. The data were split into 80% for training and 20% for testing. To address class imbalance between the majority and minority classes, training data was resampled using a three-to-one majority-to-minority class ratio. The model was trained using 100× bootstrap resampling in the outer loop to enhance generalization and estimate model stability. Within each bootstrap iteration, a grid search was employed in the inner loop to systematically explore a range of hyperparameters, specifically the regularization strength for ridge penalization, and identify the best-performing hyperparameter configuration. To further ensure robust evaluation, the inner loop applied stratified fivefold crossvalidation, which maintained class balance within each fold while testing different hyperparameter sets. This nested crossvalidation setup ensured that hyperparameter tuning of the model was independent of the outer loop resampling, minimizing the risk of overfitting and optimizing performance on unseen data. Supplementary Table 8 rank orders the most important predictor variables by beta-coefficient values.

Discussion

The Intuition brain health study was a large, virtual observational study in 23,004 US adults using direct-to-consumer App-based interactive and passive data collection from an iPhone and Apple Watch. We describe the study design, deployment and baseline study population, and a provide initial proof-of-concept modeling results to support the validity of remote MCI classification. These initial results support the feasibility, acceptability and validity of the measurement approach. In 18 months, the study recruited and enrolled a demographically diverse population with a focus on middle-to-late adulthood and a spectrum of risk for decline in cognition. With a DCT framework approach, two-thirds of the study enrolled using two key strategies: focused email campaigns based on diverse demography; and word-of-mouth, which was increasingly important for recruitment of under-represented populations by race/ethnicity. With 5 min to collect demographics and subjective cognitive health data in the Study App and 30 min to perform baseline CANTAB on a personal computing device, we demonstrate moderate-to-high MCI classification accuracy in a population of $N = 16{,}790$ aging adults as proof-of-concept to support the validity of a remote, cognitive screening approach.

To develop digital biomarkers of cognition, and thereby advance precision medicine, robust multimodal datasets (that is, 'big data') are needed to train and test models that can generalize and be ecologically valid when applied to the greater populations intended to screen, track or treat[20,53]. Consider that only 5% of US adults have ever participated in traditional clinical research trials and have disproportionately been males in early-to-middle adulthood[54]. By contrast, consider that age-related neurodegenerative diseases, such as AD, are more prevalent in females in late adulthood, individuals with lower educational attainment and socioeconomic status, and in certain racial and ethnic minoritized groups, including Hispanic/Latino and Black/African American populations[55–57]. We report on key demographics that speak to how decentralized research studies can open new avenues that invite diversity and promote health equity. Intuition enrolled 64.5% female and 31.5% of the total study population reported being from under-represented races or ethnicities, and that proportion increased to 43.3% in the combined SCC/MCI cohorts most susceptible to decline. In addition, socioeconomic and geographic enrollment barriers were lower for our study as evidenced by the 22.1% reporting annual household incomes below US$50,000 and all 50 US states represented in close relative proportion to the known distribution of regional population densities (Table 1 and Extended Data Fig. 1). Although these numbers are higher than what is typically seen in clinical research and trials, there is still greater diversity needed.

Recruiting and retaining participants are separate challenges for clinical research[58]. The attrition rate for Intuition was 10.5%, which is substantially lower than the pooled dropout rate of 49% (95% CI, 27–70) reported in systematic reviews of other observational digital App-based trials, albeit these numbers are difficult to interpret given notable heterogeneity in study size and duration[59]. In terms of passive and active adherence, data collection was robust across the first 12 months, and was marked by a group-level pattern of those aging Controls-L and SCC participants displaying the highest device use and active cognitive assessment adherence, while the Controls-EM and all MCI cohorts

exhibited lower adherence. Compared with cumulative adherence, segmental and longitudinal adherence remained high across groups. Stated differently, once a device was activated (for example, Apple Watch) or interactive cognitive assessments started (for example, CANTAB) then participants were more likely to be adherent longitudinally. With the Study App and passive device use running in the background of participants' daily lives, managing the technologies still required executive function and prospective memory to remember to charge, wear and use the devices.

In under-represented populations, adherence was slightly lower across participants of color but remained strong for passive and active cognitive data capture. The American Academy of Clinical Neuropsychology predicts that, by the year 2050, 60% of the US population will be 'untestable' with our current outdated armamentarium of cognitive assessments, which are mostly monolingual and monocultural[48]. It is encouraging to see that a more gamified, language-agnostic, smartphone-based strategy might help to address these growing concerns for brain health screening and opens new venues for adapting and validating these assessments in other languages.

Self-reported cognitive decline in the aging population is common but lacks specificity[24–26]. Our baseline findings support the acceptability, reliability and validity of remote subjective and objective interactive cognitive assessments. Typical age- and education-related differences in performance observed in controls across the adult lifespan were expected and in line with previous work on cognitive aging[60]. Test–retest reliability and performance consistency reinforces these initial results (Supplementary Fig. 6)[61]. In terms of clinical and construct validity, we can appreciate that the highest at-risk cohorts, such as SCC and MCI, displayed different cognitive performance distributions. SCC participants exhibit significant differences from Controls-L and intermediate in statistical difference in group means compared with matched MCI in terms of learning/recall, working memory, attention, and executive function. The initial proof-of-concept MCI classifier models suggest that remote MCI detection can be valid at scale. We sought to use a clinically verified cohort of MCI to first establish the validity of interactive cognitive assessment for classification. Next, we plan to explore the classification value of ecological momentary assessments and passive measurements of cognition from everyday behaviors. Exploration into the everyday digital signatures of cognitive phenotypes using multimodal mobile sensing remains a frontier of new knowledge with the potential to inform our evolving understanding of aging, cognition and early detection of neurodegeneration.

Although there are definite strengths of the Intuition study, such as the decentralized approach that enabled facile enrollment of large sample sizes including under-represented populations, there are caveats and limitations that should be acknowledged. Clinical history was limited in reliability and accuracy given the self-report approach to collect medical history in the Study App, and by the variable longitudinal engagement and adherence. We mitigated these risks in part by having recruited a core patient base of clinically confirmed MCI and by adjudicating the confidence in the self-reported MCI label with tele-research consultations (Methods). Self-report of medical and cognitive status confounds the accuracy of risk stratification given there are likely misestimations in prevalence of medical diagnoses. English-fluency and iPhone usage eligibility criteria were also constraining and might limit the generalizability of future results. Finally, the incentivization strategy with the Apple Watch 'earn it to own it' is a departure from traditional compensation strategies in research trials and brings learnings about motivated human behavior, but also comes with caveats about comparative adherence with other decentralized studies. In terms of the initial proof-of-concept classification models, interactive rather than passive cognitive data were used for the purpose of remote MCI detection. We acknowledge this speaks to the validity of the assessment approach, whereas the capacity for passive data to detect impairment remains to be determined and will be the focus of future work.

In conclusion, we provided promising early results from this large, US population-based observational brain health study that achieved target enrollment with decentralized research methods. Initial results support the feasibility, acceptability and validity to measure real-world cognition and expeditiously enroll representative groups and reach historically diverse populations that are disproportionately impacted by the illnesses that clinicians aim to prevent and treat. Our early findings are the first step in a larger research initiative aimed to equip and empower patients and clinicians with mobile tools for prevention and early detection of prevalent neurodegenerative and neuropsychiatric disorders. Remote, multimodal and frequent cognitive sampling with consumer-grade digital devices has the potential to offer scalable solutions for ecologically valid screening and for tracking cognitive health by way of meeting society where it is at in this age of information.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03475-9.

References

1. Alfalahi, H. et al. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. *Sci. Rep.* **12**, 7690 (2022).
2. Stroud, C. B., Davila, J. & Moyer, A. The relationship between stress and depression in first onsets versus recurrences: a meta-analytic review. *J. Abnorm. Psychol.* **117**, 206–213 (2008).
3. Yang, Y. et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nat. Med.* **28**, 2207–2215 (2022).
4. Adib, F., Mao, H., Kabelac, Z., Katabi, D. & Miller, R. C. Smart homes that monitor breathing and heart rate. In *Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems* (eds Begole, B & Kim, J.) 837–846 (Association for Computing Machinery, 2015).
5. Baumeister, H. & Montag, C. *Digital Phenotyping and Mobile Sensing* (Springer International Publishing, 2019).
6. Weiser, M. The computer for the 21st century. *Sci. Am.* **265**, 94–105 (1991).
7. Poslad, S. *Ubiquitous Computing Smart Devices, Smart Environments and Smart Interaction* (Wiley, 2009).
8. Macleod, E. The history of the smartphone. *Mobile Industry Review* www.mobileindustryreview.com/2016/10/the-history-of-the-smartphone.html (2016).
9. Hartanto, A. & Yang, H. Is the smartphone a smart choice? The effect of smartphone separation on executive functions. *Computers Hum. Behav.* **64**, 329–336 (2016).
10. Sarwar, M. & Soomro, T. R. Impact of smartphones on society. *Eur. J. Sci. Res.* **98**, 216–226 (2013).
11. Montag, C., Elhai, J. D. & Dagum, P. On blurry boundaries when defining digital biomarkers: how much biology needs to be in a digital biomarker? *Front. Psychiatry* **12**, 740292 (2021).
12. Dagum, P. Digital biomarkers of cognitive function. *NPJ Digital Med.* **1**, 10 (2018).
13. Montag, C. & Elhai, J. D. Digital phenotyping—a case for cognitive functions and dementia? *Digital Psychol.* **1**, 44–51 (2020).
14. Piau, A., Wild, K., Mattek, N. & Kaye, J. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild Alzheimer disease and implications for clinical care: systematic review. *J. Med. Internet Res.* **21**, e12785 (2019).

15. Mobile fact sheet. *Pew Research Center* www.pewresearch.org/internet/fact-sheet/mobile/#:~:text=The%20vast%20majority%20of%20Americans,smartphone%20ownership%20conducted%20in%202011 (2023).
16. Miller, D. et al. *The Global Smartphone: Beyond a Youth Technology* (UCL, 2021).
17. Draft guidance on decentralized clinical trials. Conducting clinical trials with decentralized elements; guidance for industry, investigators, and other interested parties. *United States Food and Drug Administration* www.fda.gov/media/167696/download (2024).
18. Goodson, N. et al. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. *NPJ Digital Med.* **5**, 58 (2022).
19. Tan, T. et al. Digital approaches to enhancing community engagement in clinical trials. *NPJ Digital Med.* **5**, 37 (2022).
20. Kelsey, M. D. et al. Inclusion and diversity in clinical trials: actionable steps to drive lasting change. *Contemp. Clin. Trials* **116**, 106740 (2022).
21. *Global status report on the public health response to dementia* (World Health Organization, 2021); www.who.int/publications/i/item/9789240033245
22. Gustavsson, A. et al. Global estimates on the number of persons across the Alzheimer's disease continuum. *Alzheimer's Dement.* **19**, 658–670 (2023).
23. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **15**, 321–387 (2019).
24. Jessen, F. et al. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's Dement.* **10**, 844–852 (2014).
25. Molinuevo, J. L. et al. Implementation of subjective cognitive decline criteria in research studies. *Alzheimer's Dement.* **13**, 296–311 (2017).
26. Jessen, F. et al. The characterisation of subjective cognitive decline. *Lancet Neurol.* **19**, 271–278 (2020).
27. Petersen, R. C. Mild cognitive impairment as a diagnostic entity. *J. Intern. Med.* **256**, 183–194 (2004).
28. Petersen, R. C. et al. Mild cognitive impairment: a concept in evolution. *J. Intern. Med.* **275**, 214–228 (2014).
29. Petersen, R. C. et al. Practice guideline update summary: mild cognitive impairment. Report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. *Neurology* **90**, 126–135 (2018).
30. Sachdev, P. S. et al. Classifying neurocognitive disorders: the DSM-5 approach. *Nat. Rev. Neurol.* **10**, 634–642 (2014).
31. Albert, M. S. et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Focus* **11**, 96–106 (2013).
32. McKhann, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7**, 263–269 (2011).
33. van der Flier, W. M., de Vugt, M. E., Smets, E. M., Blom, M. & Teunissen, C. E. Towards a future where Alzheimer's disease pathology is stopped before the onset of dementia. *Nat. Aging* **3**, 494–505 (2023).
34. Sabbagh, M. N. et al. Rationale for early diagnosis of mild cognitive impairment (MCI) supported by emerging digital technologies. *J. Prev. Alzheimer's Dis.* **7**, 158–164 (2020).
35. Sabbagh, M. N. et al. Early detection of mild cognitive impairment (MCI) in primary care. *J. Prev. Alzheimer's Dis.* **7**, 165–170 (2020).
36. Livingston, G. et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet* **396**, 413–446 (2020).
37. Casaletto, K. et al. Late-life physical activity relates to brain tissue synaptic integrity markers in older adults. *Alzheimer's Dement.* **18**, 2023–2035 (2022).
38. Rabin, L. A., Brodale, D. L., Elbulok-Charcape, M. M. & Barr, W. B. in *Clinical Cultural Neuroscience: an Integrative Approach to Cross-Cultural Neuropsychology* (ed. Pedraza, O.) 55–80 (Oxford Univ. Press, 2020).
39. Öhman, F., Hassenstab, J., Berron, D., Schöll, M. & Papp, K. V. Current advances in digital cognitive assessment for preclinical Alzheimer's disease. *Alzheimer's Dement. (Amst.)* **13**, e12217 (2021).
40. Papp, K. V. et al. Unsupervised mobile cognitive testing for use in preclinical Alzheimer's disease. *Alzheimer's Dement. (Amst.)* **13**, e12243 (2021).
41. Possin, K. L. et al. The brain health assessment for detecting and diagnosing neurocognitive disorders. *J. Am. Geriatrics Soc.* **66**, 150–156 (2018).
42. Teh, S. K., Rawtaer, I. & Tan, H. P. Predictive accuracy of digital biomarker technologies for detection of mild cognitive impairment and pre-frailty amongst older adults: a systematic review and meta-analysis. *IEEE J. Biomed. Health Inform.* **26**, 3638–3648 (2022).
43. Zhou, H. et al. Digital biomarkers of cognitive frailty: the value of detailed gait assessment beyond gait speed. *Gerontology* **68**, 224–233 (2022).
44. Chen, R. et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Teredesai, A. & Kumar, V.) 2145–2155 (Association for Computing Machinery, 2019).
45. Thompson, L. I. et al. Remote and in-clinic digital cognitive screening tools outperform the MoCA to distinguish cerebral amyloid status among cognitively healthy older adults. *Alzheimer's Dement. (Amst.)* **15**, e12500 (2023).
46. Berron, D. et al. A remote digital memory composite to detect cognitive impairment in memory clinic samples in unsupervised settings using mobile devices. *NPJ Digital Med.* **7**, 79 (2024).
47. Nicosia, J. et al. Unsupervised high-frequency smartphone-based cognitive assessments are reliable, valid, and feasible in older adults at risk for Alzheimer's disease. *J. Int. Neuropsychol. Soc.* **29**, 459–471 (2023).
48. Relevance 2050 Initiative. *American Academy of Clinical Neuropsychology* https://theaacn.org/relevance-2050/relevance-2050-initiative/#gsc.tab=0 (2024).
49. Wiese, L. K., Williams, I. C., Schoenberg, N. E., Galvin, J. E. & Lingler, J. Overcoming the COVID-19 pandemic for dementia research: engaging rural, older, racially and ethnically diverse church attendees in remote recruitment, intervention and assessment. *Gerontol. Geriatr. Med.* **7**, 23337214211058919 (2021).
50. Shaaban, C. E., Lin, H. H. S., Ren, D. & Lingler, J. H. Impact of the COVID-19 pandemic on enrollment at United States Alzheimer's Disease Research Centers. *Alzheimer's Dement.* **19**, e077849 (2023).
51. Franzen, M. D. *Reliability and Validity in Neuropsychological Assessment* (Springer Science & Business Media, 2013).
52. Manly, J. J. et al. Estimating the prevalence of dementia and mild cognitive impairment in the US: the 2016 health and retirement study harmonized cognitive assessment protocol project. *JAMA Neurol.* **79**, 1242–1249 (2022).
53. Berisha, V. et al. Digital medicine and the curse of dimensionality. *NPJ Digital Med.* **4**, 153 (2021).

54. Jiang, S. & Hong, Y. A. Clinical trial participation in America: The roles of eHealth engagement and patient–provider communication. *Digital Health* **7**, 20552076211067658 (2021).
55. Grill, J. D., Sperling, R. A. & Raman, R. What should the goals be for diverse recruitment in Alzheimer clinical trials. *JAMA Neurol.* **79**, 1097–1098 (2022).
56. Wilkins, C. H., Schindler, S. E. & Morris, J. C. Addressing health disparities among minority populations: why clinical trial recruitment is not enough. *JAMA Neurol.* **77**, 1063–1064 (2020).
57. Franzen, S. et al. Diversity in Alzheimer's disease drug trials: the importance of eligibility criteria. *Alzheimer's Dement.* **18**, 810–823 (2022).
58. Frampton, G. K., Shepherd, J., Pickett, K., Griffiths, G. & Wyatt, J. C. Digital tools for the recruitment and retention of participants in randomised controlled trials: a systematic map. *Trials* **21**, 478 (2020).
59. Meyerowitz-Katz, G. et al. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. *J. Med. Internet Res.* **22**, e20283 (2020).
60. Daffner, K. R. Promoting successful cognitive aging: a comprehensive review. *J. Alzheimers Dis.* **19**, 1101–1122 (2010).
61. Erkkinen, M. G. et al. Reliability of unsupervised digital cognitive assessment in a large adult population across the aging lifespan from INTUITION: a brain health study (S8. 009). *Neurology* https://doi.org/10.1212/WNL.0000000000203227 (2023).

[1]Apple Inc., Cupertino, CA, USA. [2]Biogen Inc., Cambridge, MA, USA. [3]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4]Intuition Study Scientific Committee, Boston, MA, USA. [5]Banner Alzheimer's Institute, Phoenix, AZ, USA. [6]University of Pittsburgh School of Nursing, Pittsburgh, PA, USA. [7]The Balm in Gilead Inc., Richmond, VA, USA. [8]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [9]Stanford School of Medicine, Palo Alto, CA, USA. [10]Penn State University, University Park, PA, USA. [11]University of Rochester School of Medicine and Dentistry, Rochester, NY, USA. [12]School of Medicine, Boston University Chobanian and Avedisian, Boston, MA, USA. ✉e-mail: pmbutler@bwh.harvard.edu

Methods

Overview

The Intuition brain health study (NCT05058950) was a prospective, observational and decentralized study in which all activities were mediated via a custom research iPhone application (Study App) and there were no in-person visits. The study was sponsored by Biogen, Inc., in collaboration with Apple, Inc. Scientific leadership for the design, analyses and communication of study results were guided by an external Scientific Committee, which was comprised of clinicians, researchers, and experts in ethics and technology and patient advocacy. The Scientific Committee members were engaged to provide independent scientific input and guidance for the duration of the study.

Participants aged 21 to 86 years residing in the United States were recruited using digital and online recruitment strategies directed by healthcare practitioners, researchers and the joint Apple–Biogen study team. Individuals were guided to the study website (http://IntuitionStudy.com) where they viewed the Institutional Review Board (IRB)-approved study information, and could elect to participate by providing e-consent through the Intuition Study App on the iPhone (Fig. 1). General eligibility required prospective participants to have an iPhone version 8 or newer running the latest version of iOS and to be willing to wear a study-provisioned Apple Watch (Supplementary Fig. 1). As part of the compensation for participation, the participants could later own the Apple Watch by completing study tasks. Three categories of participants were recruited into seven cohorts based on age and cognitive status (Supplementary Fig. 2), including individuals presumed to be cognitively intact (Controls), individuals with prominent subjective cognitive complaints (SCC) and those self-reporting or known to have a medical diagnosis of mild cognitive impairment (MCI). The largest category were the three Control cohorts ($n \approx 18{,}000+$) including at least 6,000 in Early and Middle Adulthood (Controls-EM) aged 21–59 years and 12,000 in the Late Adulthood (Controls-L) divided into those at low- or high- risk for cognitive decline based on prespecified risk factor criteria (Controls-L Low- and Controls-L High-Risk; Supplementary Table 1). The second category of participants were those with concern for new decline in cognitive function (SCC cohort) compared with 1 year before study enrollment as defined by prespecified threshold score on a validated baseline screening questionnaire CFI-14-item (total score ≥ 4) and by the age of 50–86 years ($n \approx 2{,}000$). The third category included three cohorts of MCI ($n \approx 2{,}000$) divided into those in early and middle adulthood (MCI-EM) aged 21–49 years who self-reported receiving a diagnosis of cognitive impairment, or late adulthood aged 50–86 years who either self-reported an MCI diagnosis (MCI) or who were referred/identified by clinical sites/medical record review with documented and clinically confirmed MCI status (MCI-CC). For further details on cohorts and eligibility see Supplementary Section 1. While MCI was developed as a clinically diagnosed syndrome intended to identify those at risk for progression to dementia, we took a wider view and included those at risk of any cognitive impairment at any age or cause in adulthood. The purpose was to better understand varieties of typical and atypical cognitive health issues in real-world heterogenous populations with medical, psychiatric and neurological comorbidities.

Participants were recruited regardless of whether they were pre-existing Apple Watch users, and all cohorts were eligible and encouraged to order, pair and wear the study-provisioned Apple Watch after completing baseline enrollment. The digital engagement strategy deployed in the Study App was designed to fit alongside individuals' daily activities and required, on average, about an hour per month to carry out study-related active tasks. Participants were asked to complete surveys and questionnaires on health and habits, perform cognitive assessments (for example, 'memory and thinking activities') both in and out of the Study App, and were provided with educational content intended to raise brain health awareness (Extended Data Figs. 7 and 8).

Participants were informed of the overarching aims of the study to help researchers investigate the role an Apple Watch and iPhone could play in measuring changes in thinking and memory, and by studying changes in brain health over time, which may occur normally as people age or could be an early indicator of certain forms of dementia, such as AD. With a decentralized framework, we surmised that the study addressed important challenges with the current paradigm of conducting clinical studies via direct recruitment and engagement in brain health research, and by facilitating more patient-oriented approaches integrated into everyday life, which encouraged a broader diverse audience and democratized participation. The new study design advanced and aligned with the objectives of the 21st Century Cures Act according to the US Food and Drug Administration, which seeks to promote medical advances, evolve the traditional model of trial design and probe the value of real-world data to improve brain health outcomes. With the Intuition brain health study DCT framework, burdens to participate were minimized with data collection and tracking seamlessly integrated in the devices of everyday life (for example, mobile apps and wearables). The core objectives of the Intuition study were to classify MCI using multimodal passive and interactive data collected from the Apple Watch and iPhone and to characterize cognitive trajectories in individuals at risk for longitudinal decline.

Study design

Intuition enrollment and study flow consisted of four stages described below and depicted in Fig. 1. The planned study observation window was 24 months per subject based on the time from enrollment. The Intuition brain health study opened enrollment on 20 September 2021 and, following 24 months of data collection, the study was closed earlier than planned on 20 September 2023. Due to changing program priorities at Biogen, the study was discontinued early. With at least 12 months of data from 82.3% ($n = 18{,}934$) of participants, including 92.9% of which provided at least 4 h of daily Apple Watch wear data, then the study has collected adequate data to approach the key objectives outlined in the study design and statistical analysis plan.

Recruitment stage. We worked with vendors, commercial, research and academic entities in each of these principal strategic areas to facilitate broad recruitment:

(1) Email campaigns—both broad and demographic focused approaches;
(2) Word-of-mouth study referrals;
(3) Web search and social media advertisements;
(4) Community health and advocacy events;
(5) Intuition Study website and Apple App store traffic;
(6) Referrals from study sites and identified by diagnostic codes from research, medical and/or claims databases (for example, MCI-CC).

Strategies were adaptive, for example, as targeted email campaigns recruited demographically defined cohorts, we shifted email recruitment approaches toward those demographic characteristics required for cohorts that remained open for enrollment as other cohorts filled completely. The study website contained a variety of IRB-approved materials about the study, eligibility criteria and expectations for study participation, and was updated with key information as the study proceeded. For additional information about recruitment approaches, including details about clinically confirmed MCI, see Supplementary Section 1.

Screening, e-consent and eligibility stage. Interested individuals were directed to the Apple App Store to download the Intuition Study App and initiate screening. The core initial screening eligibility criteria (Supplementary Fig. 1) were: age 21–86 years, primary resident of the United States for the study duration, educational attainment of eighth grade

or higher, fluent in spoken and written English, use of an active iPhone version 8 (released in Fall 2017) or newer running the latest iOS, access to Wi-Fi or hardwired internet with a desktop computing device (Mac or Windows) or iPad, willingness to wear an Apple Watch and have an active email address for use in study communications. After demonstrating initial eligibility potential participants advanced through by providing IRB-approved e-consent, including an explanation and overview of data that was to be collected and confirmed their understanding of the study and acknowledged their willingness to participate. Email contact information was confirmed, and identity verified. Next, participants provided self-reported responses to questions in the Study App that evaluated cognitive and risk factor status to determine potential cohort eligibility. Participants were presented with the ability to share the relevant Health Kit and Sensor Kit data streams, and to receive study notifications, before moving on to complete the Onboarding Stage.

Onboarding stage. Participants were oriented to the Study App, including the tasks, points and rewards, and profile sections. Next, an onboarding 'new user experience' ensued, which formally welcomed individuals, provided an overall study overview and explained reasons why participation and contribution to research were important. A curriculum timeline for study activities was presented, including baseline surveys on health and habits, baseline cognitive assessment (that is, CANTAB 30-min computerized battery), Watch ordering and features, and educational material on cognition and brain health. To encourage Apple Watch wear, introductions were provided for to how to pair the Watch once received and participants were incentivized to engage in a weekly 'Stand challenge' and set up sleep tracking. In addition to monthly CANTAB batteries, participants were asked every 3 months to perform 2 weeks of high-frequency 'burst' cognitive testing in the Study App on the iPhone. For these bursts, an introduction, assessment tutorial, scheduling and practice opportunities were offered.

Baseline enrollment stage and beyond. After completion of the onboarding stage, including the new user experience, participants advanced to baseline enrollment status by attempting the out-of-App CANTAB computerized cognitive battery on a personal computing device (for example, iPad, laptop or desktop computer). Completion of CANTAB triggered the Apple Watch provisioning and shipping process. Participants paired the Apple Watch with the iPhone and started sharing Watch-based passive study data and began to earn points for completing activities such as the weekly Stand challenges and sleep tracking. We deployed points for study task completion strategically to drive engagement over the course of the study and to serve as a mechanism for participant compensation. Completing baseline enrollment provided subjective and objective cognitive health data to define phenotypes for longitudinal comparison. Cycles of interactive data collection ensued (Extended Data Fig. 8), and passive data capture occurred in the background of typical daily device use. All participants accumulated and received points for their time participating in study-related activities and reached the 'Watch Goal' by completing approximately 40% of routine study activities. The 'Watch Goal' was meant to provide an opportunity to the participants to keep the Apple Watch after study completion or if they decided to withdraw voluntarily after reaching the study goal. Points could have been redeemed in an ongoing fashion through the Study App for monetary rewards, up to US$280 of possible compensation with high adherence to study tasks.

Data sources and measurement approach

Interactive cognitive measurements. Extended Data Table 1 provides an overview and description of the type of data captured with descriptions of source, activity and cadence of sampling. These cycles persi sted over the study observation window (Extended Data Fig. 8). With the DCT approach there were no traditional brick-and-mortar site-based requirements. All study data were collected digitally and included Study App engagement information and participant-entered self-report data related to demographics; health, lifestyle and habits; global and mental health and cognitive health. The overall interactive cognitive measurement approach (Supplementary Section 2) included six main areas, including:

(1) SCC surveys: in-app biannually (CFI-14, E-Cog-12);
(2) Monthly CANTAB: out-of-app CANTAB 30-min computerized battery;
(3) Quarterly Cam-Cog burst: in-app high-frequency testing for 2 weeks, three times daily;
(4) Quarterly language: in-app 5-min custom battery with recorded voice;
(5) Tele-research: out-of-app, event-based triggered by prespecified criteria, tele-visit to evaluate cognitive health status, medical comorbidities and to perform a tele-MoCA;
(6) Context of cognition: in-app baseline, quarterly and biannual surveys.

To complement the subjective cognitive assessments, participants completed computerized neuropsychological tests monthly and quarterly. Monthly assessments included five tests from the CANTAB—a tool with over 30 years of application in neuropsychological and clinical research[62]. The five tests were the PRMi, PRMd, PAL, SWM and MTS tasks. These tests were chosen specifically for their broad coverage of key cognitive domains, including visual short-term episodic, recognition and working memory, as well as processing speed, complex attention and executive functioning. All tests selected were based on unique visuospatial stimuli agnostic of language and or culturally mediated effects. Unique stimuli (for example, parallel forms) were used in each assessment to obviate overt practice effects. Previous research indicates that CANTAB tests demonstrate orthogonality among outcome measures, suggesting they may capture distinct aspects of cognition[63]. In addition, correlational analyses between CANTAB and traditional paper-and-pencil neuropsychological assessments have identified moderate relationships and some overlapping cognitive domain structures, such as PAL/PRM outcomes with episodic learning/memory[64]. Moreover, the selected tests have clinical validity for measuring progressive changes associated with human aging[65,66], exhibiting sensitivity to the early identification and trajectories of cognitive decline in both cross-sectional and longitudinal studies[66,67]. Specifically, the five CANTAB tests selected (PRMi, PRMd, PAL, SWM and MTS) have demonstrated clinical validity and performance differences across several neurodegenerative and neuropsychiatric disorders with moderate-to-high effect sizes, including MCI, AD and related dementias[67–69], Parkinson's disease[70,71] and clinical depression[72]. Different therapeutic areas benefit from unique combinations of CANTAB tasks, with the five selected for our study chosen based on their utility in stratifying age-associated cognitive changes from MCI and AD trajectories[73].

The monthly assessments were deployed through web browsers on participant's preferred personal computing device, which has been shown to be ecologically valid for remote cognitive assessment and equivalent to in-clinic, supervised environments[74,75]. Furthermore, the tasks have multiple parallel/alternate forms to reduce learning effects and incorporate stepwise difficulty levels to mitigate floor and ceiling effects, particularly among clinical cohorts, thereby enhancing the sensitivity and specificity of the cognitive assessments. Once launched in the web browser, the CANTAB battery uses an English recorded voice for instructions, but the contents of the tests are all nonverbal and have been validated in non-English speaking populations, such as Spanish[76,77].

In the quarterly Cambridge Cognition (Cam-Cog) burst assessment, participants were invited to complete an N-Back task (that is, 2-Back) in addition to the Digit Symbol Substitution Test (DSST) on their personal iPhones. These tests were captured in a high-frequency paradigm, allowing users to provide a snapshot of their cognition up to three times per day, across 2 weeks each quarter (see Extended Data

Fig. 8 for further information). Emerging research has demonstrated the utility in smartphone testing for better approximating a user's cognitive function, enhancing the ecological validity as well as the clinical utility in performance features such as the learning curve and intra-individual variability in cohort stratification[47,78]. Further information on each cognitive task is given in the following paragraphs.

The PRM task assesses visual learning and recognition memory across two phases. The immediate phase begins with sequential presentation of 18 abstract nonsemantic images to learn. The task proceeds to a forced-choice recognition test where users must select between a previous pattern and a visually similar foil. Participants then perform the delayed recognition task at the end of the full assessment battery. Total percent correct is a key outcome measure for each task phase.

PAL assesses visual-spatial learning and episodic memory, requiring participants to recall the location of an abstract nonsemantic image. The task increases in difficulty from 2-box patterns to 4-, 6-, 8- and 12-box patterns. Failing a level after four attempts terminates the task, preventing users proceeding to higher difficulty levels. Total error counts are a key outcome measure for the task.

The SWM assesses users working memory and executive function through the acquisition and manipulation of spatial information. Through the process of elimination, users must open boxes to collect hidden tokens, but tokens will never present in the same box twice. Participants must find all the tokens ideally without reopening any boxes that have previously contained one. The task increases in difficulty from 4, 6, 8 to 12 tokens. Total error counts and strategy scores are a key outcome measures for the task.

The MTS assesses processing speed and complex attention by requiring participants to find the exact match of an abstract, nonsemantic target image from an array of visually similar variants and foils. Trials are pseudorandomized across difficulty levels of 1-, 2-, 4-, 6- and 8-pattern choices requiring them to scan each option and choose the correct response as fast as possible. The task has a speed–accuracy trade off with reaction times and accuracy measures being key outcomes.

For further descriptions, see 'Cognition measurement approach' in Supplementary Section 2.

Passive measurements. Multimodal data from the iPhone and Apple Watch can measure across a diverse array of human functions, such as sensorimotor, behavioral, physiologic and autonomic. These data streams are signals with varying sampling frequencies, from event-based sensing (for example, number of running workouts) to routine scheduled sampling (for example, daily volume of calls received) and near continuous high-frequency signals (100 Hz IMU accelerometer). The Sensor Kit (https://developer.apple.com/documentation/sensorkit) system collects information using various sensors on the iPhone and Watch, and computes features derived from sensor information using proprietary algorithms. The main areas covered by the Sensor Kit include device and application usage, keyboard metrics, message and phone use, sound and speech detection, facial metrics, odometer and locations, and the x–y–z coordinates of a body's acceleration and angular velocity from triaxial accelerometer and gyroscope measurements. Health Kit records a variety of health metrics, both sensed and manually entered, and from first- and third-party sources. Examples include physical activity (for example, exercise minutes, active calories burned, step counts, and so on), different types of workouts (for example, running or rowing), time spent and energy burned, walking speed and asymmetry, heart rate and variability, VO_2 max, respiratory rate, oxygen saturation and sleep (https://developer.apple.com/documentation/healthkit).

Ethics, privacy and data storage

The Intuition study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and complied with all applicable regulations and guidance, including but not limited to International Council for Harmonization (ICH) and Good Clinical Practice (GCP) guidelines. This study was approved by the IRB, Advarra (Study ID 285PI401, Board no. 00000971). All participants in the study were provided informed consent electronically and remotely via the Study App.

Secure frameworks were developed to meet the security standards set forth in applicable law, including the deployment of technology and data security processes with vulnerability monitoring and penetration testing. Study data, including any protected health information, were stored in encrypted form at rest and in transit following National Institute of Standard and Technology guidelines outline by the Joint Task Force Transformation Initiative 2013.

The Study App used for enrollment, eligibility screening and active task administration, as well as the platform used for data collection and monitoring, utilized physical, organizational and technical safeguards designed to protect the confidentiality, security and integrity of the data collected. For example, data were encrypted for transmission and storage following guidelines recommended by the US Department of Commerce National Institute of Standard and Technology Federal Information Processing Standard Publication 140-2 that outlines security standards for securing data with health information.

Study oversight

Scientific, ethical and clinical leadership were guided by a Scientific Committee consisting of recognized leaders in the fields of clinical research, neurology, psychiatry and medicine, technology and wearable devices, real-world evidence and biostatistics, bioethics and patient advocacy. The key roles and responsibilities of the Committee were to oversee and provide input on the conduct of the trial, monitor study progress, provide guidance related to recruitment, retention, and attrition and contribute to data analyses and dissemination strategies for scientific results. The Committee members were engaged to provide independent scientific input and guidance for the duration of the study.

Study objectives

The co-primary objectives of the study were:

(1a) To develop and validate a classifier using multimodal passive sensor data and metrics derived from normal iPhone and Apple Watch usage to distinguish individuals with normal cognition from those with MCI.

(1b) To develop and validate a cognitive health score that tracks fluctuations in cognitive performance over time using multimodal passive sensor data and metrics derived from normal iPhone and Apple Watch usage.

The secondary objective of the study was:

(2) To develop a prediction model that uses multimodal passive sensor data and metrics derived from normal iPhone and Apple Watch usage to predict cognitive decline and/or conversion to MCI.

Sample sizes

Formal sample size calculation. Because the primary and secondary objectives of this study did not involve formal, prespecified hypothesis tests, traditional power analyses were not applicable. Two methods for calculating prediction model sample size were chosen: Hanley and McNeil[79]—to ensure precise estimation of the model AUROC; and Riley[80]—to guarantee precision in estimation of the overall outcome proportion (that is, MCI prevalence or incidence), low average prediction error and low likelihood of model overfitting[79,80].

For the Hanley method, sample sizes were computed to ensure an AUROC CI width of ≤0.05. For the approach based on Riley[80], sample sizes were estimated according to the author's recommendations: to ensure a margin of error of ≤0.05 in estimate of outcome proportion, ≤0.05 mean absolute prediction error and small overfitting defined by an expected shrinkage of predictor effects of 10% or less. Calculations

assumed a range of AUC values, and a study outcome proportion (MCI prevalence or incidence for the diagnostic or prognostic models, respectively) of 0.1.

Estimated sample sizes for a range of assumed model AUCs are shown in Supplementary Table 9. With a conservative assumed model AUROC of 0.7, the number of MCI (or MCI converter for the prognostic model) to be included in the study was $n = 360$ based on Hanley, and $n = 564$ based on Riley. Combined target MCI sample sizes from the population-based and clinically validated patient groups of $n = 722$ ensured sufficient numbers to develop a diagnostic model for the primary objective while holding out a large independent test set for model validation. After accounting for attrition and varied adherence (for example, 10–20%) with study protocol activities, the target MCI population size was then set at $n = 800–1,000$ for those at risk for age-related neurodegenerative causes (that is, dementia). Similarly, for those in early and middle adulthood with variable reasons for cognitive impairment, we wanted to target enrollment for $n = 800–1,000$ for adequate model building and testing. For appropriate control group sample sizes required for classifier development, we targeted $N = 6,000$ for those Controls-EM to pair with the MCI-EM and account for the impact of attrition and variable adherence. To approach the secondary objective of the study and develop prognostic models of cognitive decline and MCI conversion, we set similar target sample sizes ($n = 800–1,000$) for anticipated Controls/SCC who might decline and/or convert to MCI. Based on epidemiologic data for projections, we estimated that with 1,500 Controls aged 50–59 years, 2,000 SCC, 6,000 Controls-L Low-Risk and 6,000 Controls-L High-Risk, we anticipated that approximately 500–550 participants would decline clinically annually. As with the diagnostic model, this number of estimated converters was sufficient to develop a well-powered prognostic model for MCI conversion, while holding out a large independent test set. For further discussion about the epidemiologic calculations and reference to previous studies and sample size for MCI classification please reference Supplementary Section 3.

Statistical analyses

Overall approach. Because both the primary objective of a diagnostic MCI classifier and the secondary objective of a prognostic classifier involve primarily the use of baseline or near-baseline participant data—to predict current MCI status and future transition to MCI, respectively—the analytic approaches are similar. Participants will be split into a training dataset, which will be used for all model development and tuning activities, and a testing dataset, which will be set aside for independent model validation. Candidate statistical and machine learning models will be validated on the independent, held-out testing dataset of participants to ensure generalizability of the resulting model performance measures. In addition to assessing model performance, the importance of different features and sensor streams will be assessed to understand which sensor domains provide the most predictive utility for MCI classification and prognosis.

Special concerns must be accounted for in model development due to the nature of the study, with real-world, high-frequency data collection across a wide breadth of data modalities. These concerns include understanding and accounting for data missingness, choosing optimal sampling windows and temporal resolutions, and balancing interpretability with performance in what could be very complex models.

Intermediate steps are also necessary before achieving the primary and secondary objectives. These involve but are not limited to (1) assessment of adherence and data missingness across the active and passive data streams of the study; (2) characterization of the active unsupervised cognitive task data, including between- and across-task correlations, possible ceiling/floor effects and psychometric validation steps such as looking for expected demographic and clinical associations and assessing test–retest reliability and learning effects; (3) full examination of the passive data streams, many of which are exploratory in the context of cognition; and (4) data reduction steps, particularly for the high-volume passive sensor streams.

Initial approach to analyses and proof-of-concept MCI classifier. For this manuscript, we began MCI classification with readily interpretable statistical tests and modeling. For baseline characterization of cohorts by age and cognitive status we applied analysis of variance and pairwise *t*-tests to group means and variance for selected key subjective (CFI/E-Cog) and objective (CANTAB) measures of cognition. For model building, we started with logistic regression in a subset of validated MCI (MCI-CC + MCI tele-health confirmed, $N = 556$) versus a large diverse population aged 50–86 years both with and without cognitive complaints (SCC + Controls, $N = 16,234$). Predictor variables (total input feature space of $N = 205$) included core demographics of age, sex and education, and baseline subjective and objective cognition as measured by CFI and E-Cog total and item-level scores, and $N = 176$ CANTAB outcomes based on PRMi, PRMd, PAL, SWM and MTS assessments, all of which were scaled/standardized. The proof-of-concept MCI classifier is a logistic regression model with ridge penalization (L2 regularization). The model incorporates all baseline CANTAB outcomes (objective cognitive performance measures), along with two subjective cognition surveys: CFI and E-Cog. In addition to these cognitive assessments, the model also uses core demographic variables including age, sex and education level. The data were split into 80% for training and 20% for testing. To address class imbalance between the majority and minority classes, the training data were resampled using a 3:1 majority-to-minority class ratio. The model was trained using 100 times bootstrap resampling in the outer loop to enhance generalization and estimate the stability of the model. Within each bootstrap iteration, grid search was employed in the inner loop to systematically explore a range of hyperparameters, specifically the regularization strength for ridge penalization, and identify the best-performing hyperparameter configuration. To further ensure robust evaluation, the inner loop applied stratified fivefold crossvalidation, which maintained class balance within each fold while testing different hyperparameter sets. This nested crossvalidation setup ensured that the hyperparameter tuning of the model was independent of the outer loop resampling, minimizing the risk of overfitting and optimizing performance on unseen data. Model accuracy and mean AUROC on the test dataset are reported with 95% CIs. A list of rank ordered predictor coefficients supplement the results for model interpretability.

Reporting summary

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

Data availability

Given the language in the ICF and the contractual study agreement between Apple and Biogen, participant-level data from the study is not currently available to other researchers. Given the potential for this dataset to facilitate advances and discovery in cognition, consideration is being given to academic partnerships or other avenues for data sharing that respect the ICF and contractual agreements.

References

62. Barnett, J. H., Blackwell, A. D., Sahakian, B. J. & Robbins, T. W. The paired associates learning (PAL) test: 30 years of CANTAB translational neuroscience from laboratory to bedside in dementia research. *Curr. Top. Behav. Neurosci.* **28**, 449–474 (2016).
63. Robbins, T. W. et al. Cambridge neuropsychological test automated battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers. *Dementia* **5**, 266–281 (1994).

64. Lenehan, M. E., Summers, M. J., Saunders, N. L., Summers, J. J. & Vickers, J. C. Does the Cambridge automated neuropsychological test battery (CANTAB) distinguish between cognitive domains in healthy older adults? *Assessment* **23**, 163–172 (2016).
65. Robbins, T. W. et al. A study of performance on tests from the CANTAB battery sensitive to frontal lobe dysfunction in a large sample of normal volunteers: implications for theories of executive functioning and cognitive aging. *J. Int. Neuropsychol. Soc.* **4**, 474–490 (1998).
66. Chamberlain, S. R. et al. Differential cognitive deterioration in dementia: a two-year longitudinal study. *J. Alzheimers Dis.* **24**, 125–136 (2011).
67. Zhuang, L., Yang, Y. & Gao, J. Cognitive assessment tools for mild cognitive impairment screening. *J. Neurol.* **268**, 1615–1622 (2021).
68. Egerházi, A., Berecz, R., Bartók, E. & Degrell, I. Automated neuropsychological test battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **31**, 746–751 (2007).
69. Junkkila, J., Oja, S., Laine, M. & Karrasch, M. Applicability of the CANTAB-PAL computerized memory test in identifying amnestic mild cognitive impairment and Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* **34**, 83–89 (2012).
70. Lange, K. W. et al. L-dopa withdrawal in Parkinson's disease selectively impairs cognitive performance in tests sensitive to frontal lobe dysfunction. *Psychopharmacol. (Berl.)* **107**, 394–404 (1992).
71. Owen, A. M. et al. Fronto-striatal cognitive deficits at different stages of Parkinson's disease. *Brain* **115**, 1727–1751 (1992).
72. Rock, P. L., Roiser, J. P., Riedel, W. J. & Blackwell, A. D. Cognitive impairment in depression: a systematic review and meta-analysis. *Psychol. Med.* **44**, 2029–2040 (2014).
73. Summers, M. J. & Saunders, N. L. Neuropsychological measures predict decline to Alzheimer's dementia from mild cognitive impairment. *Neuropsychology* **26**, 498–508 (2012).
74. Ashford, M. T. et al. Unsupervised online paired associates learning task from the Cambridge neuropsychological test automated battery (CANTAB®) in the Brain Health Registry. *J. Prev. Alzheimers Dis.* **11**, 514–524 (2024).
75. Backx, R., Skirrow, C., Dente, P., Barnett, J. H. & Cormack, F. K. Comparing web-based and lab-based cognitive assessment using the Cambridge neuropsychological test automated battery: a within-subjects counterbalanced study. *J. Med. Internet Res.* **22**, e16792 (2020).
76. Green, R. et al. Assessment of neuropsychological performance in Mexico City youth using the Cambridge neuropsychological test automated battery (CANTAB). *J. Clin. Exp. Neuropsychol.* **41**, 246–256 (2019).
77. Giaquinto, F., Battista, P. & Angelelli, P. Touchscreen cognitive tools for mild cognitive impairment and dementia used in primary care across diverse cultural and literacy populations: a systematic review. *J. Alzheimer's Dis.* **90**, 1359–1380 (2022).
78. Papp, K. V. et al. Early detection of amyloid-related changes in memory among cognitively unimpaired older adults with daily digital testing. *Ann. Neurol.* **95**, 507–517 (2024).
79. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
80. Riley, R. D. et al. Calculating the sample size required for developing a clinical prediction model. *Br. Med. J.* **368**, m441 (2020).

Acknowledgements

We thank the participants for their involvement in this research study. We thank our colleagues at Cambridge Cognition, specifically F. Cormack and N. Taptiklis, for their innovation in digital measurement of cognition. In addition, we acknowledge the expert guidance and support of behavioral neurologist, Y. Zabar.

Author contributions

P.M.B. wrote the manuscript, performed analysis and served as the study medical lead. J.Y., R.B., M.H., A.B., J.P.-A., P.S., S.M., G.C., A.J., H.H.S., P.S.-C., D.R., A.S., N.S. and G.D. were involved in data collection, generation, analysis and interpretation. A.G., R.H., M.T.B., H.L., H.P., M.P. and S.B. made substantive conceptual and intellectual contributions to the study design, deployment and interpretation of results. M.G.E., J.B.L., J.H.L., P.P., Y.T.Q., S.J.S., M.S., A.P.P. and R.A. served on the study scientific committee providing oversight and guidance for this project. All authors contributed to the drafting and revision of the manuscript.

Competing interests

The following authors are employed by Apple, Inc.: P.M.B., J.Y., M.H., P.S., S.M., H.H.S., N.S., M.T.B., H.L., H.P. and M.P. The following authors are employed by Biogen, Inc.: P.M.B., R.B., M.H., A.B., J.P.-A., G.C., A.J., P.S.-C., D.R., A.S., G.D., A.G., R.H., S.B. J.B.L. is a full-time employee of Banner Health. Banner Alzheimer's Institute receives funding from Eli Lilly for its collaborative partnership on TRAILBLAZER-ALZ 3; she reports receiving grants from the NIA unrelated to this project. A.P.P. has received personal fees from Acadia Pharmaceuticals, Athira, BMS, Cognitive Research Corp, IQVIA, Lundbeck, Novartis, ONO Pharmaceuticals, Otsuka, WCG, WebMD and Xenon, and grants to institution from Athira, Biogen, Cassava, Eisai, Eli Lilly, Genentech/Roche, Vaccinex, NIA, NIMH and DOD; he is a member of the Scientific Advisory Board of Alzheon, Athira, and Cognition Therapeutics. Y.T.Q. serves as consultant for Biogen on other unrelated projects. R.A. serves as scientific advisor to Signant Health and NovoNordisk. M.G.E., J.B.L., J.H.L., P.P., Y.T.Q., S.J.S., M.S., A.P.P. and R.A. all served as consultants to Biogen, Inc. on this project as members of the Intuition Study Scientific Committee.

Additional information

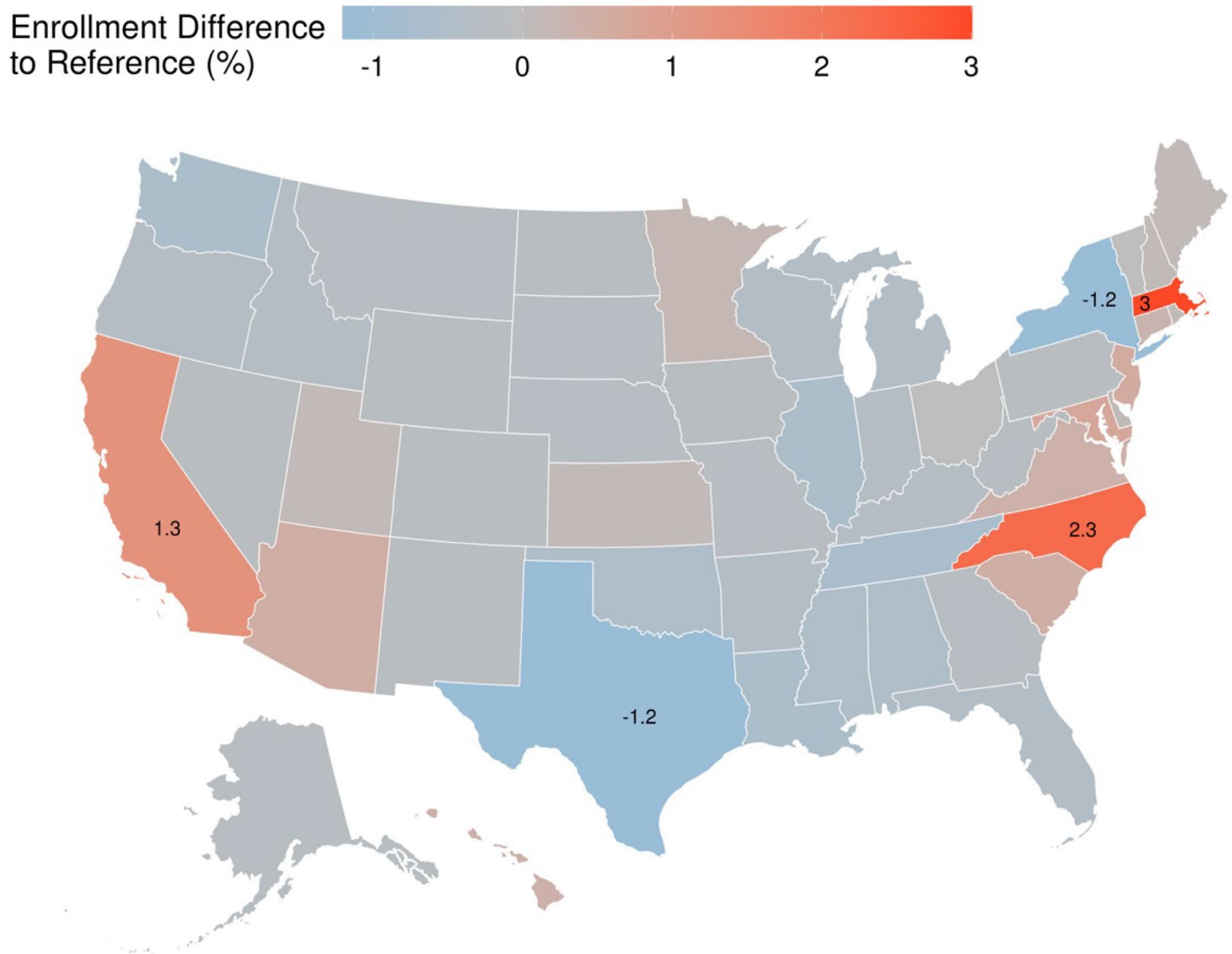
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03475-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03475-9.

Correspondence and requests for materials should be addressed to Paul Monroe Butler.

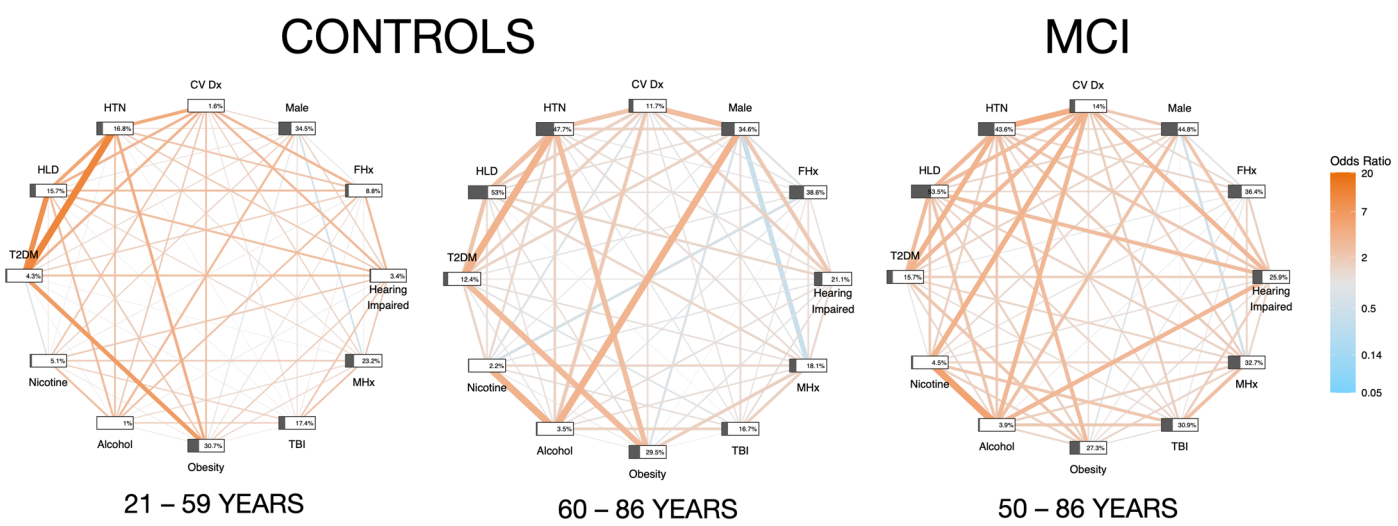
Peer review information *Nature Medicine* thanks Louisa Thompson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jerome Staal, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-03475-9



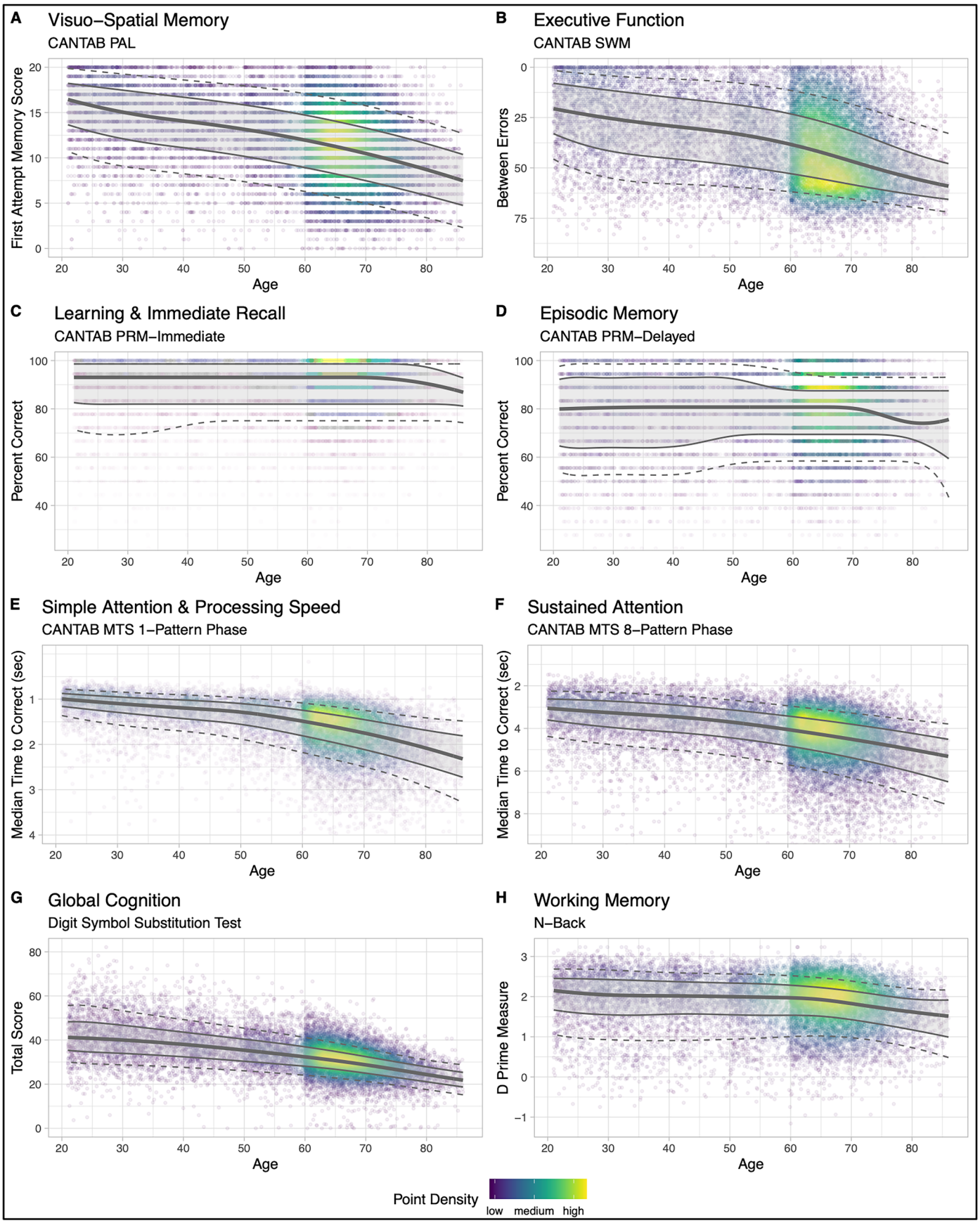
Extended Data Fig. 1 | Geographic Diversity in Enrollment across all U.S. States. Participant enrollment rate map compared to reference population. This map outlines percentage of the participant population coming from each state. These proportions are compared to the 2022 US Census population proportions for each state. Orange positive valued states are locations where over enrollment occurred relative to the general population levels. In terms of absolute deviation, only a handful of states had noticeable over or under enrollment. California and North Carolina over enrolled by 1.3% and 2.4% respectively compared to their general population. Texas and New York were the largest under enrollers which under enrolled by 1.2% each. The remainder of the states were all within 1% of the reference levels. For a complete list of state-based enrollment and reference population statistics see Online Supplementary Table 2.

https://doi.org/10.1038/s41591-024-03475-9



Extended Data Fig. 2 | Risk Factor Odds Ratio Connection Plots. Each connection plot displays the pairwise odds ratios between binary risk factors. The width of the connections is proportionate to the strength of the association and the colors display the directionality of the odds ratio with blue and orange signifying lower (<1 negative associations) or higher (>1 positive associations) values, respectively. At each node along the edge, the prevalence for each risk factor is cited in reference to the total cohort. The connections between nodes identifies risk factors that are highly associated with each other and likely to occur concomitantly. HTN = hypertension; HLD = hyperlipidemia; T2DM = Type 2 Diabetes Mellitus; CV Dx = cardiovascular disease; FHx = family history of dementia, 1st degree relative; MHx = Mental health history; TBI = traumatic brain injury. See the Methods section for a complete description of the risk factor definitions. Nicotine refers to any active use of nicotine products. Alcohol refers to a history of heavy consumption, 20 units per week for 10 years or longer. Obesity is defined by BMI ≥ 30. Hearing impairment is any history of hearing issues on the medical review of systems. TBI refers to any history of TBI without regard to severity or frequency. Mental health history is any remote or active mental illness. Cardiovascular disease is endorsement of any of the following: heart attack, atrial fibrillation, angioplasty, stent placement, or endarterectomy, cardiac bypass or other blood vessel bypass grafting procedure, pacemaker and/or defibrillator placement, congestive heart failure, angina, heart valve replacement or repair, or peripheral vascular disease.

https://doi.org/10.1038/s41591-024-03475-9

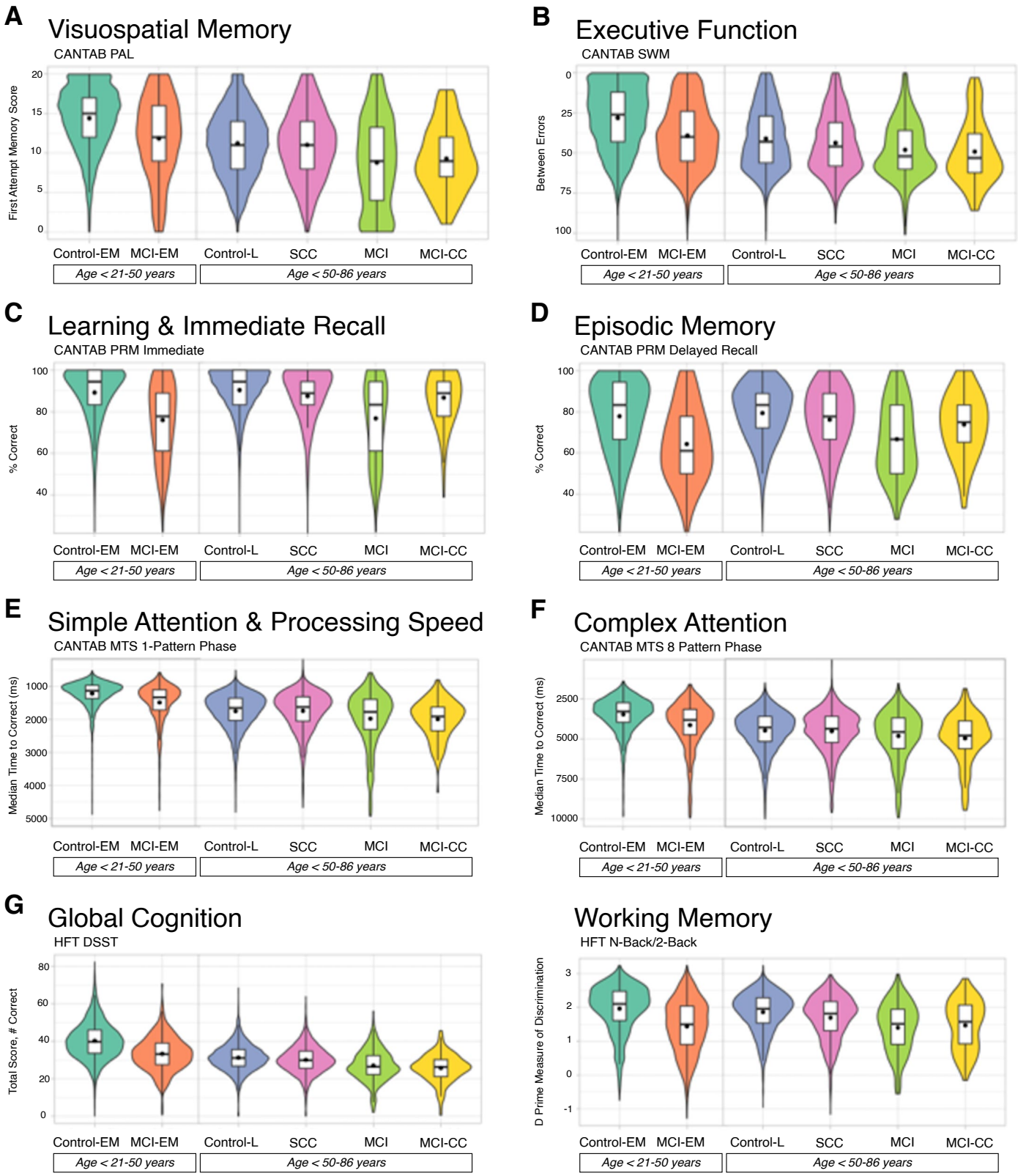


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-024-03475-9

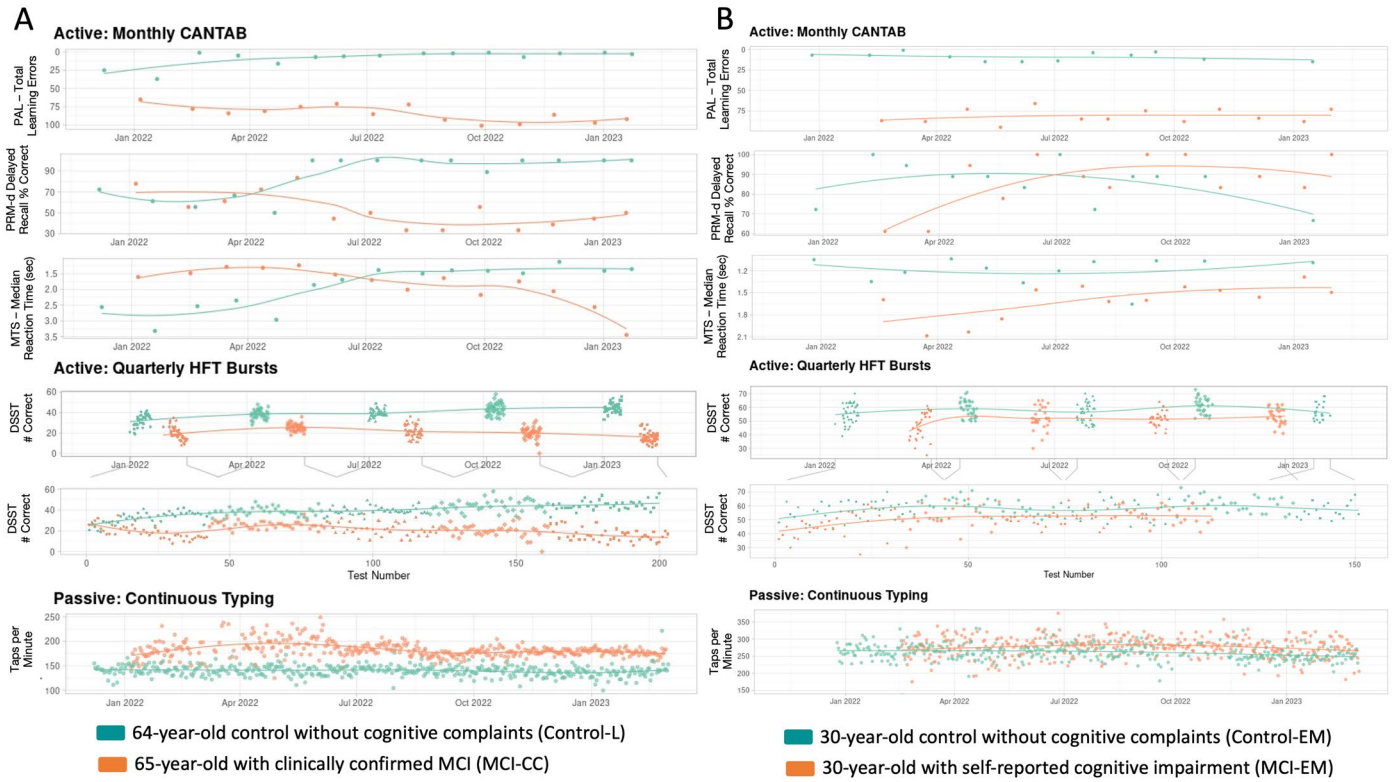
Extended Data Fig. 3 | Cognitive Baseline Performance in Controls by Age. Quantile curves of baseline cognitive performance on 8 representative measures plotted as smooth functions by age for all study control participants. CANTAB = Cambridge Neuropsychological Test Automated Battery from Cambridge Cognition; PAL = paired associates learning; SWM = spatial working memory (SWMBE46812; SWM Between Errors= errors by selecting boxes already chosen with tokens); PRM = pattern recognition memory; MTS = match-to-sample; All plots A-H show control subjects across the aging lifespan with sample outcomes from CANTAB (**a**-**f**) at baseline and burst 1 means for Plots **g**-**h**. Density plots are depicted with dashed lines indicating the 10^{th} and 90^{th} percentile, solid lines representing the 25^{th} and 75^{th} percentile, and with the age-associated inter-quartile range shown with the light gray shading. The median is denoted by the thick solid line. N = 18,845 control participants.

https://doi.org/10.1038/s41591-024-03475-9



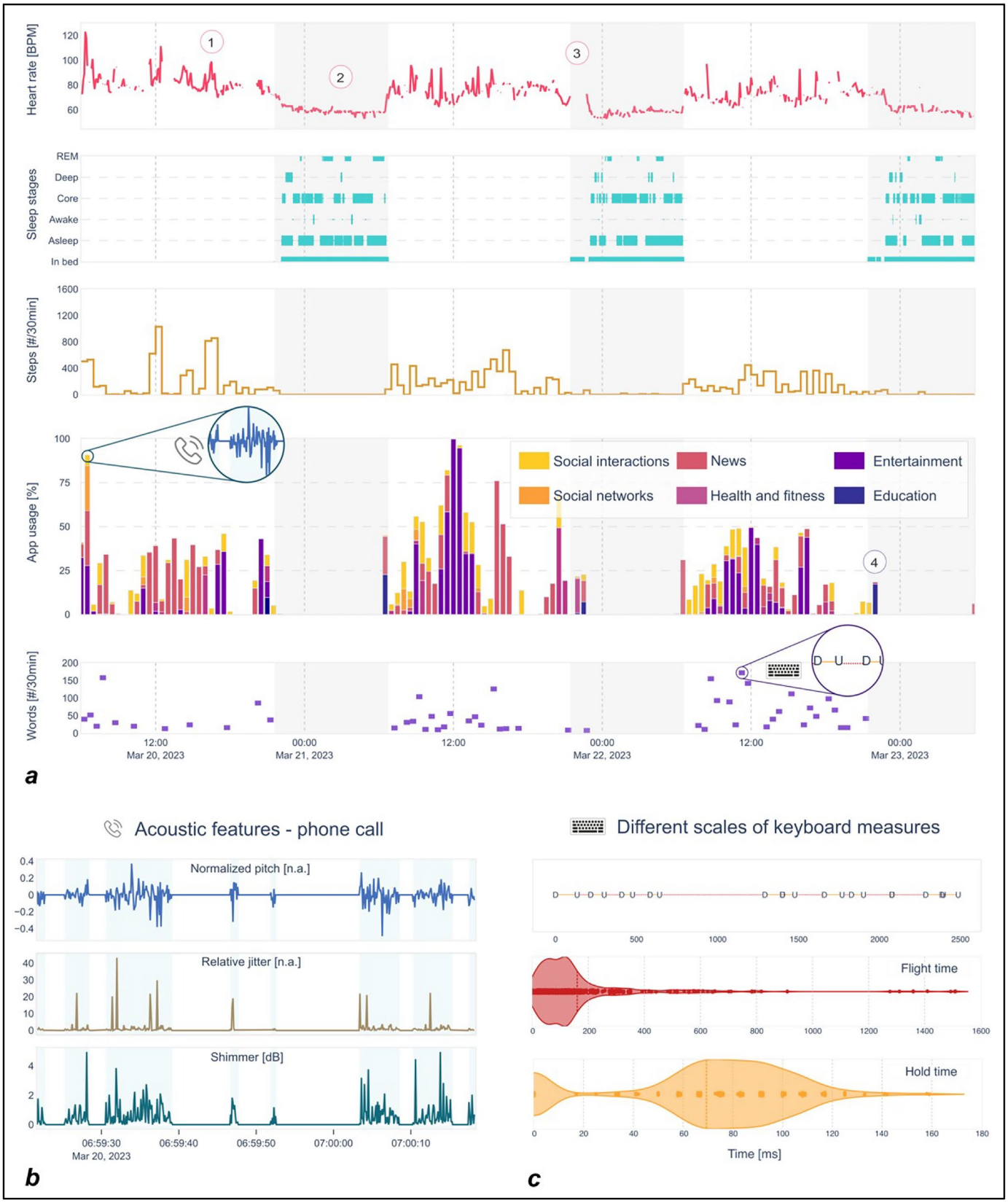
Extended Data Fig. 4 | Cognitive Baseline Performance by Cohort and Age. Box and violin plots of baseline cognitive performance for 8 representative measures from participants grouped by age and cohort status. *CANTAB = Cambridge Neuropsychological Test Automated Battery from Cambridge Cognition; PAL = Paired Associates Learning; SWM = Spatial Working Memory; SWM Between Errors= errors by selecting boxes already chosen with tokens; PRM = Pattern Recognition Memory; MTS = Match-To-Sample; DSST = Digit Symbol Substitution Test; 2-Back = N-Back. Cohorts displayed include Control-EM = Early and Middle Adulthood; MCI-EM = Mild Cognitive Impairment Early and Middle Adulthood; Control-L = Control Late Adulthood at Low/High Risk for cognitive decline; SCC = Subjective Cognitive Complaint; MCI = mild cognitive impairment self-report; MCI-CC = MCI clinically confirmed. Violin plots show CANTAB baseline outcomes in Panel A-F and high frequency burst outcomes averaged from burst 1 for Panels G-H. The box denotes the interquartile range (IQR), bold line the median, and bold dot the mean. Whiskers extend to the largest/smallest observations no further than 1.5 times the IQR from each side of the box.*

https://doi.org/10.1038/s41591-024-03475-9



Extended Data Fig. 5 | The Intersection of Passive Sensing and Active Assessment of Cognition from Sample MCI and Control Participants. Temporal alignment over 12-months from illustrative examples of demographic matched individuals with and without cognitive impairment. Active assessment outcomes from monthly CANTAB demonstrate longitudinal cognitive performance in learning, consolidation and delayed recall, and processing speed. Quarterly burst high frequency assessment outcomes reflect global cognition. Passive sensing of cognition through iPhone typing dynamics reflects cognition deployed in real world settings. Note: All cognitive outcomes are shown with respect to y-axis depicts higher and lower performance in cognition as such. Panel **a** depicts 12-months of multimodal longitudinal cognitive data in a demographically matched late adulthood control and patient with clinically confirmed amnestic multidomain MCI. Notable trends include baseline deficits in learning and global cognition which worsen over time and then the emergence of new deficits in attention and recall. In the Study App, the participant with MCI reported concern for steadily worsening cognition, suggesting an age-related neurodegenerative process. Panel **b** depicts a young/middle adulthood participant with cognitive impairment reported in the Study App as an abrupt onset deficit in learning/memory, which was reported as stable/unchanging over time. Baseline deficits in learning are apparent compared to a demographically matched control and remain stable across longitudinal assessments. In the setting of amnestic deficits, one can still appreciate familiarization, learning effects, and improvements in performance on recall, attention, and global cognitive function. Outcomes include Paired Associates Learning (PAL) Total Adjusted Errors, Pattern Recognition Memory Delayed (PRM-d) Recall = percent correct on PRM delayed, Match-To-Sample (MTS) Median Correct Response Time in seconds, Digit Symbol Substitution Test (DSST) Total Number Correct, Passive Cognition = Taps per Minute of Keyboard Characters when Typing Messages on the iPhone.

https://doi.org/10.1038/s41591-024-03475-9

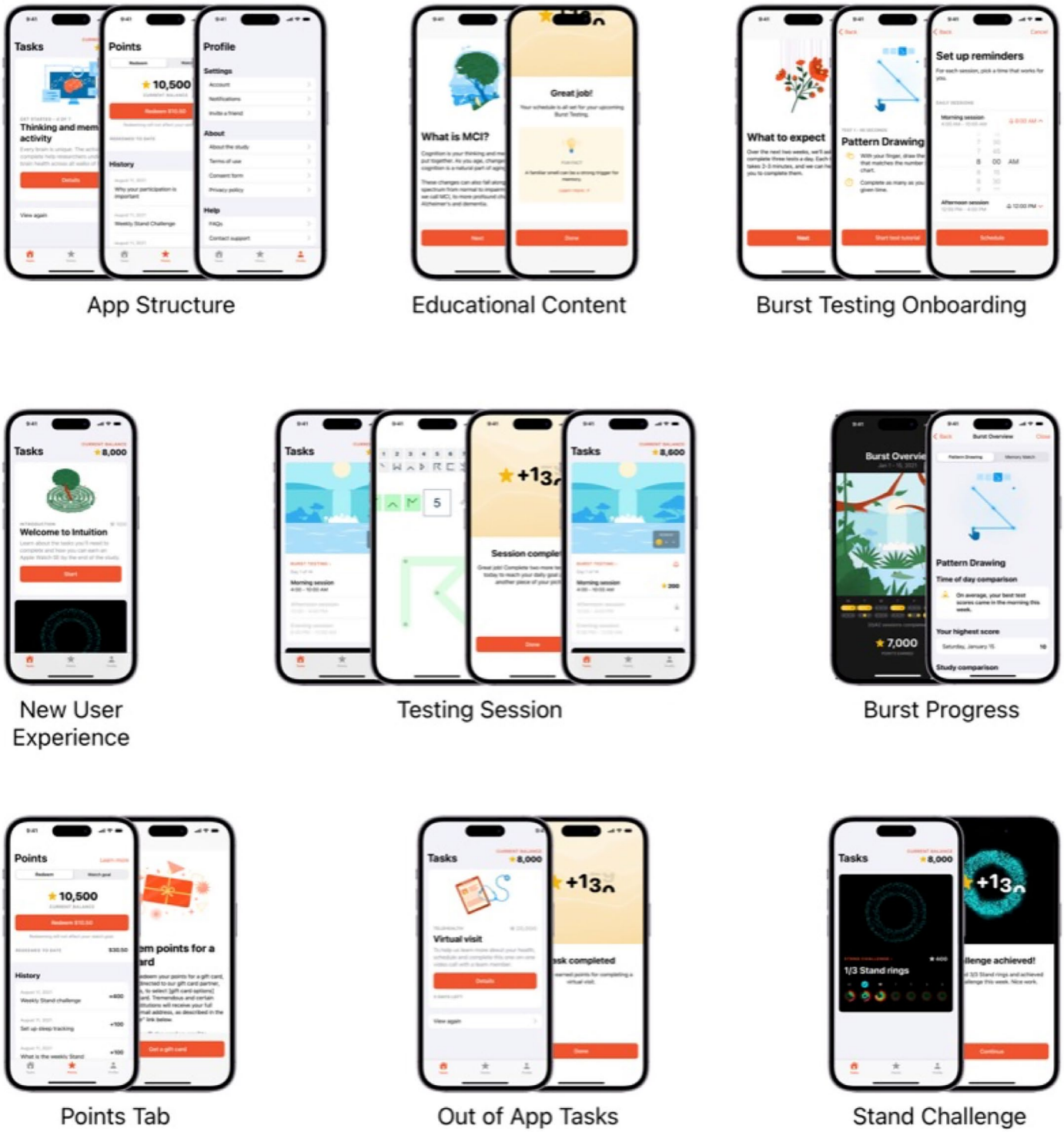


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41591-024-03475-9

Extended Data Fig. 6 | Examples of Multimodal Passive Data Collected via iPhone and Apple Watch in a Sample Study Participant. Time-aligned physiologic, autonomic, motor, and behavioral data are depicted across 3 panels. *Panel A* shows data related to circadian rhythms, including heart rate (row 1), sleep stages (row 2), step count (row 3) and App-use behavior (row 4). *Panel B* depicts motor speech behavior, including acoustic properties of voice like pitch, jitter, and shimmer (row 1-3 respectively). *Panel C* plots temporal dynamics of typing behaviors from iPhone keyboard use, including press-and-release called hold and flight times. In Panel **a**, row 1 heart rate collected from Apple Watch PPG (photoplethysmography) sensors, are tagged for periods of (1) activity, (2) quiescence, and (3) poor skin contact or Watch removal. The 30-minute time-aligned rows have light grey backgrounds when the sensors indicate the participant is in bed. Additional insights can be observed when interpreting multiple time aligned data streams, such as heart rate peaks as the number of steps increases and slows during suspected sleep. Panel **b** illustrates extracted acoustic features of voice from microphone collected data such as frequency adjusted measure of periodicity related to vocal pitch, and moment by moment variations in the fundamental frequency (that is, jitter) and amplitude (that is, shimmer) in vocal intensity. Panel **c** shows millisecond scale typing dynamics when a user is using the virtual iPhone keyboard. Row 1 shows the time scale of pushing down on the screen to type or 'D' for depressing a button and when the individual releases between touches and is up or 'U.' Row 2 and 3 show millisecond-scale dynamics for a sample release time ('U' or flight time) and button depress ('D' or hold time).

https://doi.org/10.1038/s41591-024-03475-9



Extended Data Fig. 7 | Intuition Study App and Participant Experience. The Intuition bespoke Study App was downloaded from the App store by a prospective enrollee who then underwent screening for eligibility and then e-consent. For those consented participants to onboard and orient to the study the experience included familiarization with the App structure and study activities. This multi-panel figure provides sample snapshots of the participant experience.

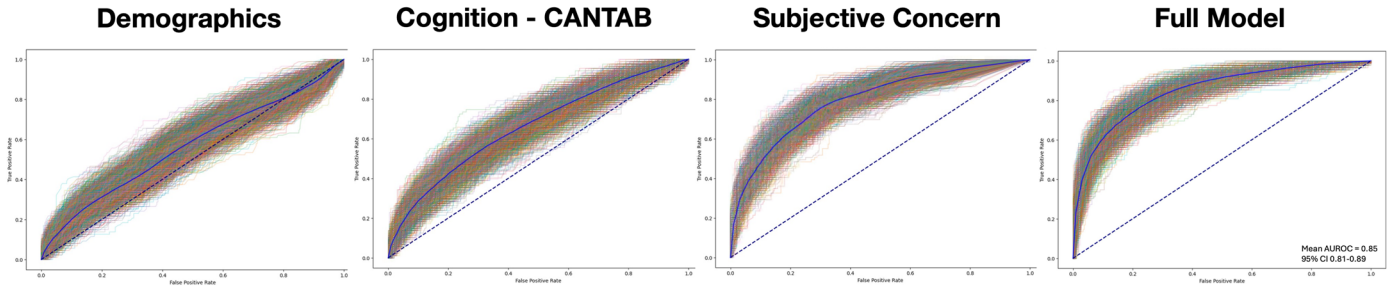
https://doi.org/10.1038/s41591-024-03475-9

		Study Start	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
Study Onboarding	Eligibility screening	✓												
	Informed consent	✓												
	Medical history	✓												
	Psychosocial status	✓												
	Apple Watch provisioning	✓												
Cognitive Assessments	CANTAB battery	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Cambridge Cognition test on iPhone		✓			✓			✓			✓		
	Speech and language battery		✓			✓			✓			✓		
Surveys	PHQ-2, GAD-2	✓			✓			✓			✓			✓
	Lifestyle (exercise, sleep, diet, substance use)	✓			✓			✓			✓			✓
	Weight				✓			✓			✓			✓
	Demographics update							✓						✓
	Cognitive Function Instrument							✓						✓
	Medical history/medication update							✓						✓
	Global cognitive function (ECog-12)	✓						✓						✓

Experience repeats for 24 months

Extended Data Fig. 8 | Study Data Sampling Cycles and Schedule of Activities. Cadence of key onboarding and longitudinal interactive cognitive and self-report assessments over 12 months of study activities.

https://doi.org/10.1038/s41591-024-03475-9



Comparison of MCI Classifier Logistic Regression Model Performance

Statistic	Demographics	CANTAB	Subjective Concerns	Full Model
Sensitivity	55.9%	60.4%	72.1%	80.2%
Specificity	54.7%	61.0%	70.6%	78.7%
TPR	0.58	0.62	0.72	0.80
FPR	0.46	0.40	0.30	0.21
PPV	29.1%	34.1%	44.9%	55.6%
NPV	78.8%	82.2%	88.4%	92.3%
Accuracy	55.0%	61.0%	71.0%	79.1%

Extended Data Fig. 9 | Primitive MCI Classification Models Using Demographics, CANTAB Cognitive Performance, Subjective Cognitive Concerns, and a Full Baseline Model. ROC = Receiver Operating Curve, AUROC = Area Under the Receiver Operating Curve, CI = Confidence Interval, SCC = Subjective Cognitive Complaints, MCI = Mild Cognitive Impairment. N = 16,790 total participants in the analysis with all controls age 50 years and above were included alongside participants with CFI-defined SCC. MCI cases were those clinically confirmed cases and self-reported MCI that was confirmed by a tele-research visit, including a tele-MoCA to confirm impairment. MCI classifier model was a logistic regression model with ridge penalization (L2) using all baseline CANTAB outcomes (objective) plus 2 subjective cognition surveys (CFI + E-Cog) and core demographics (age, sex, education). Majority-to-minority class sampling was 3:1 with 100x bootstrapping with nested cross-validated.

https://doi.org/10.1038/s41591-024-03475-9

Extended Data Table 1 | Study data sources

Domain	Device	Description	Data streams	Sampling frequency
Demographics, Health, and Habits	Study App iPhone	Demographic information	Age, sex, gender, race, ethnicity, education, marital status, income, geography, height, weight, and handedness	Baseline
		Social history, habits, and lifestyle	Diet, caffeine, exercise, sleep, substance use, mobility, and change in weight	Baseline + Quarterly
		Medical history and risk factors	Review of systems, medical history and risk factors for cognitive decline, COVID-19, women's health, and medications	Baseline + Biannual
Global and Mental Health	Study App iPhone	Mental health history	History of mental illness, including MDD, GAD, PTSD, OCD, bipolar and schizophrenia, or on psychotropic treatment	Baseline
		Stress and mood symptoms	PSS, PHQ-2 and GAD-2; depression and anxiety symptom screen	Baseline + Quarterly
		Global health and quality-of-life	PROMIS v1.2 Global Health and General Self-Efficacy, and Major Experiences of Discrimination survey	Baseline + Biannual
Cognitive Health	Study App iPhone	Subjective cognition and MCI diagnosis	CFI-14, ECog-12, global function and ADLs, and MCI diagnosis and work-up survey	Baseline + Biannual
		High-frequency cognitive assessments	DSST and N-Back tasks performed 3-times daily	Quarterly*
	Personal Computing Device + Wi-Fi	Cognitive assessment battery	CANTAB with i/d PRM, PAL, SWM, and MTS tasks	Baseline + Monthly
		Tele-research visit	Cognitive history, tele-MoCA, and MCI label confidence	Baseline + Biannual
Communication, Social Function, & Device Use	Study App iPhone	Speech and language battery[a]	Recorded voice with connected speech, phonemic and semantic fluency, picture description and passage reading	Quarterly
	iPhone + Watch[1,2]	Keyboard use	Typing behaviors (e.g., speed, corrections, pauses)	Every 15 minutes
		Locations of interest	Distance from home, locations of interest and time-of-day	
		Messages and phone calls	Volume of incoming/outgoing messages, calls and contacts	Daily
		App and device usage	Screen unlocks, App usage, and device charging	Event-based
		Speech and sounds	Speech acoustics (e.g., shimmer, jitter), speaking rate, and sound detection	
		Facial metrics	Mathematical representation of gaze and facial pose, movements, and expression	
Motor Function	iPhone + Watch [1,2]	Accelerometer	Linear acceleration in three-dimensional space	100 Hz
		Gyroscope	Angular velocity of the device in three-dimensional space	100 Hz
		Pedometer	Measures of mobility (e.g., step count, distance, walking speed, stand time)	Event-based
		Workouts	Physical activity (e.g., type, duration, energy expenditure)	
Physiology and Autonomics	Watch [1,2]	Sleep and wakefulness	Sleep time and stages (e.g., core, deep, REM)	Event-based
		Thermal sensor + PPG + Acceleration	Heart rate, resting heart rate, variability, blood oxygen saturation, respiratory rate, VO2 max, body temperature	0.1-0.25 Hz, dynamic
	iPhone + Watch [1,2]	Mindfulness breathing sessions	Use and duration of controlled breathing sessions	Event-based

***Quarterly (14 days - Morning, Afternoon, Evening), [a]Files of sound [1,2]For updated SensorKit[1] and HealthKit[2] features, please refer Apple Documentation.**
Glossary: *CognitionKit High Frequency Testing (HFT):* N-Back (N-Back Memory Test), DSST (Digit Symbol Substitution Test), *Cambridge Neuropsychological Test Automated Battery (CANTAB):* i/d PRM (immediate/delayed Pattern Recognition Memory Test), PAL (Paired Associates Learning test), SWM (Spatial Working Memory test), MTS (Match-to-Sample test), PHQ-2 (2-item Patient Health Questionnaire for mood), GAD-2 (2-item Generalized Anxiety Disorder), CFI-14 (14-item Cognitive Function Instrument), ECog-12 (12-item everyday cognition), PSS (Perceived Stress Scale), PROMIS-10 (Patient-Reported Outcome Measurement Information System), MoCA (Montreal Cognitive Assessment), REM (Rapid Eye Movement sleep phase), ADL: Activities of Daily Living, PPG: Photo-plethysmography.