JMIR Neurotechnology

For research exploring how technologies (e.g. information technology, neural engineering, neural interfacing, clinical data science, robotics, eHealth/mHealth) can be applied in clinical neuroscience (e.g., neurology, neurosurgery, neuroradiology) to prevent, diagnose, and treat neurological disorders.
Volume 5 (2026) ISSN 2817-092X Editors-in-Chief: Pieter Kubben MD, PhD

Contents

Viewpoint

Tracking Cognitive Health With Wearables in Telerehabilitation Female Participants: Could Nighttime Sleep Measures Be Used as Sex-Specific Digital Endpoints? (e81318)
Stephanie Zawada, Louis Faust, Emma Fortune. . . . 2

Original Papers

A Pocket Laboratory for Functional Neuroimaging Research Using Mobile Visual Oddball, Multimodal Electroencephalography, and Functional Near-Infrared Spectroscopy Imaging: Instrument Validation Study (e78217)
Peter Rokowski, Meltem Izzetoglu, Luis Gomero, Roee Holtzer. . . . 13

Feasibility of Interleaving Computerized Cognitive Training With Repetitive Transcranial Magnetic Stimulation: Pilot Studies in Mild Cognitive Impairment Due to Alzheimer's Disease and Stroke (e81437)
Laura Campbell, Stephanie Aghamoosa, Olivia Horn, Holly Fleischmann, James Lopez, Katrina Rbeiz, Katrina Madden, Michael Antonucci, Gonzalo Revuelta, Lisa McTeague, Andreana Benitez. . . . 30

Diagnostic Accuracy of GPT-4 With Vision in Neuroradiology Board-Style Exam Questions: Cross-Sectional Case-Based Study (e69708)
Tom Sussan, Rebekah Brawley, Joshua Eckroth, James Mossell, Tao Weitao. . . . 50

xHD-Vox, an Automated Speech Model for Estimating Motor and Cognitive Scores in Huntington Disease: Development and Longitudinal Validation (e83838)
Tiphaine Le Ludec, Andres Gil-Salcedo, Hadrien Titeux, Robin Louiset, Clément Le Moine Veillon, Renaud Massart, Anne-Catherine Bachoud-Lévi. . . . 65

Self-Guided Beta-Responsive Visual Neurofeedback in a Public Exhibition: Feasibility and Usability Randomized Controlled Trial (e89256)
Melanie Baumgartner, Raphael Weibel, Tobias Hoesli, Florian von Wangenheim, Joseph Ollier. . . . 84

XSL•FO
RenderX

Tracking Cognitive Health With Wearables in Telerehabilitation Female Participants: Could Nighttime Sleep Measures Be Used as Sex-Specific Digital Endpoints?

Stephanie J Zawada[1,2], MS, PhD; Louis Faust[2], PhD; Emma Fortune[1,2], PhD

[1]Division of Health Care Delivery Research, Mayo Clinic, 200 1st Street SW, Rochester, MN, United States

[2]Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Stephanie J Zawada, MS, PhD
Division of Health Care Delivery Research, Mayo Clinic, 200 1st Street SW, Rochester, MN, United States

Abstract

While changes in brain structure are common across the lifespan, it is difficult to differentiate benign variations from early disease pathogenesis, especially in patients participating in home-based rehabilitation. Cognitive decline is frequently linked with normative aging, but its early detection can facilitate preventative interventions, particularly in patients at high risk of cognitive impairment and dementia, such as older women. Although women have fewer modifiable risk factors for dementia than men, wearables have the potential to establish new digital endpoints that facilitate the management and/or prevention of abnormal brain aging. Sleep, which is necessary for maintaining overall brain health, is one behavior tracked via wearables, for which an ever-growing body of pilot and validation phase studies exists, yet endpoints defining optimal sleep are not one-size-fits-all, as individual chronotype variability necessitates new strategies to personalize care. Though sex-based differences in circadian rhythm are well established, little is understood about which sleep measures and thresholds are uniquely important to cognitive health in women, particularly those with high comorbid burden. In this viewpoint, we discuss recent findings on the use of wearables to track sleep and cognitive health in women, while highlighting challenges and opportunities for health outcomes and clinical trial researchers seeking to implement meaningful digital endpoints in future telerehabilitation programs.

(JMIR Neurotech 2026;5:e81318) doi:10.2196/81318

KEYWORDS

cardiopulmonary health; cognition; dementia; digital health; sleep; wearables; women's health

Introduction

With the rapid adoption of wearables in home-based rehabilitation, also known as telerehabilitation, an emerging need exists to establish clinically meaningful digital endpoints, “precisely defined variable[s] intended to reflect an outcome of interest that is statistically analyzed to address a particular research question” [1]. For telerehabilitation participants with neurologic or functional deficits, such as those with recent hospitalizations for stroke, chronic obstructive pulmonary disease, or coronary artery disease, research has shown the effectiveness of interventions to monitor and track exercise through digital endpoints derived from wearable devices, such as steps per day [2-4]. Beyond physical activity monitoring, the application of wearables to track cognitive health, defined as the “improvement, maintenance, or minimal decline of cognitive function and absence, delay of onset, or slowing the progression of dementia,” is on the rise, yet it remains understudied in telerehabilitation [5-8].

The monitoring of cognitive health during telerehabilitation has typically relied on videoconferencing, requiring clinical assessments of participant status; however, the link between cognitive health and sleep, which can be passively observed using wearables, is well documented [9,10]. Historically, research investigating the sleep-cognition relationship has primarily featured univariate correlations between scores obtained from cognitive assessments and self-reported sleep duration [11]. Both short and long sleep durations, outside the optimal window (7 - 9 h per night), have been associated with cognitive decline; however, the application of wearable devices has introduced new inconsistencies regarding this association [12]. Though self-report bias remains a known confounder in analyses relying on survey instruments, arbitrary thresholds, such as device-specific cutoffs for nighttime versus daytime sleep, used in sleep data processing might contribute to conflicting results regarding wearable-derived sleep duration and cognitive health [13]. The generalizability of these results to telerehabilitation populations remains unknown [14].

Despite limitations, wearable devices offer an enhanced approach to elucidating the sleep-cognition link by capturing objective measures beyond sleep duration, from sleep fragmentation scores to nighttime blood oxygen saturation, offering a never-before-seen multidimensional view of sleep

XSL•FO
RenderX

[13]. To realize the potential of multidimensional sleep data, their use in high-acuity populations, such as patients undergoing rehabilitation, requires the validation and specific consideration of known risk factors for cognitive decline, such as age and cardiovascular disease (CVD) [15]. With the goal of 1 day operationalizing sleep measures as digital endpoints reflective of cognitive health, we encourage researchers to take a thoughtful approach to processing sleep data, particularly in female populations, as hormonal shifts across the lifespan contribute to evolving sleep chronotypes. Considering that women have fewer modifiable risk factors for dementia than men but are at a higher risk of cognitive impairment, new strategies to promote cognitive health in women are urgently needed [16,17]. The application of sleep monitoring during telerehabilitation could be one approach to unlock new discoveries in women's cognitive health—if meaningful digital endpoints can be created.

Such insights would be exceptionally helpful for clinicians overseeing telerehabilitation cohorts, wherein subtle cognitive deficits might slowly emerge and not be accurately captured in surveys due to participant self-report bias or cognitive decline impacting self-management [18]. Rather than providing a comprehensive review of wearables in the telerehabilitation space in this viewpoint, we aim to emphasize the aspects of a few recent wearables studies involving sleep measures beyond sleep duration alone that may be translated into clinical trials and outcomes research, preparing investigators to formulate exploratory sleep digital endpoints related to cognition and women's health in telerehabilitation populations.

Cognitive Performance and Sleep Onset and Regularity in Women

One recent study by Swanson et al [19] examined relationships between cognitive performance and sleep measures in older adult women who enrolled in the Study of Women's Health Across the Nation. In the subsample of participants who completed the at-home wearables study during a follow-up visit in 2015 - 2016 (n=1177; mean age=65 y), participants were instructed to wear an Actiwatch-2 (Philips Respironics) on their nondominant wrist for 7 days to record critical sleep measures: timing (average midpoint from sleep onset to wake) and regularity (midpoint SD) [20]. Since the Actiwatch-2 uses accelerometer sensors to record movement, its sleep onset and regularity measures are derived from periods of nonmovement during nighttime [21]. Cognitive performance was measured via motor, visual, learning, and memory questions in these validated assessments: the East Boston Memory Test, Symbol Digit Modalities Test, and Digit Span Backwards exam.

Metabolic burden in participants was high, with 63.5% being hypertensive and 17.6% being diabetic. The majority of participants had 2 or more comorbidities (54.9%) and a waist circumference indicating central obesity (59.9%). Importantly, at baseline, the average participant was at intermediate risk for a heart attack or stroke within a decade (mean atherosclerotic cardiovascular disease risk score (ASCVD)=9%). The authors adjusted linear regression models for relevant covariates to explore associations between real-world sleep measures and cognitive performance, reporting β coefficients representing the change in cognitive measure for a 1-unit change in sleep measure. Irregular sleep timing was linked with improved working memory (β=.50; *P*=.004) and worse delayed (β=−.36; *P*=.006) and immediate verbal memory (β=−.29*; P*=.02). While late sleep timing, defined as a midpoint after 4:00 AM, was associated with decreased processing speed (β=−1.80; *P*=.008), early sleep timing (a midpoint before 2:00 AM) was associated with worse delayed verbal memory (β=−.37*; P*=.047). Notably, a sensitivity analysis revealed that the magnitude of the effect of sleep irregularity on working memory was greater in participants with hypertension (interaction β=−3.35, *P*=.04). In addition, as the magnitude of ASCVD risk increased, so did the strength of the association between early sleep timing and delayed verbal memory (interaction β=−8.83; *P*=.03), drawing attention to the impact of CVD risk factors on sleep and cognitive decline in older women.

These findings highlight the potential for wearable-derived sleep measures to elucidate cognitive health; however, they specifically focused on older adult women. Hormones influence sleep characteristics, particularly in women, and their associations, as well as their directionalities and magnitudes, may not generalize well to younger cohorts. These results support existing literature that finds irregular sleep associated with gray matter atrophy and increased CVD risk via high β-amyloid burden, both potential contributors to cognitive decline [22].

Only ASCVD and hypertension revealed significant interactions with sleep-cognition associations in Swanson et al's [19] study, both of which are linked with increasing white matter hyperintensity volume in middle age [23,24]. It is plausible that the lack of associations for some factors, such as diabetes, which is a known risk factor for dementia, represents an age-related baseline for otherwise "healthy" women. The circadian clock controls blood pressure and the hypothalamic-pituitary-adrenal axis, which regulates blood glucose levels. As age and dementia risk increase, the circadian clock weakens [25]. After hypertension emerges, vascular changes influence the development of white matter hyperintensities, amyloid-β deposits, and cerebrovascular disease–related atrophy, all contributing to cognitive decline [26]. Downstream, these changes, hypothalamic-pituitary-adrenal dysfunction, may, in some cases, lead to diabetes or central obesity via insulin resistance, which then modulates the association between some sleep measures and cognition [27,28]. Therefore, digital endpoints related to sleep might require adjusting for comorbid conditions beyond CVD, but such discussion is outside the scope of this viewpoint.

Sex-Specific Cardiovascular Disease Risk and Sleep Patterns

One method to further elucidate the link between sleep and cognitive health in women during telerehabilitation could involve the thoughtful consideration of CVD risk when designing sleep digital endpoints.

XSL•FO
RenderX

Expanding on the CVD results presented by Swanson et al [19], the results from Nikbakhtian et al [29] highlight the importance of considering sex-specific CVD markers when tracking sleep remotely. They investigated the role of sleep onset in CVD development using a large population cohort study, the UK Biobank (n=103,712; 57.9% female; aged 43 - 79 y), which captured 1 week of real-world sleep data via the wearable Axivity AX3 accelerometer (Open Lab, Newcastle University).

In contrast to male participants, who mostly experienced sleep onset later in the night, women usually experienced sleep onset between 10:00 and 11:00 PM, the sleep onset time associated with the lowest incidence of CVD. Adjusting for relevant risk factors, sleep regularity, and sleep duration, Cox proportional hazards models generated hazard ratios (HRs), summarizing the risk of CVD diagnosis occurring in a cohort stratified by sleep onset time. Here, the associations between sleep onset time and CVD risk persisted, yielding Cox proportional hazards ratios (HRs) of 1.25 (CI 1.02 - 1.52; *P*=.03), 1.24 (CI 1.10 - 1.39; *P*<.005), and 1.12 (CI 1.01 - 1.25; *P*=.04) for sleep onset times of >12:00 AM, <10:00 PM, and 11:00-11:59 PM, respectively. In sex-specific models, the associations between sleep onset times of <10:00 PM (HR=1.63; CI 1.20 - 2.21; *P*<.005) and >12:00 AM (HR=1.63; CI 1.20 - 2.21; *P*<.005) were more pronounced in female participants. In contrast, only the association between sleep onset time of <10:00 PM and CVD was significant in male participants.

CVD and Alzheimer disease, the most common form of dementia, often share markers of systemic pathogenesis, as amyloid-β deposits can accumulate in the heart muscle, vessels, and brain as the diseases progress [30]. Moreover, hypertension contributes to brain structural changes implicated in cognitive decline, a finding more pronounced in menopausal and post-menopausal women, partly due to hormonal changes [31]. Thus, particular attention should be paid to sex-specific CVD risk when developing sleep digital endpoints for cognitive health, as CVD risk factors might help predict cognitive decline.

In addition to plausible mechanisms by which CVD contributes to cognitive decline, systematic reviews have reinforced the need for the proactive management of cognitive health in patients with CVD [32]. For example, Eggermont et al [32] concluded that interventions stimulating cognitive function should be routinely considered when developing treatment protocols for patients with CVD. Specific to telerehabilitation, for which a limited number of studies on CVD and cognition are published (n=9), Dabbaghipour et al's [33] systematic review outlined the need for rehabilitation programs to track cognitive health with the goal of improving treatment adherence and quality of life in participants. Combined, these reviews highlight the need for cognitive health monitoring to consider CVD risk factors, especially hypertension, though neither review assessed the role of sleep. One potential way to operationalize CVD risk in the development of sleep digital endpoints might be to stratify endpoints based on hypertension severity, with patients experiencing worse hypertension in need of aggressive sleep interventions [31,34-37].

Cognitive Impairment and Pulmonary Function During Sleep

The links between pulmonary markers and cardiovascular as well as cognitive health are well documented [32,38]. Not only is low blood oxygen saturation associated with a high risk of cognitive impairment, but it is also implicated in CVD exacerbations, forcing changes in cardiovascular and brain structures that affect sleep patterns and quality [39,40]. One commonly studied pulmonary marker is peripheral capillary oxygen saturation (SpO_2). Readily recorded by photoplethysmography (PPG) sensors in many wearable devices, SpO_2 serves as an indirect measure of lung diffusion, quantifying the concentration of oxygen molecules transported in blood from inhalation [41].

Extending the results of Swanson et al's [19] work, Ding et al [42] used the SleepImage Ring to track SpO_2 during sleep. In this study of older adult participants (n=62; 67.7% women; mean age=74 y; 80.5% cognitively intact), intraclass correlation coefficients (ICCs), which measure agreement across multiple timepoints, were reported for measures recorded over a 3-day sleep monitoring period, with ICCs equal to or greater than 0.70 indicating reliability. Using SleepImage Ring data from real-world settings, mean SpO_2 during sleep (ICC range: 0.75 - 0.77) was stable over 3 nights in participants without cognitive impairment; however, this finding did not persist in participants with cognitive impairment (ICC range=0.68 - 0.83). In addition to sleep onset and regularity measures, mean SpO_2 during sleep may inform digital endpoints that help researchers differentiate between individuals with and without cognitive impairment.

The findings from another exploratory study conducted in clinical settings, with a middle-aged cohort of obstructive sleep apnea (OSA) patients (n=207; 44.4% female; mean age=49), extend the link in Ding et al's [42] study between pulmonary function during sleep and cognitive health [43]. Thorisdottir et al [43] used data from the Embla PSG device (Flaga) obtained over a 1-night observational period and scores obtained from the Rey Auditory Verbal Learning Test. Reporting β coefficients, they found that mean SpO_2 during sleep was significantly associated with both immediate recall (β=−.171; *P*<.022) and total recall (β=−.188; *P*<.007).

Considering the results of these studies and prior work demonstrating that verbal memory is impacted by blood oxygen saturation, it is plausible that nighttime SpO_2 might specifically facilitate the indirect observation of verbal memory [44-46]. Though the external validity of Thorisdottir et al's [43] findings beyond patients with OSA is unknown, the underdiagnosis of OSA in the general population is high, with estimates of 40% to 80% of patients with CVD also experiencing OSA [47,48]. As such, the results of sleep study analyses that exclude patients with an OSA diagnosis may inadvertently include those with undiagnosed OSA. Due to the limited number of large population studies with comparable wearable devices, a thoughtful approach to assessing the current sleep-cognition

XSL•FO
RenderX

literature available, on populations with and without OSA, is vital.

Ding et al's [42] study focused on older adults, while that of Thorisdottir et al [43] involved a middle-aged cohort. As such, the interpretation of SpO_2 measures might necessitate stratification by age, especially considering that Thorisdottir et al's [43] study was conducted in clinical settings and may not generalize to the real world. The findings by Tam et al [49] using the SleepImage Ring offer support for age-stratification of pulmonary measures from wearables when formulating endpoints. Tam et al [49] hypothesized that sleep quality in women decreases after menopause. They conducted the first 1-night study with composite variables derived from built-in accelerometer sensors to capture movement in addition to SpO_2 from the PPG sensor.

Tam et al [49] performed subgroup analyses with one-way ANOVA, comparing the means of different measures across age-stratified cohorts to detect a statistically significant difference. Tam et al [49] used the SleepImage Ring and proprietary measures derived directly from the ring's data stream. In this cohort (n=1444; 48.8% female; mean age=54), Tam et al [49] observed a drop in the Sleep Quality Index, a composite measure using both PPG and accelerometer sensor data to score sleep fragmentation, sleep quality, sleep stability, and sleep periodicity, in women 51 years of age and older. After the age of 50, women also experienced a greater increase than men in the apnea-hypopnea index, which uses SpO_2 to quantify shallow and paused breathing episodes, as well as the Arousal Index, which relies on accelerometer data to quantify shifts from deep sleep to near wakefulness. Though the authors did not investigate cognitive health in participants across stratified age groups, future researchers could address this gap to better understand if composite sleep measures offer new insights beyond unidimensional measures, such as sleep duration, into cognitive changes across the lifespan [49]. Furthermore, an increase in the Arousal index is associated with a higher risk of CVD, suggesting that an age of 50 years might be an appropriate threshold for researchers to explore for adjusting sleep digital endpoints for CVD risk factors [50].

Multidimensional View of Cognitive Health and Sleep

The potential for these wearable-derived endpoints, should they prove to be more clinically meaningful than conventional self-reported sleep endpoints, to provide more accurate analyses can advance the design of remote monitoring programs in telerehabilitation and beyond. They can also inform the development of more effective sleep interventions targeting populations with high comorbid burden, such as telerehabilitation patients.

To date, the only peer-reviewed multivariate analysis of sleep and cognition necessitated a multi-step statistical analysis. Qin et al [51] used the Oura Ring (Oura) for 14 - 28 days to capture 23 sleep measures and investigate associations between sleep measures and cognitive domains in adults who participated in the Singapore Chinese Health Study. A total of 7 cognitive domains (verbal memory, attention, visuospatial ability, visual memory, language, executive function, and processing speed) were assessed via the following validated instruments: Rey Auditory Verbal Learning Test, Boston Naming Test, Associative Learning Test, Brief Visuospatial Memory Test-Revised, Color Trail Tests 1 and 2, Design Fluency Test, WAIS-III, and Symbol Digit Modality Test. First, scores were standardized from each test to T scores (mean 50, SD 10). Then, for cognitive domains with more than 1 test score, the reported score was summarized as the average of all test scores for that domain.

After adjusting for confounding variables, including age and sex, Qin et al [51] applied a series of statistical techniques to identify significant associations between sleep and cognitive health. First, they used partial least squares correlation (PLSC), a statistical technique using 2 matrices—in this case, a matrix of multiple sleep measures and a matrix of multiple cognition scores—to highlight optimal relationships between sleep, as a multidimensional concept, and cognition, also as a multidimensional concept. Their analysis of data obtained from adult Chinese participants over the age of 65 years (n=773; 51.1% female; mean age=75; 1.2% with cognitive impairment) found that the first PLSC component contributed the majority of covariance (82%) between cognition and sleep scores (r=0.2; P<.001).

After generating the covariance matrix, they applied singular value decomposition to generate latent variables representing the maximum covariance between sleep and cognitive measures. For each latent variable with a significant correlation value (Pearson's r), 5000 bootstrap tests were run to identify measures contributing to the first PLSC component. They identified robust contributions to the first PLSC component for 11 sleep measures and 3 cognitive domains. Qin et al [51] found 8 sleep regularity measures (Sleep Regularity Index [SRI], Sleep Fragmentation Index [SFI] Intraindividual Standard Deviation [iSD], Wake-After-Sleep-Onset (WASO) iSD, Efficiency iSD, Total Sleep Time iSD, Time in Bed iSD, Wake Time iSD, Bedtime iSD) and 3 continuity measures (Efficiency, WASO, SFI) contributed to the sleep-cognition relationship. All but SRI (which tracked whether an individual was awake or asleep at the same time each day), SFI (which quantified nighttime movement and sleep epochs lasting less than 1 min), and SFI iSD were extracted directly from the Oura dataset. Executive function, verbal memory, and processing speed were identified as relevant cognitive domains, each of which is associated with age-related decline [52].

Finally, to probe for associations between specific sleep measures and cognitive domains, a partial correlation analysis was performed, identifying only weak correlation coefficients between sleep measures and cognitive domains. Among sleep regularity measures, a higher SRI was associated with better processing speed (r=0.17) and improved executive function (r=0.17). Additionally, as sleep efficiency (efficiency iSD) was less regular, processing speed declined (r=0.15). In contrast, among sleep continuity measures, only high sleep efficiency was associated with better processing speed (r=0.11).

XSL•FO
RenderX

Using multidimensional, objective sleep measures, Qin et al's [51] results represent a significant step forward in elucidating the relationship between sleep and cognition; however, their analyses did not include sleep measures derived from relevant biomarkers, like blood pressure or pulmonary function, highlighting a gap in the literature for future researchers to fill.

Challenges and Recommendations

Despite the use cases for sleep digital endpoints in telerehabilitation highlighted above, no systematic reviews or meta-analyses have been published on this topic. The disparate study methods, device types, and sleep measures used contribute to the lack of evidence synthesis on the topic. As such, the statistical significance, magnitude, and directionality of the findings presented in this viewpoint should be interpreted as exploratory and not absolute, with future research required to use these measures in practice and real-world settings. Fundamentally, to develop useful sleep digital endpoints, robust and reliable associations between sleep measures and cognition must be established.

This viewpoint highlights multiple trends in the sleep-cognition space to achieve this goal. First, more large cohort studies using wearables that collect multidimensional data from across relevant biomarkers—including sleep biomarkers like SpO_2 in addition to sleep timing measures—must be conducted. Longitudinal sleep data captured beyond 1-week periods should be prioritized [53-55]. Accurate longitudinal monitoring is especially relevant when conducting aging research in cognitive impairment and dementia, when prevalence rates change drastically over short epochs. For instance, dementia prevalence is <1% in adults under the age of 65 years [56]. After the age of 65 years, that number shifts to between 3% and 11% and climbs to over 30% in adults over 85 years of age [56]. When reviewing studies with less than 1 week of data, the results should be interpreted cautiously, as the limited samples may not be representative of an individual's sleep routine. Enrolling larger cohorts for longer periods will facilitate the development of standardized protocols for processing sleep wearable data and, subsequently, validating digital endpoints and related thresholds in telerehabilitation populations [57].

Second, across the recent studies presented in this piece, measures of sleep regularity were the most frequently associated with cognitive health. Therefore, we encourage researchers to prioritize studying sleep regularity across the lifespan. Considering that 2 individuals could sleep for the same duration during the night while experiencing separate disturbances, sleep regularity measures quantify individual sleep habits across days more sufficiently than duration measures alone.

Finally, the findings outlined in this viewpoint offer an exploratory rationale for the development of sleep digital endpoints stratified by age. With stark changes in pulmonary function and cardiovascular risk linked to sleep patterns in menopausal women, age might serve as a proxy for changes in hormone concentration. Alternatively, the inclusion of blood biomarker data to stratify sleep digital endpoints might offer more precise monitoring in telerehabilitation programs. Regardless, establishing thresholds relying on sleep data to bifurcate healthy versus impaired cognition necessitates a closer look by researchers. These issues are complicated by the fact that individual sleep chronotypes change across the lifespan, necessitating robust methods that account for within-person variability, rather than relying exclusively on population-based thresholds [58]. Such an approach is appropriate for dynamic monitoring in real-world settings, automatically accounting for variability in the environment.

Our viewpoint outlines multiple opportunities for researchers to advance knowledge in the pursuit of developing sleep digital endpoints in telerehabilitation programs. Beyond the lack of systematic reviews and meta-analyses assessing sleep measures relevant to telerehabilitation, no research into cognitive health and SpO_2 measures during sleep stratified by age has been published. Additionally, no multivariate analyses assessing sleep measures and cognitive measures have included nighttime biomarker measures, like SpO_2 or blood pressure levels. Of note, for studies exclusively focused on older adults, the associations between sleep measures and cognitive domains might be subject to survival bias and require additional investigation to confirm results across age cohorts. Addressing these gaps will create a starting point for more robust research into the sleep-cognition relationship, advancing the promise of wearables.

These studies may have implications for the future of sleep monitoring research in detecting cognitive decline in women. First, the influence of cardiovascular and pulmonary risk factors on the association between sleep and cognition underscores the importance of considering comorbid burden when interpreting cognitive health data [59]. Next, these remote monitoring approaches can reach women in rural areas who might otherwise be unable to participate in in-clinic monitoring, thereby addressing barriers and adherence issues by introducing enhanced flexibility in rehabilitation care [60].

Beyond the potential of monitoring research on sleep and cognition, numerous real-world barriers limit the implementation of sleep digital endpoints, even when they are validated in research. Historically, most validation research on wearable device measures was performed using research-grade devices. In recent years, with the growing acceptance of consumer-grade devices, the shift to validating consumer-grade devices has introduced new complexities. For instance, consumer-grade wearables often generate proprietary composite measures with limited documentation. Also, the interpretation of results derived from wearable devices requires special attention paid to the device version, as software and hardware updates might introduce unknown variability in the data collected [61].

Another barrier to real-world implementation of sleep digital endpoints is devising feasible strategies to integrate wearable data streams into clinical and research workflows. Should researchers pinpoint meaningful sleep measures and thresholds for digital endpoints tracking cognitive health in women, the data collection, wrangling, and analysis processes are resource intensive. If a feasible and cybersecure data pipeline infrastructure for wearables is designed and provides actionable information to clinical and research teams, the demand for evidence-based interventions to help participants will increase

XSL•FO
RenderX

[62,63]. For telerehabilitation participants, this might require interventions that can be delivered outside of clinical settings, from home health visits and/or digital tools. Recently, novel applications of digitally delivered cognitive behavioral therapy have shown promise with regard to improving sleep measures in patients. Cognitive behavioral therapy interventions might confer a protective effect on cognitive health during telerehabilitation, one day being measured by digital endpoints [64-66]. Virtual reality–based support tools and virtual companions could also be offered in tandem with conventional telerehabilitation programs [67-69]. Notably, unclear reimbursement and payment pathways for remote monitoring of sleep, particularly in the United States, remain a key bottleneck to its widespread adoption [70].

The application of artificial intelligence (AI), such as machine learning methods, to swiftly detect trends in patient sleep and cognition can augment workflows and make the implementation of digital endpoints a reality. For sleep digital endpoint implementations that rely on AI to review 24/7 data from patients, threshold levels will likely evolve, as more data are fed into databases on which AI is trained [71]. This requires additional review and analysis of evolving AI tools. In addition to the workflow barriers, identifying patients who want to participate in sleep monitoring, as well as determining how to troubleshoot tech problems remotely, can be challenging. At the time of this publication, wearable monitoring research over longitudinal periods appears promising in older adults, even in those with cognitive impairment and dementia [28,72].

Multiple study design limitations complicate the interpretation of sleep data from wearables. First, the studies surveyed in this viewpoint reflect significant heterogeneity with regard to methodologies and primary endpoints [73-75]. Therefore, deriving actionable information for women's health from objective sleep data remains elusive. Also, the proportion of participants in sleep studies with wearables has been overwhelmingly male (61%) [76]. Next, the location of wear for wearables, such as wrist versus ring finger, is sometimes limited during rehabilitation due to functional deficits and can influence results. Since the body of literature in this space is continually growing, there is a need for reviews summarizing and comparing the results by location of wear, particularly in female participants.

Summarizing the recent studies outlined in this viewpoint, we recommend that researchers consider the sleep measures and cognitive variables in Table 1 as a starting point to design digital endpoints from measures obtained via wearable sensors. Researchers should also consider the impact of age and CVD risk factors, including hypertension, as they develop clinically meaningful thresholds specific to female patients. Theoretically, a rudimentary sleep digital endpoint for cognitive health in female patients between the ages of 50 and 60 years old without hypertension might establish a target sleep onset between 10 PM and 12 AM with stable mean SpO_2 during sleep 7 days per week. This outcome of interest incorporates multidimensional sleep measures (sleep onset and SpO_2), adjusted for hypertensive status and age, setting an optimal threshold for monitoring cognitive health indirectly. To be operationalized in clinical settings, values that fall outside the aforementioned thresholds might trigger clinician review. While plausible based on the findings reviewed in this viewpoint, this digital endpoint is purely hypothetical. Thus, research beyond what is currently available is required to formulate sleep digital endpoints in light of comorbid burden.

The list of measures in Table 1 is by no means exhaustive, with new digital measures ripe for application to tracking cognitive health emerging every year. For example, 1 exploratory sleep measure to investigate across the lifespan is nighttime movement by age. In an illustrative example using an arm-worn sensor (Everion) to track sleeping habits in patients with multiple sclerosis, Moebus et al [77] found that age correlated strongly with movement during the night, a measure unstudied in the context of female cognitive health and telerehabilitation. Other novel measures include sound-based measures of sleep, using a smartphone-compatible microphone, and applying AI algorithms to classify sleep stages in real-world settings [78].

Table . Sleep measures and related cognitive variables outlined in recent research.

Wearable-derived sleep measure	Cognitive variable (instrument)
Sleep onset (time) and regularity (standard deviation)	• Working memory (Digit Span Backward) • Verbal memory (East Boston Memory Test) • Processing speed (Symbol Digit Modalities Test)
SpO_2[a] (mean, min, max)	• Cognitive impairment (Montreal Cognitive Assessment) • Rey auditory verbal learning
Sleep Regularity Index (Oura-specific)	• Processing speed (Colour Trail Test 1; Symbol Digit Modality Test)
Sleep efficiency	• Executive function (Colour Trail Test 2; Design Fluency Test)

[a]SpO_2: oxygen saturation.

Overall, these studies build upon current literature by highlighting potentially modifiable sleep measures, such as sleep onset, sleep regularity, sleep continuity, or SpO_2 during sleep, that can inform the development of digital endpoints for cognitive health. The age- and sex-specific nature of the results as well as CVD considerations outlined provide an early-stage rationale for developing more personalized digital endpoints for sleep stratified by known risk factors for cognitive decline [37].

XSL•FO
RenderX

Conclusions

Historically, telerehabilitation programs have relied on physical activity endpoints to monitor and evaluate participants in real-world settings; however, wearable-derived sleep measures might augment conventional telerehabilitation protocols by providing insights into cognitive health, addressing a critical gap for older female participants who are at higher risk of cognitive decline. Ultimately, the studies outlined above offer an early-stage rationale for investigating emerging sleep measures as multidimensional endpoints for cognitive health in female patients. We recommend that the contributions of sleep and nighttime pulmonary measures to cognitive decline in women be explored further using wearables and adjusting for cardiovascular factors across middle and late adulthood. As no systematic reviews or meta-analyses have been published on this topic, the specific findings presented in this viewpoint should be interpreted circumspectly. Nonetheless, the combined findings demonstrate how researchers might hypothetically apply real-world sleep data from wearables to develop clinically meaningful digital endpoints of cognitive health for women, while being mindful of adjustments necessary to account for comorbid burden. By addressing the gaps in knowledge outlined in this viewpoint, researchers can make strides in transforming sleep measures into digital endpoints for cognitive health interventions, enabling proactive care at home to delay or stop the onset of cognitive decline in women during critical periods, such as rehabilitation.

Acknowledgments

We did not use artificial intelligence to produce or author this manuscript.

Funding

The funding for this study was provided by NIH R56 HL173214 and the Robert D and Patricia E Kern Center for the Science of Health Delivery. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

SJZ is an associate editor of *JMIR Neurotechnology* at the time of this publication. EF and LF have no conflicts of interest to report.

References

1. Leyens L, Batchelor J, De Beuckelaer E, Langel K, Hartog B. Unlocking the full potential of digital endpoints for decision making: a novel modular evidence concept enabling re-use and advancing collaboration. Expert Rev Pharmacoecon Outcomes Res 2024 Jul;24(6):731-741. [doi: 10.1080/14737167.2024.2334347] [Medline: 38747565]
2. Miller A, Collier Z, Reisman DS. Beyond steps per day: other measures of real-world walking after stroke related to cardiovascular risk. J Neuroeng Rehabil 2022 Oct 14;19(1):111. [doi: 10.1186/s12984-022-01091-7] [Medline: 36242083]
3. Armstrong M, Winnard A, Chynkiamis N, Boyle S, Burtin C, Vogiatzis I. Use of pedometers as a tool to promote daily physical activity levels in patients with COPD: a systematic review and meta-analysis. Eur Respir Rev 2019 Dec 31;28(154):190039. [doi: 10.1183/16000617.0039-2019] [Medline: 31722891]
4. Zhong W, Liu R, Cheng H, et al. Longer-Term effects of cardiac telerehabilitation on patients with coronary artery disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2023 Jul 28;11:e46359. [doi: 10.2196/46359] [Medline: 37505803]
5. Udeh-Momoh C, Watermeyer T, Female Brain Health and Endocrine Research (FEMBER) consortium. Female specific risk factors for the development of Alzheimer's disease neuropathology and cognitive impairment: call for a precision medicine approach. Ageing Res Rev 2021 Nov;71:101459. [doi: 10.1016/j.arr.2021.101459] [Medline: 34508876]
6. Beverly Hery CM, Hale L, Naughton MJ. Contributions of the Women's Health Initiative to understanding associations between sleep duration, insomnia symptoms, and sleep-disordered breathing across a range of health outcomes in postmenopausal women. Sleep Health 2020 Feb;6(1):48-59. [doi: 10.1016/j.sleh.2019.09.005] [Medline: 31699635]
7. Jedrziewski MK, Lee VMY, Trojanowski JQ. Physical activity and cognitive health. Alzheimers Dement 2007 Apr;3(2):98-108. [doi: 10.1016/j.jalz.2007.01.009] [Medline: 18379636]
8. Cormack F, McCue M, Skirrow C, et al. Characterizing longitudinal patterns in cognition, mood, and activity in depression with 6-week high-frequency wearable assessment: observational study. JMIR Ment Health 2024 May 31;11:e46895. [doi: 10.2196/46895] [Medline: 38819909]
9. Cacciante L, Pietà CD, Rutkowski S, et al. Cognitive telerehabilitation in neurological patients: systematic review and meta-analysis. Neurol Sci 2022 Feb;43(2):847-862. [doi: 10.1007/s10072-021-05770-6] [Medline: 34822030]
10. Leong RLF, Chee MWL. Understanding the need for sleep to improve cognition. Annu Rev Psychol 2023 Jan 18;74:27-57. [doi: 10.1146/annurev-psych-032620-034127] [Medline: 35961036]

XSL•FO
RenderX

11. Cudney LE, Frey BN, McCabe RE, Green SM. Investigating the relationship between objective measures of sleep and self-report sleep quality in healthy adults: a review. J Clin Sleep Med 2022 Mar 1;18(3):927-936. [doi: 10.5664/jcsm.9708] [Medline: 34609276]
12. Ahn EK, Yoon K, Park JE. Association between sleep hours and changes in cognitive function according to the morningness-eveningness type: a population-based study. J Affect Disord 2024 Jan 15;345:112-119. [doi: 10.1016/j.jad.2023.10.122] [Medline: 37865346]
13. Qin S, Leong RLF, Ong JL, Chee MWL. Associations between objectively measured sleep parameters and cognition in healthy older adults: a meta-analysis. Sleep Med Rev 2023 Feb;67:101734. [doi: 10.1016/j.smrv.2022.101734] [Medline: 36577339]
14. Jones AK, Yan CL, Rivera Rodriquez BP, Kaur S, Andrade-Bucknor S. Role of wearable devices in cardiac telerehabilitation: a scoping review. PLoS One 2023;18(5):e0285801. [doi: 10.1371/journal.pone.0285801] [Medline: 37256878]
15. Doherty C, Baldwin M, Keogh A, Caulfield B, Argent R. Keeping pace with wearables: a living umbrella review of systematic reviews evaluating the accuracy of consumer wearable technologies in health measurement. Sports Med 2024 Nov;54(11):2907-2926. [doi: 10.1007/s40279-024-02077-2] [Medline: 39080098]
16. Martin J, Gordon EH, Reid N, Hubbard RE, Ward DD. Sex differences in modifiable dementia risk factors: findings from the Rush Memory and Aging Project. Alzheimers Dement 2025 Jul;21(7):e70506. [doi: 10.1002/alz.70506] [Medline: 40696813]
17. Moutinho S. Women twice as likely to develop Alzheimer's disease as men—but scientists do not know why. Nat Med 2025 Mar;31(3):704-707. [doi: 10.1038/s41591-025-03564-3] [Medline: 40087515]
18. Van Der Donckt J, Vandenbussche N, Van Der Donckt J, et al. Mitigating data quality challenges in ambulatory wrist-worn wearable monitoring through analytical and practical approaches. Sci Rep 2024 Jul 30;14(1):17545. [doi: 10.1038/s41598-024-67767-3] [Medline: 39079945]
19. Swanson LM, Hood MM, Thurston RC, et al. Sleep timing, sleep timing regularity, and cognitive performance in women entering late adulthood: the Study of Women's Health Across the Nation (SWAN). Sleep 2025 May 12;48(5):zsaf041. [doi: 10.1093/sleep/zsaf041] [Medline: 39955263]
20. Santoro N, Sutton-Tyrrell K. The SWAN song: Study of Women's Health Across the Nation's recurring themes. Obstet Gynecol Clin North Am 2011 Sep;38(3):417-423. [doi: 10.1016/j.ogc.2011.05.001] [Medline: 21961710]
21. Actigraph watch for sleep assessment (PCS3). Carnegie Mellon University. URL: https://www.cmu.edu/common-cold-project/measures-by-study/health-practices/actigraph-watch-sleep-assessment-pcs3.html [accessed 2025-12-09]
22. Qin S, Chee MWL. The emerging importance of sleep regularity on cardiovascular health and cognitive impairment in older adults: a review of the literature. Nat Sci Sleep 2024;16:585-597. [doi: 10.2147/NSS.S452033] [Medline: 38831959]
23. d'Arbeloff T, Elliott ML, Knodt AR, et al. White matter hyperintensities are common in midlife and already associated with cognitive decline. Brain Commun 2019;1(1):fcz041. [doi: 10.1093/braincomms/fcz041] [Medline: 31894208]
24. Pålhaugen L, Sudre CH, Tecelao S, et al. Brain amyloid and vascular risk are related to distinct white matter hyperintensity patterns. J Cereb Blood Flow Metab 2021 May;41(5):1162-1174. [doi: 10.1177/0271678X20957604] [Medline: 32955960]
25. Hood S, Amir S. The aging clock: circadian rhythms and later life. J Clin Invest 2017 Feb 1;127(2):437-446. [doi: 10.1172/JCI90328] [Medline: 28145903]
26. Barone MTU, Menna-Barreto L. Diabetes and sleep: a complex cause-and-effect relationship. Diabetes Res Clin Pract 2011 Feb;91(2):129-137. [doi: 10.1016/j.diabres.2010.07.011] [Medline: 20810183]
27. Wingenfeld K, Wolf OT. HPA axis alterations in mental disorders: impact on memory and its relevance for therapeutic interventions. CNS Neurosci Ther 2011 Dec;17(6):714-722. [doi: 10.1111/j.1755-5949.2010.00207.x] [Medline: 21143429]
28. Gabb VG, Blackman J, Morrison H, et al. Longitudinal remote sleep and cognitive research in older adults with mild cognitive impairment and dementia: prospective feasibility cohort study. JMIR Aging 2025 May 28;8:e72824. [doi: 10.2196/72824] [Medline: 40435500]
29. Nikbakhtian S, Reed AB, Obika BD, et al. Accelerometer-derived sleep onset timing and cardiovascular disease incidence: a UK Biobank cohort study. Eur Heart J Digit Health 2021 Dec;2(4):658-666. [doi: 10.1093/ehjdh/ztab088] [Medline: 36713092]
30. Troncone L, Luciani M, Coggins M, et al. Aβ amyloid pathology affects the hearts of patients with Alzheimer's disease: mind the heart. J Am Coll Cardiol 2016 Dec 6;68(22):2395-2407. [doi: 10.1016/j.jacc.2016.08.073] [Medline: 27908343]
31. Liu L, Hayden KM, May NS, et al. Association between blood pressure levels and cognitive impairment in older women: a prospective analysis of the Women's Health Initiative Memory Study. Lancet Healthy Longev 2022 Jan;3(1):e42-e53. [doi: 10.1016/s2666-7568(21)00283-x] [Medline: 35112096]
32. Eggermont LHP, de Boer K, Muller M, Jaschke AC, Kamp O, Scherder EJA. Cardiac disease and cognitive impairment: a systematic review. Heart 2012 Sep;98(18):1334-1340. [doi: 10.1136/heartjnl-2012-301682] [Medline: 22689718]
33. Dabbaghipour N, Javaherian M, Moghadam BA. Effects of cardiac rehabilitation on cognitive impairments in patients with cardiovascular diseases: a systematic review. Int J Neurosci 2021 Nov;131(11):1124-1132. [doi: 10.1080/00207454.2020.1773823] [Medline: 32449872]

XSL•FO
RenderX

34. Yano Y, Inokuchi T, Hoshide S, Kanemaru Y, Shimada K, Kario K. Association of poor physical function and cognitive dysfunction with high nocturnal blood pressure level in treated elderly hypertensive patients. Am J Hypertens 2011 Mar;24(3):285-291. [doi: 10.1038/ajh.2010.224] [Medline: 21088668]
35. Yu JH, Kim REY, Park SY, et al. Night blood pressure variability, brain atrophy, and cognitive decline. Front Neurol 2022;13:963648. [doi: 10.3389/fneur.2022.963648] [Medline: 36119712]
36. Yano Y, Butler KR, Hall ME, et al. Associations of nocturnal blood pressure with cognition by self-identified race in middle-aged and older adults: the GENOA (Genetic Epidemiology Network of Arteriopathy) study. J Am Heart Assoc 2017 Oct 27;6(11):e007022. [doi: 10.1161/JAHA.117.007022] [Medline: 29079569]
37. Wang Y, Zhou X, Guo Z, Fang X, Liu F, Shen L. Consideration of stratification in confirmatory trials with time-to-event endpoint. Contemp Clin Trials 2024 Jun;141:107434. [doi: 10.1016/j.cct.2024.107434] [Medline: 38215875]
38. Duggan EC, Graham RB, Piccinin AM, et al. Systematic review of pulmonary function and cognition in aging. J Gerontol B Psychol Sci Soc Sci 2020 Apr 16;75(5):937-952. [doi: 10.1093/geronb/gby128] [Medline: 30380129]
39. Thakur N, Blanc PD, Julian LJ, et al. COPD and cognitive impairment: the role of hypoxemia and oxygen therapy. Int J Chron Obstruct Pulmon Dis 2010 Sep 7;5:263-269. [doi: 10.2147/copd.s10684] [Medline: 20856825]
40. James SK, Erlinge D, Herlitz J, et al. Effect of oxygen therapy on cardiovascular outcomes in relation to baseline oxygen saturation. JACC Cardiovasc Interv 2020 Feb 24;13(4):502-513. [doi: 10.1016/j.jcin.2019.09.016] [Medline: 31838113]
41. Tamura T. Current progress of photoplethysmography and SPO2 for health monitoring. Biomed Eng Lett 2019 Feb;9(1):21-36. [doi: 10.1007/s13534-019-00097-w] [Medline: 30956878]
42. Ding H, Madan S, Searls E, et al. Exploring nightly variability and clinical influences on sleep measures: insights from a digital brain health platform. Sleep Med 2025 Jul;131:106532. [doi: 10.1016/j.sleep.2025.106532] [Medline: 40306226]
43. Thorisdottir K, Hrubos-Strøm H, Karhu T, et al. Verbal memory is linked to average oxygen saturation during sleep, not the apnea-hypopnea index nor novel hypoxic load variables. Sleep Med 2024 Nov;123:29-36. [doi: 10.1016/j.sleep.2024.08.028] [Medline: 39232262]
44. Hrubos-Strøm H, Nordhus IH, Einvik G, et al. Obstructive sleep apnea, verbal memory, and executive function in a community-based high-risk population identified by the Berlin Questionnaire Akershus Sleep Apnea Project. Sleep Breath 2012 Mar;16(1):223-231. [doi: 10.1007/s11325-011-0493-1] [Medline: 21350844]
45. Alomri RM, Kennedy GA, Wali SO, Ahejaili F, Robinson SR. Differential associations of hypoxia, sleep fragmentation, and depressive symptoms with cognitive dysfunction in obstructive sleep apnea. Sleep 2021 Apr 9;44(4):zsaa213. [doi: 10.1093/sleep/zsaa213] [Medline: 33045082]
46. Kainulainen S, Duce B, Korkalainen H, et al. Severe desaturations increase psychomotor vigilance task-based median reaction time and number of lapses in obstructive sleep apnoea patients. Eur Respir J 2020 Apr;55(4):1901849. [doi: 10.1183/13993003.01849-2019] [Medline: 32029446]
47. Motamedi KK, McClary AC, Amedee RG. Obstructive sleep apnea: a growing problem. Ochsner J 2009;9(3):149-153. [Medline: 21603432]
48. Yeghiazarians Y, Jneid H, Tietjens JR, et al. Obstructive sleep apnea and cardiovascular disease: a scientific statement from the American Heart Association. Circulation 2021 Jul 20;144(3):e56-e67. [doi: 10.1161/CIR.0000000000000988] [Medline: 34148375]
49. Tam J, Ferri R, Mogavero MP, Palomino M, DelRosso LM. Sex-specific changes in sleep quality with aging: insights from wearable device analysis. J Sleep Res 2025 Aug;34(4):e14413. [doi: 10.1111/jsr.14413] [Medline: 39543848]
50. Shahrbabaki SS, Linz D, Hartmann S, Redline S, Baumert M. Sleep arousal burden is associated with long-term all-cause and cardiovascular mortality in 8001 community-dwelling older men and women. Eur Heart J 2021 Jun 1;42(21):2088-2099. [doi: 10.1093/eurheartj/ehab151] [Medline: 33876221]
51. Qin S, Ng EKK, Soon CS, et al. Association between objectively measured, multidimensional sleep health and cognitive function in older adults: cross-sectional wearable tracker study. Sleep Med 2025 Aug;132:106569. [doi: 10.1016/j.sleep.2025.106569] [Medline: 40393112]
52. Forbes M, Lotfaliany M, Mohebbi M, et al. Depressive symptoms and cognitive decline in older adults. Int Psychogeriatr 2024 Nov;36(11):1039-1050. [doi: 10.1017/S1041610224000541] [Medline: 38623851]
53. Druiven SJM, Riese H, Kamphuis J, et al. Chronotype changes with age; seven-year follow-up from the Netherlands study of depression and anxiety cohort. J Affect Disord 2021 Dec 1;295:1118-1121. [doi: 10.1016/j.jad.2021.08.095] [Medline: 34706423]
54. Carpi M, Fernandes M, Mercuri NB, Liguori C. Sleep biomarkers for predicting cognitive decline and Alzheimer's disease: a systematic review of longitudinal studies. J Alzheimers Dis 2024;97(1):121-143. [doi: 10.3233/JAD-230933] [Medline: 38043016]
55. Evans MA, Buysse DJ, Marsland AL, et al. Meta-analysis of age and actigraphy-assessed sleep characteristics across the lifespan. Sleep 2021 Sep 13;44(9):zsab088. [doi: 10.1093/sleep/zsab088] [Medline: 33823052]
56. National Institutes of Health. 2023 Alzheimer's disease facts and figures. Alzheimers Dement 2023 Apr;19(4):1598-1695. [doi: 10.1002/alz.13016]
57. Baumert M, Cowie MR, Redline S, et al. Sleep characterization with smart wearable devices: a call for standardization and consensus recommendations. Sleep 2022 Dec 12;45(12):zsac183. [doi: 10.1093/sleep/zsac183] [Medline: 35913733]

XSL•FO
RenderX

58. Foy BH, Petherbridge R, Roth MT, et al. Haematological setpoints are a stable and patient-specific deep phenotype. Nature New Biol 2025 Jan 9;637(8045):430-438. [doi: 10.1038/s41586-024-08264-5]
59. Zawada SJ, Ganjizadeh A, Conte GM, Demaerschalk BM, Erickson BJ. Accelerometer-measured behavior patterns in incident cerebrovascular disease: insights for preventative monitoring from the UK Biobank. J Am Heart Assoc 2024 Jun 4;13(11):e032965. [doi: 10.1161/JAHA.123.032965] [Medline: 38818948]
60. Fischer MJ, Scharloo M, Abbink JJ, et al. Participation and drop-out in pulmonary rehabilitation: a qualitative analysis of the patient's perspective. Clin Rehabil 2007 Mar;21(3):212-221. [doi: 10.1177/0269215506070783] [Medline: 17329278]
61. Collins T, Woolley SI, Oniani S, et al. Version reporting and assessment approaches for new and updated activity and heart rate monitors. Sensors (Basel) 2019 Apr 10;19(7):1705. [doi: 10.3390/s19071705] [Medline: 30974755]
62. Abbasi AB, Curtis LH, Califf RM. The promise of real-world data for research—what are we missing? N Engl J Med 2025 Jul 24;393(4):318-321. [doi: 10.1056/NEJMp2416479] [Medline: 40689459]
63. Silva-Trujillo AG, González González MJ, Rocha Pérez LP, García Villalba LJ. Cybersecurity analysis of wearable devices: smartwatches passive attack. Sensors (Basel) 2023 Jun 8;23(12):5438. [doi: 10.3390/s23125438] [Medline: 37420605]
64. Kyle SD, Hurry MED, Emsley R, et al. The effects of digital cognitive behavioral therapy for insomnia on cognitive function: a randomized controlled trial. Sleep 2020 Sep 14;43(9):zsaa034. [doi: 10.1093/sleep/zsaa034] [Medline: 32128593]
65. Harvey PD. Digital therapeutics to enhance cognition in major depression: how can we make the cognitive gains translate into functional improvements? Am J Psychiatry 2022 Jul;179(7):445-447. [doi: 10.1176/appi.ajp.20220441] [Medline: 35775161]
66. Prather AA, Krystal AD, Emsley R, et al. The effectiveness of digital cognitive behavioral therapy to treat insomnia disorder in US adults: nationwide decentralized randomized controlled trial. JMIR Ment Health 2025 Dec 4;12:e84323. [doi: 10.2196/84323] [Medline: 41343796]
67. Prats-Bisbe A, López-Carballo J, García-Molina A, et al. Virtual reality–based neurorehabilitation support tool for people with cognitive impairments resulting from an acquired brain injury: usability and feasibility study. JMIR Neurotech 2024;3:e50538. [doi: 10.2196/50538]
68. Kokorelias KM, McMurray J, Chu C, et al. Technology-enabled recreation and leisure programs and activities for older adults with cognitive impairment: rapid scoping review. JMIR Neurotech 2024;3:e53038. [doi: 10.2196/53038]
69. Portacolone E, Feddoes DE. Should artificial intelligence play a role in cultivating social connections among older adults? AMA J Ethics 2023 Nov 1;25(11):E818-E824. [doi: 10.1001/amajethics.2023.818] [Medline: 38085584]
70. Hwang D, Melius BN. The affordable care act and the future of sleep medicine. In: Narcolepsy: A Clinical Guide: Springer; 2016:417-435. [doi: 10.1007/978-3-319-23739-8_31]
71. Postnova S. Optimizing alertness in a 24-hour society: the role of personalized digital tools. Sleep 2025 Nov 10;48(11):zsaf185. [doi: 10.1093/sleep/zsaf185] [Medline: 40652314]
72. Hackett K, Xu S, McKniff M, Paglia L, Barnett I, Giovannetti T. Mobility-based smartphone digital phenotypes for unobtrusively capturing everyday cognition, mood, and community life-space in older adults: feasibility, acceptability, and preliminary validity study. JMIR Hum Factors 2024 Nov 22;11:e59974. [doi: 10.2196/59974] [Medline: 39576984]
73. Berryhill S, Morton CJ, Dean A, et al. Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. J Clin Sleep Med 2020 May 15;16(5):775-783. [doi: 10.5664/jcsm.8356] [Medline: 32043961]
74. Wallace ML, Lee S, Stone KL, et al. Actigraphy-derived sleep health profiles and mortality in older men and women. Sleep 2022 Apr 11;45(4):zsac015. [doi: 10.1093/sleep/zsac015] [Medline: 35037946]
75. Anchouche K, Elharram M, Oulousian E, et al. Use of actigraphy (wearable digital sensors to monitor activity) in heart failure randomized clinical trials: a scoping review. Can J Cardiol 2021 Sep;37(9):1438-1449. [doi: 10.1016/j.cjca.2021.07.001] [Medline: 34256087]
76. Nebel RA, Aggarwal NT, Barnes LL, et al. Understanding the impact of sex and gender in Alzheimer's disease: a call to action. Alzheimers Dement 2018 Sep;14(9):1171-1183. [doi: 10.1016/j.jalz.2018.04.008] [Medline: 29907423]
77. Moebus M, Hilty M, Oldrati P, Barrios L, Holz C, PHRT Author Consortium. Assessing the role of the autonomic nervous system as a driver of sleep quality in patients with multiple sclerosis: observation study. JMIR Neurotech 2024;3:e48148. [doi: 10.2196/48148]
78. Kim JW, Kim S, Cho E, et al. Evaluation of sound-based sleep stage prediction in shared sleeping settings. Sleep Med 2025 Aug;132:106533. [doi: 10.1016/j.sleep.2025.106533] [Medline: 40315671]

Abbreviations

AI: artificial intelligence
ASCVD: atherosclerotic cardiovascular
CVD: cardiovascular disease
HR: hazard ratio
ICC: intraclass correlation coefficient
iSD: intraindividual standard deviation
OSA: obstructive sleep apnea

XSL•FO
RenderX

PLSC: partial least squares correlation
PPG: photoplethysmography
SFI: sleep fragmentation index
SpO_2: oxygen saturation
SRI: sleep regularity index

Edited by S Brini; submitted 12.Aug.2025; peer-reviewed by E Mahmoudi, Sunny, CL Au; revised version received 10.Dec.2025; accepted 22.Dec.2025; published 05.Mar.2026.

Please cite as:
Zawada SJ, Faust L, Fortune E
Tracking Cognitive Health With Wearables in Telerehabilitation Female Participants: Could Nighttime Sleep Measures Be Used as Sex-Specific Digital Endpoints?
JMIR Neurotech 2026;5:e81318
URL: https://neuro.jmir.org/2026/1/e81318
doi:10.2196/81318

© Stephanie J Zawada, Louis J Faust, Emma Fortune. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 5.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

A Pocket Laboratory for Functional Neuroimaging Research Using Mobile Visual Oddball, Multimodal Electroencephalography, and Functional Near-Infrared Spectroscopy Imaging: Instrument Validation Study

Peter Rokowski[1], BS, MSc; Meltem Izzetoglu[1], BS, MSc, PhD; Luis Gomero[1], BS, MSEE; Roee Holtzer[2,3], BS, BA, MA, PhD

[1]Department of Electrical and Computer Engineering, Villanova University, 800 E. Lancaster Ave, Villanova, PA, United States

[2]Ferkauf Graduate School of Psychology, Yeshiva University, New York, NY, United States

[3]Department of Neurology, Albert Einstein College of Medicine, New York, NY, United States

Corresponding Author:
Peter Rokowski, BS, MSc
Department of Electrical and Computer Engineering, Villanova University, 800 E. Lancaster Ave, Villanova, PA, United States

Abstract

Background: The need to observe brain activity in more natural environments, that is, outside of laboratory settings, is critical for understanding cognition. Wearable low-cost neuroimaging modalities (electroencephalography [EEG] and functional near-infrared spectroscopy [fNIRS]) are portable, noninvasive, and robust to motion artifacts but lack similarly portable tools for use in ecologically valid studies. Smartphones are ubiquitous, programmable, wireless, and thus strong candidates for "pocket laboratories" companion platforms that travel with study subjects. Therefore, we developed the Wearable Cognitive Assessment and Augmentation Toolkit (WearCAAT), a cross-platform neuroimaging task platform that integrates external sensors via the lab streaming layer (LSL) and supports over 100 sensor types. We validated our implementation with healthy human participants under multimodal neuroimaging conditions prior to analysis of data collection in ongoing clinical settings.

Objective: This study aimed to validate WearCAAT, as a platform for functional neuroimaging research, via analysis of human participant data collected during our ongoing National Institutes of Health–funded study.

Methods: We analyzed data from healthy college-aged (ages 18 - 30 y) adult participants, who completed a battery of shortened neurocognitive tasks (each lasting 4 min) in WearCAAT, while outfitted with research-grade multimodal EEG and fNIRS sensors. We indicated validity via the presence of task-related behavioral responses and their neuroimaging correlates. As a representative example, we analyzed the visual oddball task due to its well-documented poststimulus features for EEG and fNIRS. We extracted behavioral responses, mean response accuracies, and response times for infrequent (target) and frequent (standard) stimuli classes. We examined, poststimulus, P300, positive amplitude deflection around 300 (ms) in EEG and increased average oxygenated hemoglobin (HbO) levels in fNIRS.

Results: We enrolled a total of 57 (male individuals: n=27, 47%; female individuals: n=30, 53%; mean age 22, SD 3.4 y) participants for data collection. We excluded the first 4 (7%) participants from our analysis due to technical errors. Our analysis revealed increased mean response times for infrequent (target) stimuli (mean 718, SD 148 ms) compared to frequent (standard) stimuli (mean 542, SD 122 ms) with the Wilcoxon test ($Z=6.33$; $P<.001$; $r=0.87$); higher P300 amplitudes over midline regions (parietal and temporal) for EEG; and increased oxygenated hemoglobin over the prefrontal cortex for fNIRS. All participants completed the full battery and reported no usability concerns or app crashes. Similarly, we observed no data loss or corruption that would negatively impact analyses.

Conclusions: WearCAAT-provided outcomes from our study, which analyzed multimodal neuroimaging data collected during a mobile app–based visual oddball task, matched expectations from the literature. While full validation is ongoing for other tasks, we demonstrated initial validity of our app for neurocognitive imaging use. Our app and approach represent the first attempt at dedicated neuroimaging mobile-pocket laboratory and contribute to greater studies in ecological validity.

(JMIR Neurotech 2026;5:e78217) doi:10.2196/78217

KEYWORDS

pocket laboratory; neuroimaging; functional near-infrared spectroscopy; fNIRS; smartphone data collection; mobile brain and body imaging

XSL•FO
RenderX

Introduction

Functional neuroimaging entails the use of noninvasive brain monitoring sensors during tailored neurocognitive tasks to measure cortical activity in different brain regions corresponding to targeted cognitive domains [1]. Neuroimaging research offers views into the brain and is critical to the understanding of brain development, injury, and disease or impairment. High-precision techniques such as functional magnetic resonance imaging have limitations that include participant refusal, high cost, and requirements on staying still in the supine position [2]. This presents an interesting challenge that cognitive neuroscientists have grappled with since the early days of cognitive studies on "ecological validity" [3], wherein experimental conditions, particularly those imposed by the nature of laboratory settings and high-precision imaging modalities such as functional magnetic resonance imaging, impact the observed phenomena [4]. A complementary solution to the challenge of ecological validity is to use wearable, low-power, wireless technologies to combat such challenges facing traditional approaches [5]. Noninvasive wearable neuroimaging modalities include functional near-infrared spectroscopy (fNIRS) to monitor changes in cerebral hemodynamics related to cognitive activity and electroencephalography [EEG] to monitor neural activations via changes in electrophysiology. Both methods enable observations of human participants in more "natural" environments, and their complementary nature allows overcoming of their individual limitations on temporal and spatial resolution when used together [6].

Development and advances in commercially available mobile devices, such as tablets and smartphones, enable portable platforms for conducting human experiments that can further mobilize wearable sensing modalities. The resulting concept "pocket laboratories" describes this paradigm well and poses a unique solution to ecological validity by enabling human participant research in nonstatic environments that could potentially travel with a participant. Researchers in human behavior and medicine have leveraged the widespread adoption of mobile devices as pocket laboratories for more "natural" environments, which are classified under mobile health apps. For example, smartphones and tablets were used in conjunction with neuroimaging devices to measure interactions with websites, apps, and each other, to glean insight into health behaviors such as alcohol consumption [7] and Alzheimer disease detection [8,9]. In addition to the conveniences and benefits that mobile devices offer, they also provide useful hardware for human participant research such as internet connectivity and integrated sensors (ie, gyroscopes, accelerometers, and GPS) [10,11] and are extensible through Bluetooth connections with external wearable health sensors such as the ones embedded in smartwatches. With an estimated 4.7 billion smartphone users by 2024 [12], the potential for participant recruitment is vast, which can further improve not only our understanding of cognition but also allow for the early detection of different conditions, that is, cognitive impairment and monitoring of treatment outcomes through larger studies and populations otherwise unattainable.

Despite the practical uses of mobile devices in clinical research, there's a notable gap: a platform for conducting general functional neuroimaging research using mobile devices. While accepted tools such as the NIH Toolbox [13] developed by the National Institutes of Health provide a platform for gathering psychometrics on human participants and are widely used in clinical settings (225 journals and conferences as of 2022 [14]), they lack compatibility with functional neuroimaging sensing modalities. Other apps are too limited in scope and lack integration beyond basic proof of concepts and require significant effort to extend to new scenarios. We believe this is because app development is difficult and requires deep technical knowledge and funding that is outside the scope of normal external funding vehicles [15]. In response to this, we developed a framework for a functional neuroimaging pocket laboratory and provided implementation in the Wearable Cognitive Assessment and Augmentation Toolkit (WearCAAT).

WearCAAT is a cross-platform mobile app, used on both iOS and Android, in conjunction with external single or multimodal sensors, integrated via the lab streaming layer (LSL) [16]. LSL adds signal synchronization capabilities, equivalent on mobile devices to desktop systems [17]. However, validation of our framework and implementation is still outstanding. There are numerous challenges in translating a desktop software capability to mobile devices, especially in the domain of functional neuroimaging. Touchscreens are dual-purpose tools that share the responsibility of presenting stimuli and capturing responses via "soft" buttons. Mobile operating systems are sandboxed in nature and typically prevent access to high-precision time-aware clocks, as well as limit multithreading capabilities. A full end-to-end test for our paradigm is necessary to understand the limits and abilities of pocket laboratories in functional neuroimaging.

Methods

Ethical Considerations

Participants signed informed consent before completing cognitive tasks using WearCAAT on an iPad. The consent and collection protocol were reviewed and approved by the Biomedical Research Alliance of New York, LLC (BRANY) [18], external institutional review board (1R01AG077018-01), on March 24, 20. Participant privacy was covered under the certificate of confidentiality by the National Institutes of Health that states that researchers will not disclose or use information that may identify participants in any federal, state, or local civil, criminal, administrative, legislative, or other action, suit, or proceeding, even if there is a court subpoena (with exceptions being federal, state, or local law that requires disclosure, or the explicit approval of individual participants to release their name and/or personally identifiable information). Participants were compensated US $20.

Procedure

We examined whether neurocognitive tasks provided by WearCAAT, on commercial mobile devices, reliably elicited cognitive engagement and whether the corresponding biological markers were detectable and identifiable in neuroimaging data. This required that (1) behavioral responses aligned with

XSL•FO
RenderX

established task-specific patterns in the literature, and (2) these responses enabled the extraction of physiologically meaningful signals from neuroimaging modalities. Failure to meet both criteria across tasks constituted a negative inconclusive finding, whereas partial success supported the technical validity of our implementation and integration. Consequently, we scoped our analysis on the visual oddball paradigm [19], also built in WearCAAT, which is well studied in both EEG and fNIRS modalities for attention monitoring with established expected behavioral, neural, and hemodynamic outcomes [20-23].

First, we hypothesized that participants' behavioral data during the visual oddball task, as built in WearCAAT, would exhibit a longer response time (RT) to infrequent (target) stimuli than to frequent (standard) stimuli. Second, we hypothesized that neural correlates for cognition, as measured by EEG and fNIRS during the performance of the visual oddball task, would be detectable in their respective sensing modality when examined using time stamps obtained from the behavioral data. For EEG, the WearCAAT-implemented visual oddball task would evoke a higher P300 subcomponent (positive deflection in amplitude around 300 ms after stimulus) in event-related potentials (ERPs; stimulus-locked activations in EEG) to the infrequent (target) stimuli as compared to the frequent (standard) ones in the midline region. For fNIRS, the average oxygenated hemoglobin (HbO) would positively increase in response to infrequent (target) stimuli as opposed to the frequent (standard) stimuli, in the right prefrontal cortex (PFC).

The remainder of this section describes WearCAAT and the relevant features of this study, followed by the participant information, data collection protocol, the visual oddball task as presented in WearCAAT, and the signal processing pipeline to extract the EEG- and fNIRS-specific components that support or reject our hypotheses.

WearCAAT: An Overview

WearCAAT implements task-based neurocognitive monitoring, wherein participants perform 1 of 11 built-in tasks to elicit known cognitive effects in different domains, including attention, vigilance, working and episodic memory, response inhibition, set shifting, and conflict resolution. Currently implemented tasks in WearCAAT and their cognitive effects are presented in Table 1. We built WearCAAT using C# and Extensible Application Markup Language (XAML) from the Multi-App User-Interface (MAUI) framework [24] with .NET 8. MAUI provides cross-platform (supporting Android and iOS phones and tablets) app design in a unified project. LSL is integrated using the "slim bindings" approach for high-performance C and C++ libraries to be leveraged through different programming languages, giving direct access to the necessary libraries.

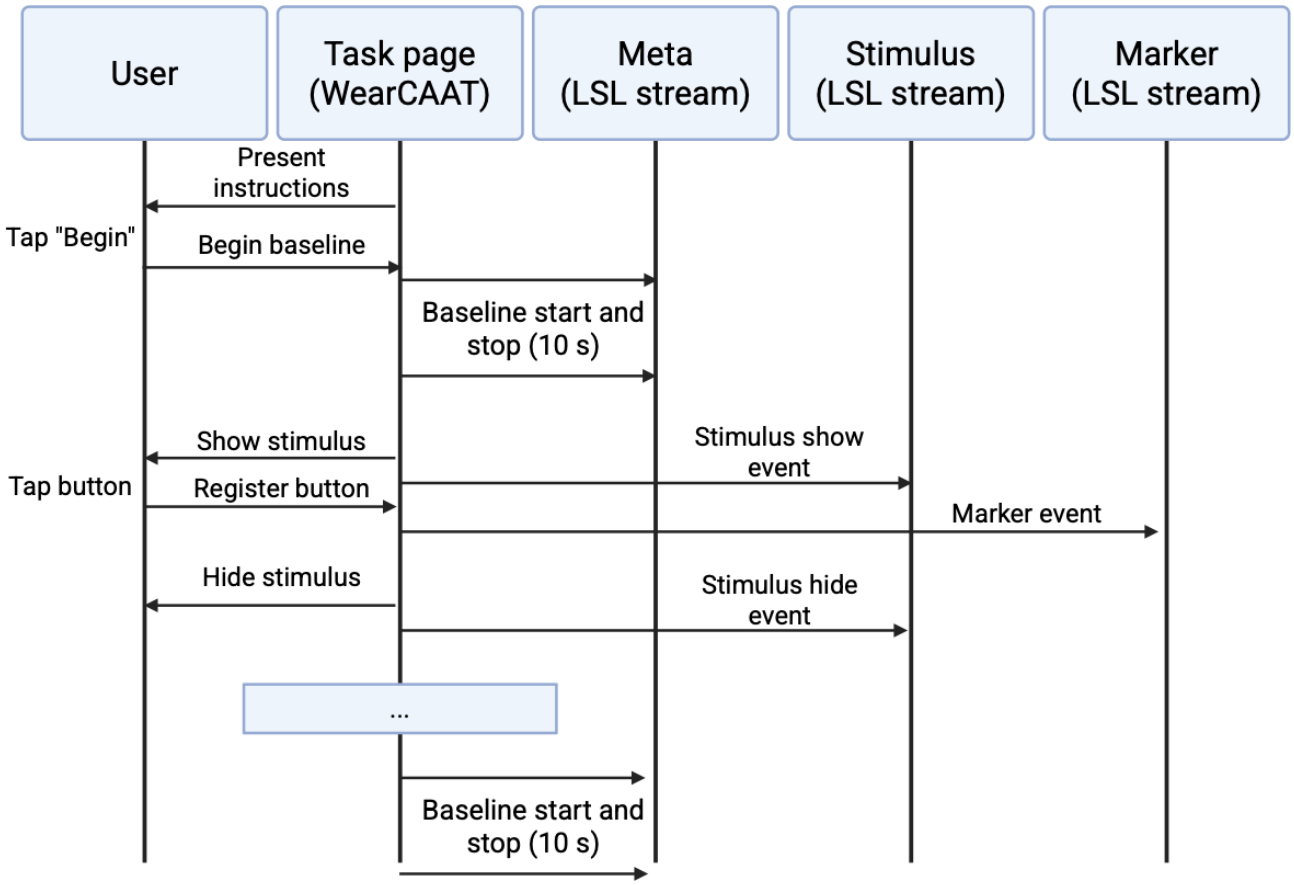
Table . Currently existing cognitive task battery implemented in the Wearable Cognitive Assessment and Augmentation Toolkit. Cognitive battery has a total runtime of approximately 1.5 hours, and each task had a runtime of 4 minutes with a 30-second rest period in between, which can be implemented in a randomized order.

Task name	Cognitive effect
Psychomotor vigilance task [25,26]	Attention or vigilance
Visual oddball paradigm [19]	Attention or working memory
Go/no-go [27]	Response inhibition
N-back (n = [0, 1, 2]) [28,29]	Working memory
Stroop [30]	Selective attention
Flanker [31,32]	Conflict resolution
Wisconsin card sorting [33]	Set shifting
Verbal memory recognition [34]	Episodic memory
Bluegrass [35]	Episodic memory
Resting (eyes = [open, closed]) [36]	Default mode net

Neurocognitive tasks followed the sequence described in Figure 1. The participant first read the instructions, then started the assessment via tapping the "begin" button, upon which the screen was blank for a configurable resting baseline time window before the task's logic loop began. The task expired after the set amount of time and was followed by a second baseline period. Timing information was determined using a stopwatch object, which counted monotonically from the start, as is common in psychometrics platforms [16]. Tasks were configurable in the app to allow further flexibility and experimentation, specifically regarding stimulus type, timing, interstimulus interval information, stimulus presentation ratios, and more, depending on the task.

XSL•FO
RenderX

Figure 1. Task sequence diagram. Baseline periods begin immediately, with all events related to the task going through to LSL streams. “Task page” represents the logic controller behind the task in WearCAAT (created in BioRender [37]). LSL: lab streaming layer; WearCAAT: Wearable Cognitive Assessment and Augmentation Toolkit.



Task events were broken down into three categories as follows: (1) metadata pertaining to task information and configuration for auditing purposes; (2) stimulus event appearances, types, etc; and (3) user markers, button presses, or other responses. Each category streamed data to a corresponding LSL stream outlet, which sent the data wirelessly to the recording platform. Figure 1 depicts the sequence of events and the respective streams for each task.

Participants

We recruited 57 (male individuals: n=27, 47%; female individuals: n=30, 53%) participants, aged 18 to 30 (mean age 22, SD 3.4 y) years, from the undergraduate and graduate student bodies at Villanova University via flyers posted in common university spaces. We detail participant demographic data in Table 2. Exclusion criteria included current or past severe neurological or psychiatric disorders and significant vision or hearing impairments. Participants first attended an initial screening, where we collected demographic data and relevant medical histories via a written survey, measured the participant’s head circumference to determine neuroimaging device cap size, and scheduled a follow-up data collection session. After participants signed the informed consent form, we collected behavioral and neuroimaging data from them while they used WearCAAT in a session that took approximately 1.5 hours.

Table . Demographic data of recruited college-aged participants.

Demographics	Male (n=27), n (%)	Female (n=30), n (%)	Total (N=57), n (%)
Asian	8 (30)	7 (23)	15 (26)
Black or African American	1 (4)	0 (0)	1 (2)
Hispanic	2 (7)	5 (17)	7 (12)
White	13 (48)	13 (43)	26 (46)
White and Black or African American	2 (7)	0 (0)	2 (4)
White and Hispanic	1 (4)	5 (17)	6 (11)

XSL•FO
RenderX

Data Collection

We administered all 11 abbreviated tasks to participants via WearCAAT in sequence, where the first 2 were the resting tasks (eyes opened and eyes closed) and the remaining 9 tasks were presented in randomized order (Table 3). All participants completed the entire task battery built in WearCAAT in 1 sitting while wearing a full head cap housing the multiple optodes. The electrodes formed a combined wireless fNIRS-EEG system (NIRSport2, NIRx Medizintechnik GmbH and Smarting-mbt wireless EEG, mBrainTrain, respectively) [38,39]. Using our hybrid neuroimaging system, we collected 51-channel fNIRS (43 long and 8 short distances) and 32-channel EEG data from the frontal, temporal, and parietal regions of the brain simultaneously. We used 3 distinct flexible cap sizes as provided by NIRx company [40]—small (56 cm), medium (58 cm), and large (60 cm)—all with 128 slits for probe locations identified according to the 10-20 international system to accommodate for different head sizes, improve comfort, and ensure fNIRS and EEG measurements with good coupling from similar head locations. The complete layout of our protocol is detailed in Figure 2; the LSL data streams and their respective interfaces are all coordinated through a Wireless Area Network (WAN) hosted on a private router with no external internet connection nor devices.

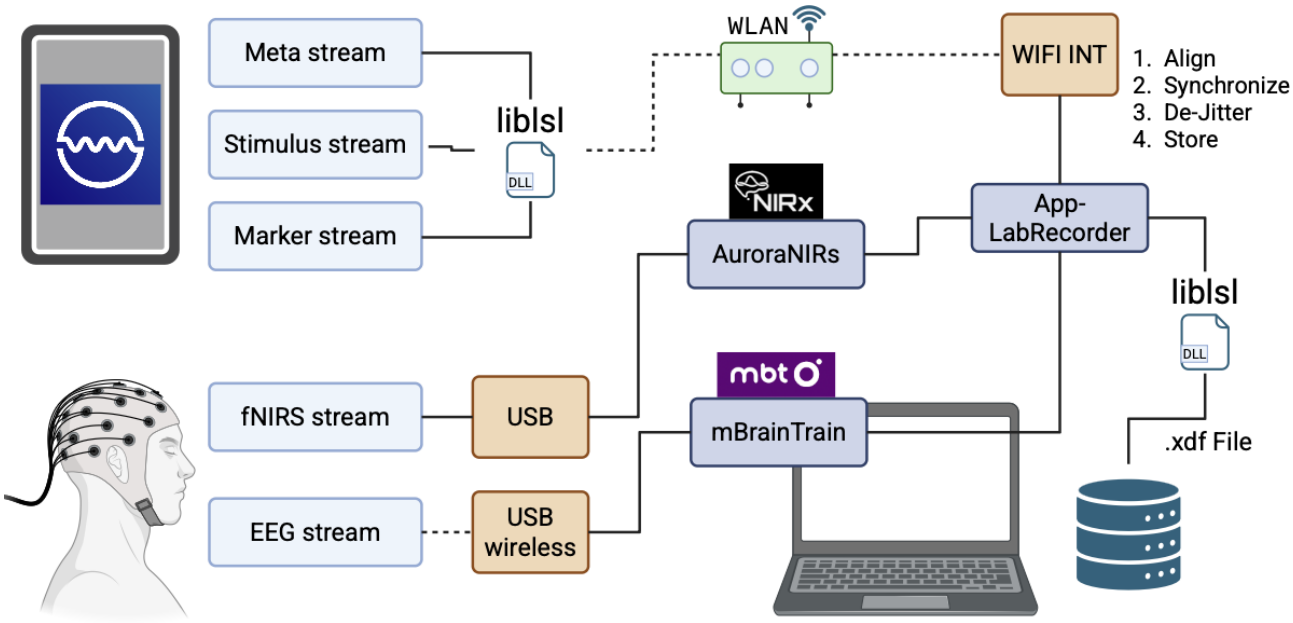
Table . The full mapping of electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) sensor locations to functional regions of interest (ROI). For our analysis, we considered regions over the prefrontal cortex, specifically the "frontal right" ROI for fNIRS. The "frontal" (Fz), "parietal" (Pz), and "temporal" (Cz) midlines are areas we focus on to observe the P300 for EEG.

ROI	EEG channels	fNIRS channels
Frontal left	• AFP1 • AFF5 • F3 • F1	• FPZ-FP1, FPZ-AF3 • FF7-AF3 • F5-AF3, F5-F7 • AF3-AFz
Frontal right	• AFP2 • AFF6h • F4 • F2	• FPZ-FP2, FPZ-AF4 • AF8-FP2, AF8-AF4 • F6-AF4, F6-F8 • AF4-AFz
Temporal left	• FTT7h • TTP7h	• FT8-T8 • TP8-T8 • C6-T8
Temporal right	• FTT8h • TTP8h	• FT7-T7 • TP7-T7 • C5-T7
Parietal left	• P1, P7 • CPP5h • TPP8h	• P5-P3, P5-C5 • P3 • CP3
Parietal right	• P2, P8 • CPP6h • TPP8h	• P6-P4, P6-CP6 • P4 • CP4
Frontal midline	• Fz	—[a]
Parietal midline	• Pz	—
Temporal midline	• Cz	—

[a]Not available.

XSL•FO
RenderX

Figure 2. Lab streaming layer event pipelines in the Wearable Cognitive Assessment and Augmentation Toolkit; buttons, stimuli, and metadata streamed wirelessly using the embedded liblsl library to the laptop via an onboard wireless interface; fNIRS connected via wired USB-cable collected using AuroraNIRS; EEG streamed via dedicated wireless dongle; data are saved to XDF files. DLL: Dynamic-Link Library; EEG: electroencephalogram; fNIRS: functional near-infrared spectroscopy; WLAN: wireless-local area network; XDF: extensible data format.



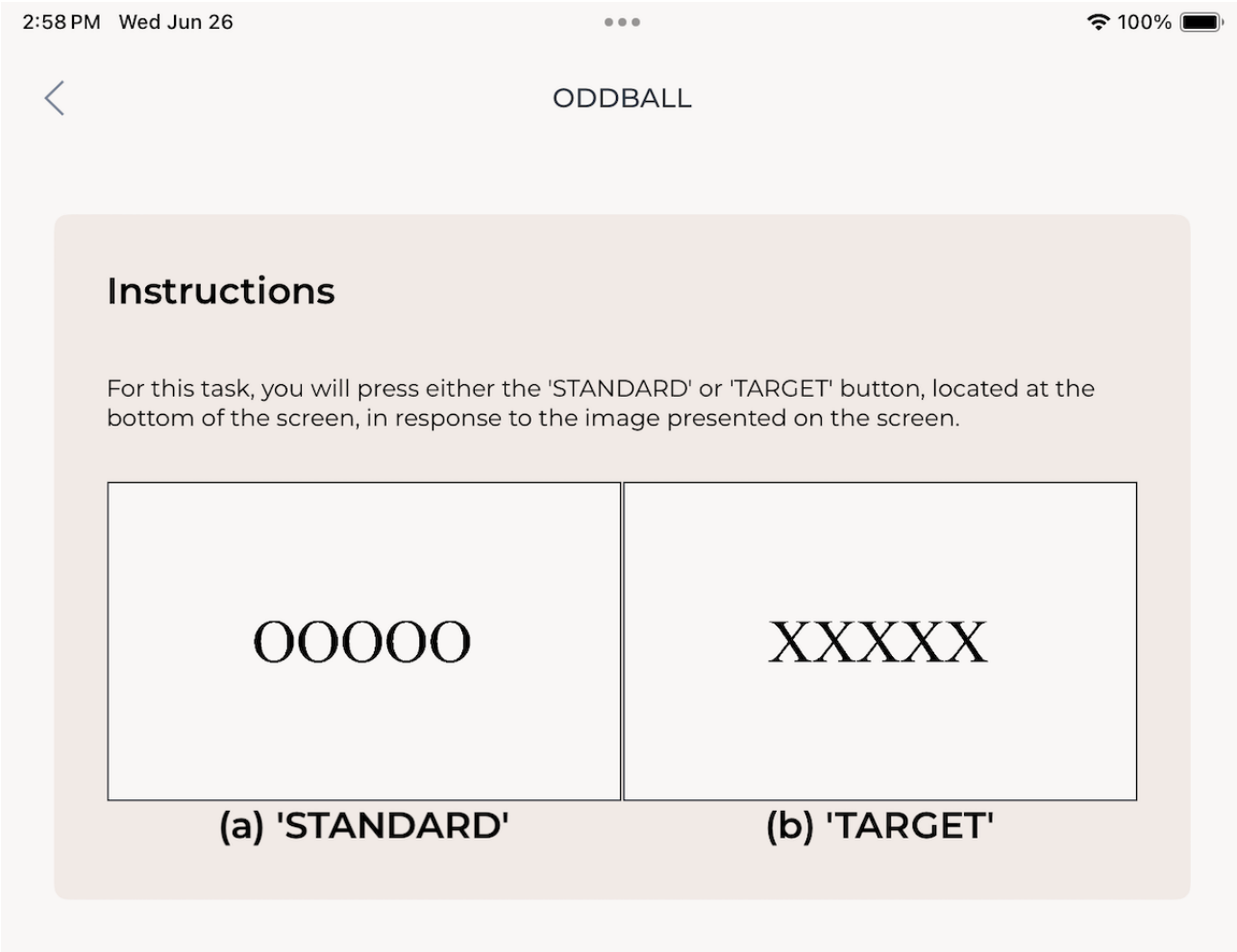
Behavioral responses were recorded in WearCAAT where task events (eg, stimulus presentation times, user responses via button presses, etc) were time stamped in WearCAAT and wirelessly streamed through LSL to a laptop (Windows 10) running App-LabRecorder [41]. Concurrent EEG and fNIRS data were also streamed wirelessly to the same laptop, which synchronized all clock times and removed jitter automatically. While WearCAAT supports Android, we only used an Apple iPad Pro (sixth generation) as our mobile platform due to the consistency of iOS devices. Using different operating systems and devices from different manufacturers might introduce errors that are harder to quantify [42], and that is out of scope for this body of work.

Evaluation Protocol: Visual Oddball Paradigm

To provide empirical support and initial validation, we reported our neuroimaging and behavioral outcomes obtained from the abbreviated visual oddball task implemented via WearCAAT. In this task, participants were presented with either 1 of 2 types of visual stimuli where each consisted of 5 repeated letters and asked to respond by tapping 1 of the 2, parallel and equally sized, buttons at the bottom of the screen with the index finger on their dominant hand. The target stimulus ("XXXXX") matched with the left-most button labeled "TARGET," and the standard stimulus ("OOOOO") matched to the right-most button. The interstimulus interval was 2 seconds, with stimuli presentations lasting 0.5 seconds and the screen remaining blank for 1.5 seconds. Target stimuli appeared infrequently relative to the standard, with at least 7 to 21 standard stimuli appearing between each target presentation to reduce participant expectation. On average, each participant witnessed 11.87 (SD 1.15) standard stimuli between each successive target stimulus. The total duration of the task was 4 minutes and occurred between two 10-second baseline periods. Participants received instructions verbally from the experimenters and in the app before beginning each task, as depicted in Figure 3. Participants began the task by tapping the "begin" button, which started the baseline period followed by the oddball sequence.

XSL•FO
RenderX

Figure 3. In-app instructions for the visual oddball task and examples of stimuli. (**A**) STANDARD stimulus on the left (OOOOO) and (**B**)TARGET stimulus on the right (XXXXX). Stimuli are presented as rectangles with black font on a white background.



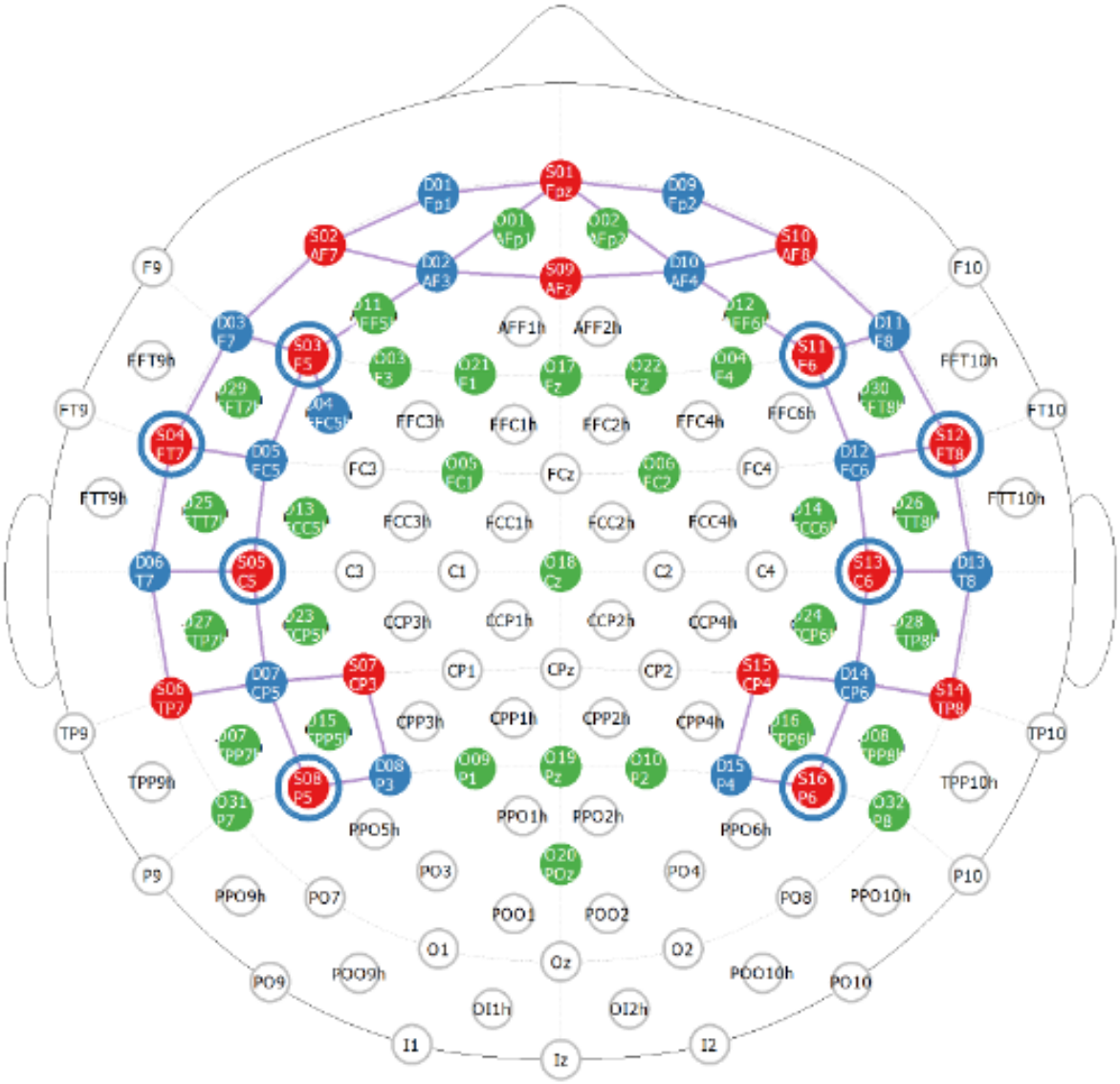
Signal Processing

Overview

The simultaneous 51-channel fNIRS and 32-channel EEG data were collected from the full head in frontal, temporal, and parietal locations as shown in Figure 4, with regions of interest (ROI) detailed in Table 3. Our fNIRS and EEG data processing pipeline for artifact removal (motion, physiological, environmental, or equipment noise) and data conversion (hemodynamic response extraction) was performed offline using custom-built MATLAB (version R2024; MathWorks, Inc) codes [43] and in accordance with published best practices [44-46]. We used the 8 fNIRS short channels to remove skin artifacts. EEG data were additionally processed for the removal of eye blinks, eye movement, muscle artifacts, power line noise, and limiting the data within the range of 0 Hz to 45 Hz.

XSL•FO
RenderX

Figure 4. Our hybrid cap design based on the modified 10-20 layout provided by NIRx. Electroencephalogram sensors are depicted in green, functional near-infrared spectroscopy (fNIRS) source in blue, and fNIRS detectors in red. fNIRS short channel locations are shown with blue circles around the red sources.



Given the large number of channels (32 for EEG and 43 long separations for fNIRS), we organized the data into ROIs based on both anatomical and functional considerations (detailed in Table 3). To manage the data efficiently, we averaged the responses for each stimulus (target and standard) separately across all 57 participants, focusing on "valid" data from each participant, channel, and stimulus. We determined a channel to be invalid based on impedance for EEG (>30 kΩ) [47], and Scalp Coupling Index (<0.4) for fNIRS [48]. Finally, we averaged the score per channel across all participants and grouped the channels based on their respective ROIs, reducing the dimensionality of the dataset.

WearCAAT Markers: User Behavioral Responses

We gathered the response accuracy by evaluating the correct and incorrect user responses and extracting the reaction times to target and standard stimuli. We determined significance between target and standard responses using the Wilcoxon test, a nonparametric alternative to the paired 2-tailed *t* test when the data do not follow a normal distribution.

EEG: Event-Related Potential

We extracted stimulus-locked data epochs (ERPs) from EEG recordings using 0.2-second prestimulus and 1-second poststimulus interval and performed a baseline correction by subtracting the mean of the prestimulus data from the whole data epoch. We then removed intrinsic response variability by averaging multiple trials of epochs within the task to obtain an averaged ERP for each stimulus (target and standard), separately. Finally, we gathered the ERP features by extracting the positive and negative peak amplitudes with their respective timings, specifically focusing on the P300 component (positive

XSL•FO
RenderX

deflection around 300 ms after the stimulus), which was shown to be higher in target response as compared to the standard one in the visual oddball task in healthy young adults [49,50].

fNIRS: Oxygenated Hemoglobin

We processed our fNIRS data by removing motion artifacts and baseline shifts through wavelet and spline filters [44]. We further removed physiological signals, cardiac, respiratory, and Mayer waves, with a finite impulse response low-pass filter using the cutoff frequency 0.08 Hz [51]. Finally, we converted the light intensity measurements into changes in HbO and deoxygenated hemoglobin using the modified Beer-Lambert Law with published coefficients from the literature [52]. After using short channel recordings to remove potential skin blood flow artifacts from long channel recordings using a general linear model [53], we extracted 20-second poststimulus data epochs and applied baseline correction using the 1-second prestimuli onset. Notably, since HbO was the most used fNIRS measure in studies implementing the visual oddball task that was indicative of cognitive activity–related changes in attention domain–specific apps [23], we focused our results and comparisons to only HbO outcomes in this study.

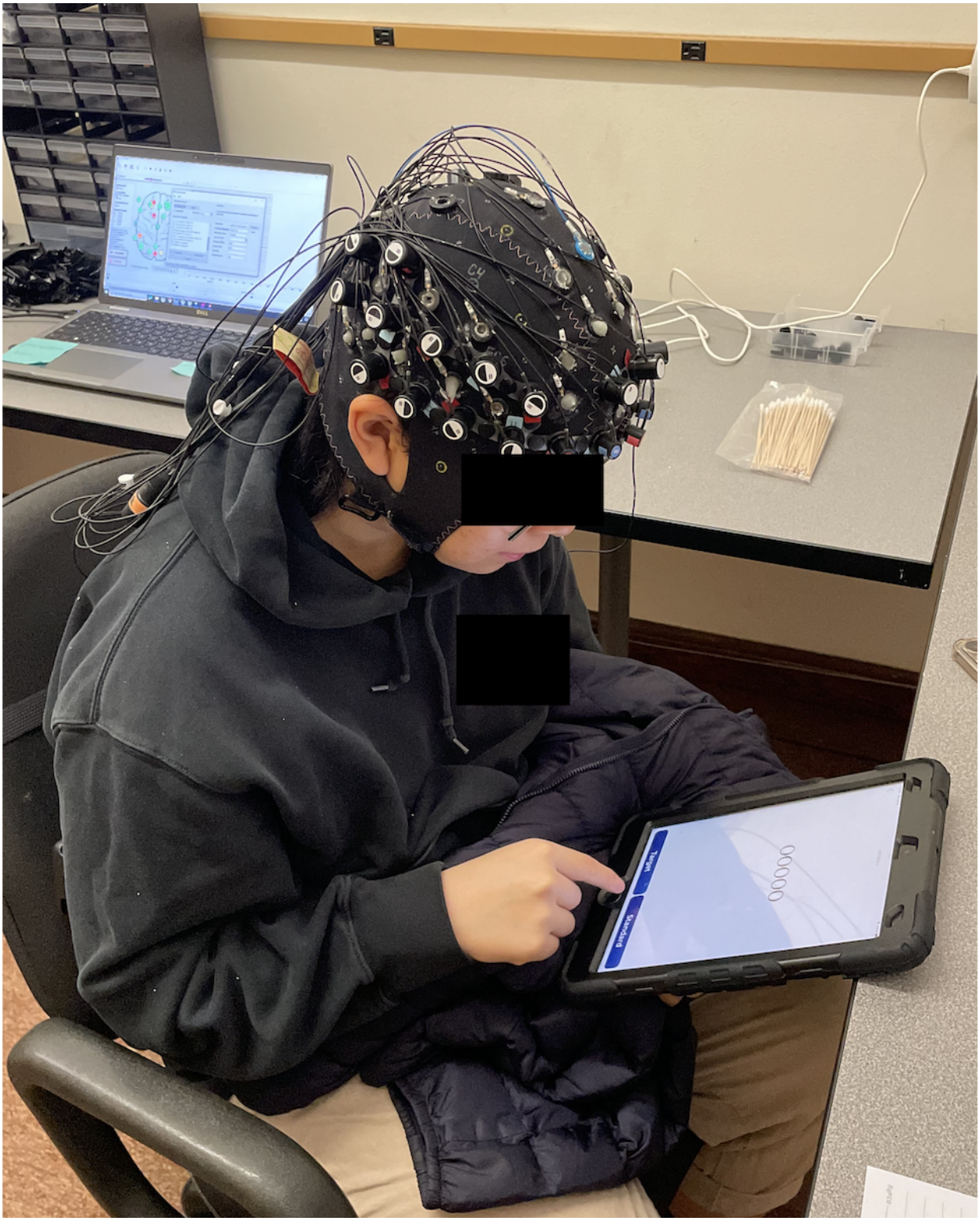
Results

Overview

We enrolled 57 participants in data collection, and all of them performed all 11 tasks in 1 sitting. We observed zero participant dropout with no app crashes or corrupted data. We excluded data from our first 4 (7%) participants; 2 due to poor impedance from improper cap setup, and 2 after a patch to WearCAAT that altered the timing logic to improve the responsiveness of the touch screen during timed loops. WearCAAT collected and synchronized multiple concurrent streams of task-related data (1 stream for stimulus; 1 for each button press; and 1 for metadata and task events, such as task start and stop and baseline start and stop) with no data loss in participant responses or disconnects from the recording server. We also observed no additional loss of information or signal content from the combined fNIRS-EEG sensors as well. Participants reported no complaints or concerns with the WearCAAT app, the instructions provided either in app or verbally, or the overall data collection procedure, indicating a low participant burden. An example participant can be seen sitting comfortably during collection in Figure 5.

XSL•FO
RenderX

Figure 5. Participant during data collection, responding to a standard (infrequent) stimulus presented via the Wearable Cognitive Assessment and Augmentation Toolkit.



User Response Times

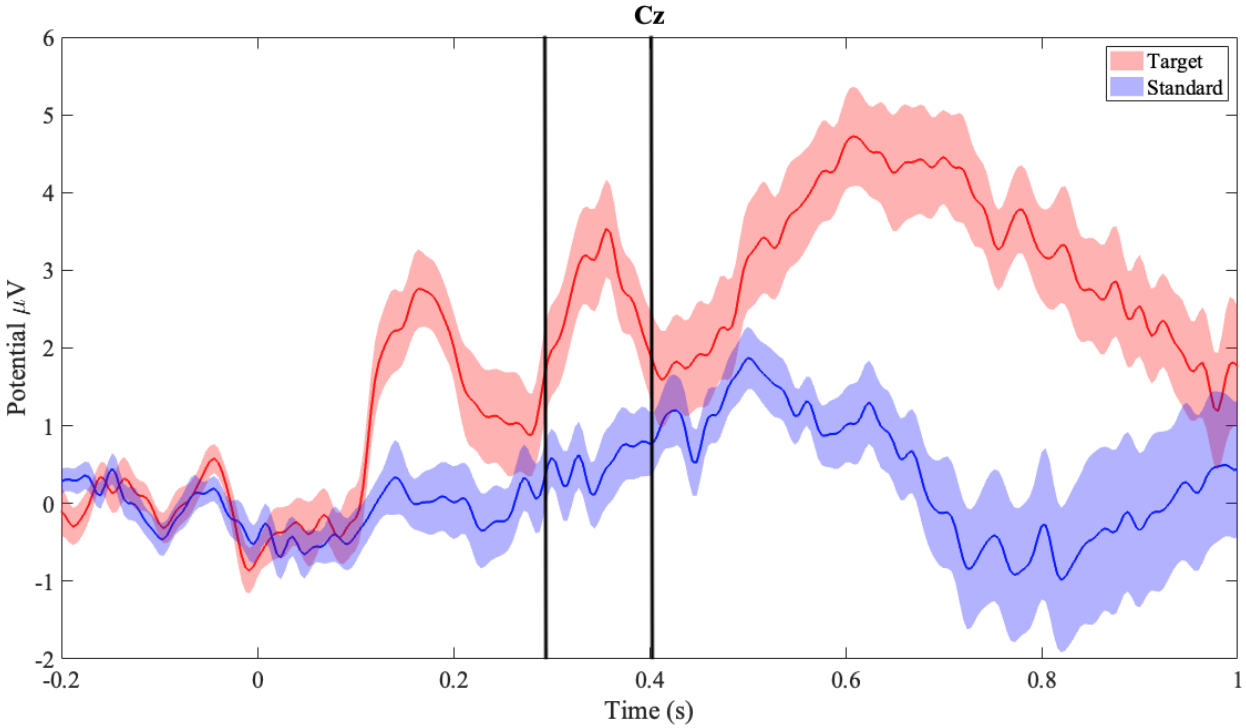
We found that the mean RT for target stimuli was 718 (SD 148) milliseconds and, for standard stimuli, it was 542 (SD 122) milliseconds. The Wilcoxon test revealed significant differences in RT to target (infrequent) stimuli as compared to the standard (frequent) ones (Z=6.33; *P*<.001; *r*=0.87). These outcomes indicated that the participants took longer to identify the target stimulus than the standard stimulus. Participants' mean percent accuracy for identifying stimuli was 94.58 (SD 7.412) for the target stimulus and 99.12 (SD 1.447) for the standard stimulus, suggesting that participants identified and responded to the frequent stimulus more correctly as compared to the infrequent ones overall.

XSL•FO
RenderX

EEG: Extracted P300

All participant-averaged (SEM) ERP waveforms obtained using WearCAAT in an iPad for the abbreviated visual oddball task for target (red) and standard (blue) stimulus in the parietal midline (Pz) and central midline (Cz) regions are presented in Figure 6, separately. Our results showed higher P300 amplitude in response to target stimuli as compared to the standard one, especially in the midline regions on the Cz and Pz locations, in line with the published literature on computerized presentation of a regular length (approximately 20 min) visual oddball task. The Cz, presented in Figure 6, recorded a positive peak within the 250-millisecond to 400-millisecond intervals at 356 milliseconds, having an amplitude of 3.5397 (SD 0.6258) µV for the target stimulus and at 372 milliseconds with an amplitude of 0.74667 (SD 0.4929) µV for the standard stimulus.

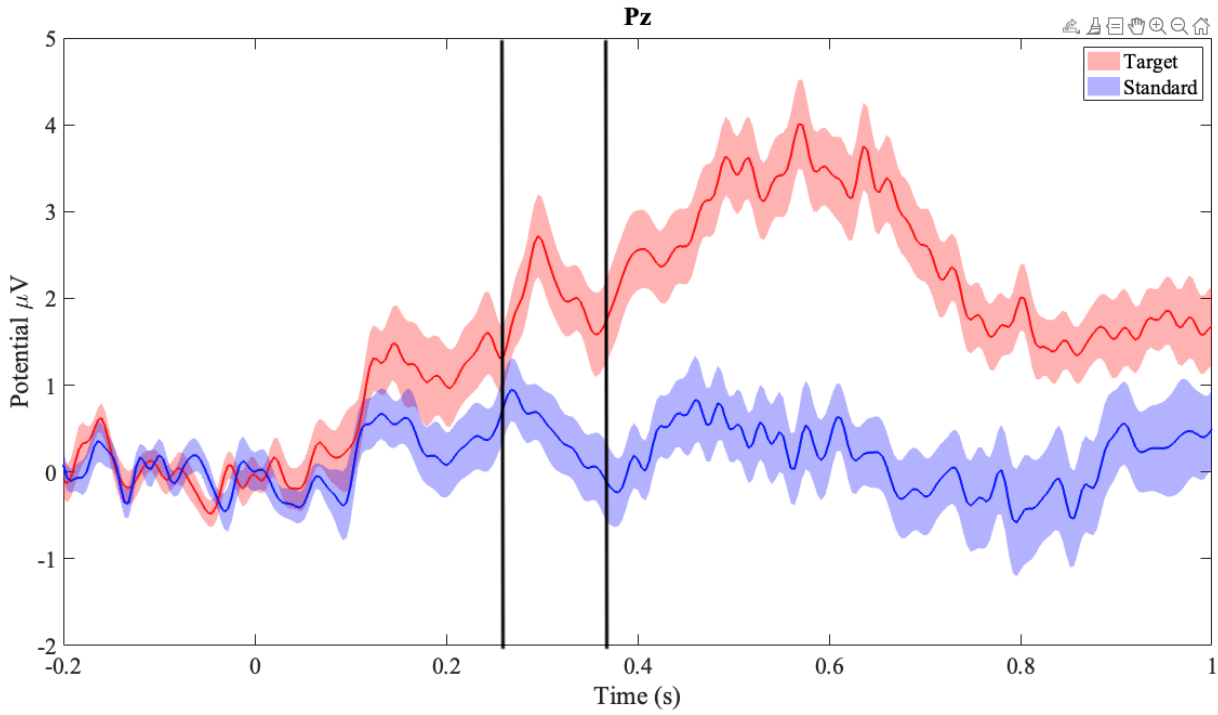
Figure 6. Mean amplitudes across participants in the central midline (Cz). Clouded regions represent the standard error of the mean. Red represents target (infrequent) stimulus and blue represents (frequent) stimulus responses.



Similarly, the Pz, presented in Figure 7, recorded a peak at 296 ms with an amplitude of 2.7219 (SD 0.4756) µV in the target stimulus and a peak at 268 ms with an amplitude of 0.949 (SD 0.3637) µV in the standard stimulus.

XSL•FO
RenderX

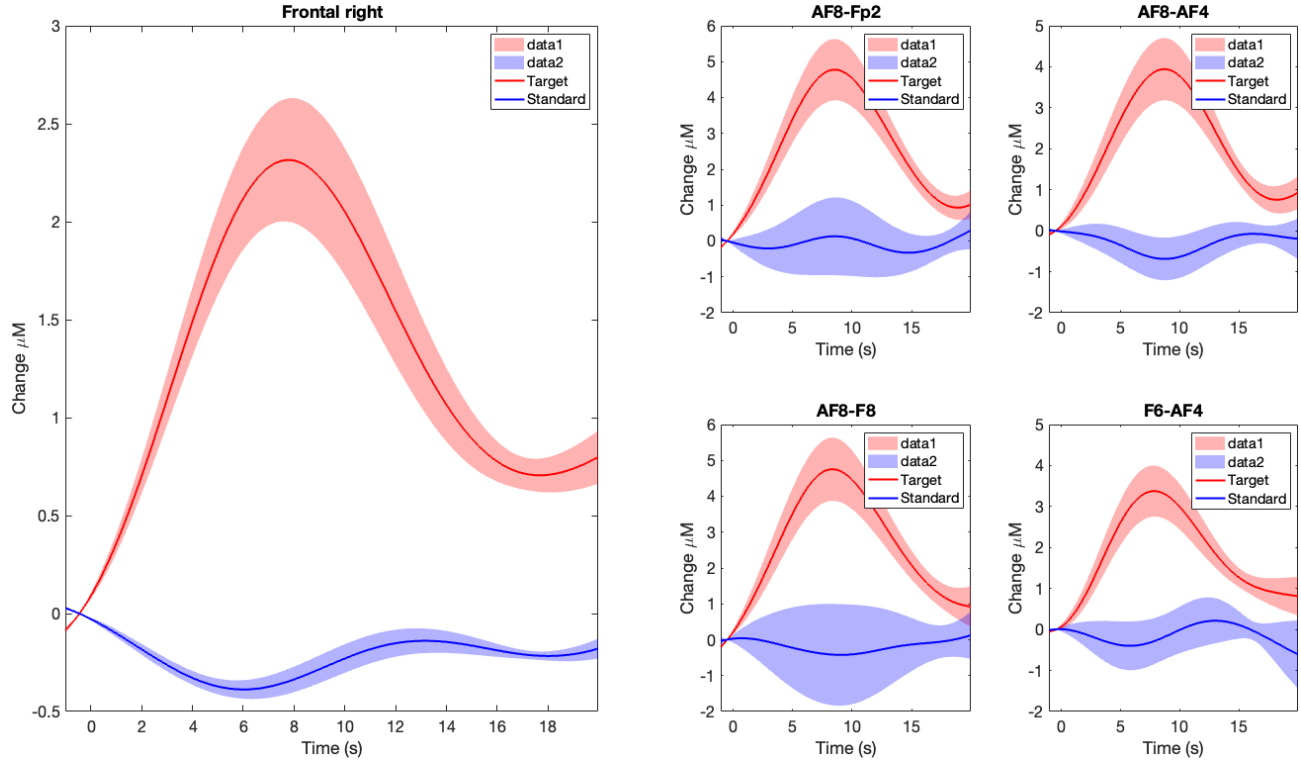
Figure 7. Mean amplitudes across participants in the parietal midline (Pz). Clouded regions represent the standard error of the mean. Red represents target (infrequent) stimulus and blue represents (frequent) stimulus responses.



fNIRS: Trends in Oxygenated Hemoglobin

As presented in Figure 8, the overall participant-averaged (SEM) HbO activations in the right PFC demonstrated a clear, positive increase in response to the target (red) stimulus peaking at approximately 9 seconds and then returning to lower values following the shape of a typical hemodynamic response function. The standard stimulus (blue) did not generate an increased HbO pattern.

Figure 8. Mean oxygenated hemoglobin activations in the frontal right region of interest across all participants. Clouded regions represent the standard error of the mean. Red represents the target (infrequent) stimulus and blue represents the (frequent) stimulus responses. The aggregate frontal right forehead region is displayed on the left, and the individual sensors of interest are displayed in the quadrant on the right.



XSL•FO
RenderX

Same trends on HbO in response to target and standard stimuli were observed in all right frontal fNIRS channels constituting the right frontal ROI, as also presented separately in Figure 8.

Discussion

Principal Findings

We presented our analysis for the abbreviated visual oddball task as presented and collected using WearCAAT and our procedures. We found indications that electrophysiological and hemodynamic activation patterns for the brain observed with simultaneously collected fNIRS and EEG data follow expected trends, despite the shorter runtime (4 min as opposed to 20 min in commonly implemented versions of the task) and the mobile app platform (as opposed to a computer screen).

Our initial hypothesis for the behavioral responses was supported. We observed greater RTs to target (infrequent) versus standard (frequent) stimuli. The Wilcoxon signed-rank test demonstrated the significance for RT between infrequent and frequent stimuli. These findings are consistent with previously reported values in the literature for the visual oddball task [54,55].

Interestingly, we also observed that periods of responses for target stimuli were greater than some reported values, whereas the standard stimuli RTs were much closer. This may be attributed, in part, to the physical differences in iPad "soft" buttons and the typical hardware switches commonly used with the visual oddball task [20]. Traditional desktop setups report participants using 2 distinct controllers (1 per hand), which dedicate a controller response per stimulus type. In contrast with our study, participants used a singular index finger to switch between button presses. The typical delay reported between stimulus types could be exacerbated by physical delays introduced by a participant needing to move his finger from hovering over 1 button to another one on the opposite side of the iPad screen. As future work, we will ensure that participants are instructed on how to hold the iPad, with relevant findings from the literature.

Our second hypothesis regarding EEG was supported by the successful extraction of P300 subcomponents from the ERP waveform, described as a positive deflection in amplitude in response to the target stimulus, appearing around 300 milliseconds after the stimulus from our collected data. We obtained higher P300 amplitude in the midline ROIs (Pz and Cz) for the infrequent (target) stimuli as compared to the frequent (standard) stimuli, following the expected outcome as reported in the literature [20,21,54,56]. Our average latencies for P300 peaks in Pz and Cz were also within previously reported bounds [57].

We noted the visual jaggedness of the P300 signals, seen in Figures 6 and 7, which we expected to be smoother, as reported in the literature. This could be caused by a combination of the shortened task times and the grand averaging technique used for analysis. Typical studies report task lengths of 20 minutes or greater for the visual oddball task, whereas this study's task length was 4 minutes. Longer task times would produce 5 times more trials for both target and standard stimuli per participant, the averages of which would smooth out irregularities and potential physiological artifacts in the time series. Further study using WearCAAT, EEG, and the visual oddball task with longer task times would provide more clarity on the matter.

Our second hypothesis regarding fNIRS was also supported by the positive average HbO increases measured in the right PFC in response to infrequent (target) stimuli as compared to the frequent (standard) stimuli. Specifically, we observed the increases in HbO in the frontal right ROI, which was widely reported in the fNIRS literature where computerized and traditional length visual oddball task was used [20,23]. In fact, such findings were prominent in all right frontal channels when considered separately as well as demonstrating an attention domain–specific global activation in right PFC as measured by fNIRS.

On usability, we point to the smoothness of data collection. Specifically, the lack of complaints from the participants and experimenters, combined with the zero-dropout rate and app crashes, is to be noted. Given that users' major concern with mobile health apps is the perceived bugginess and clunkiness of apps [15], we incorporated haptics and button color changes as feedback to users' actions. We assume the responsiveness and perceived functionality of our WearCAAT implementation is tolerable to young adults who are most fluent and comfortable in app use. However, because we did not perform a formal qualitative post–data collection survey, our interpretation is limited to "no complaints were reported." This limitation ought to be accounted for in future studies with formal participant surveys after participation to add qualitative metrics for perceived clunkiness and usability.

Conclusions

In this study, we provided evidence for the technical validation of mobile devices in task-based functional neuroimaging research via the analysis of multimodal EEG-fNIRS and behavioral data collected during an abbreviated mobile visual oddball task from 57 healthy young adults. Specifically, our goal was to evaluate whether behavioral effects, higher mean responses to infrequent (target) versus frequent (standard) stimuli, were present across participants. We also determined if the P300 component obtained from the ERP waveform on the midline and increases in measured HbO over the right PFC, as measured by fNIRS for target stimuli as compared to the standard ones, can be simultaneously captured using the visual oddball task as implemented in our mobile app WearCAAT. All desired features were elicited using an abbreviated visual oddball task on a mobile platform, which demonstrated the validity of WearCAAT functionality and synchrony for functional neuroimaging studies.

While future work entails the validation of more tasks implemented in the current iteration of WearCAAT, and comparisons of fNIRS and EEG features for young versus older adults, this work supports the use of mobile platforms for cognitive neuroimaging.

WearCAAT will soon be easily accessible through both Google Play and Apple App Stores. It is our hope that the wide range of reconfigurable neurocognitive tasks, usability, and ease of

XSL•FO
RenderX

use with extant neuroimaging setups will enable nontechnical users to leverage mobile pocket laboratories in future studies and begin to answer outstanding questions in ecological validity. We believe the validation of technical ability as reported in this experiment lends confidence to the pocket lab paradigm and informs future studies into human behavior, in and out in the wild.

Funding

This research was supported by a grant from the National Institutes of Health (grant R01AG077018) awarded to MI. The funding source has no role in the design, data collection and analysis, interpretation of the study, writing, or decision to submit for publication.

Authors' Contributions

PR contributed to app and task designs, app development, writing the original draft, writing, reviewing, and editing. MI contributed to the study, task and app designs, data analysis, interpretation, writing, reviewing, and editing. LG contributed to data collection, data analysis and interpretation, writing, reviewing, and editing. RH contributed to study and task designs, writing, reviewing, and editing.

Conflicts of Interest

None declared.

References

1. Khanna N, Altmeyer W, Zhuo J, Steven A. Functional neuroimaging: fundamental principles and clinical applications. Neuroradiol J 2015 Apr;28(2):87-96. [doi: 10.1177/1971400915576311] [Medline: 25963153]
2. Holtzer R, Epstein N, Mahoney JR, Izzetoglu M, Blumen HM. Neuroimaging of mobility in aging: a targeted review. J Gerontol A Biol Sci Med Sci 2014 Nov;69(11):1375-1388. [doi: 10.1093/gerona/glu052] [Medline: 24739495]
3. Neisser U. Cognition and Reality: Principles and Implications of Cognitive Psychology: W.H. Freeman; 1976. URL: http://archive.org/details/cognitionreality00neis [accessed 2024-05-01]
4. Bronfenbrenner U. Toward an experimental ecology of human development. American Psychologist 1977;32(7):513-531. [doi: 10.1037//0003-066X.32.7.513]
5. von Lühmann A, Zheng Y, Ortega-Martinez A, et al. Towards Neuroscience of the Everyday World (NEW) using functional near-infrared spectroscopy. Curr Opin Biomed Eng 2021 Jun;18:100272. [doi: 10.1016/j.cobme.2021.100272] [Medline: 33709044]
6. Jungnickel E, Gramann K. Mobile Brain/Body Imaging (MoBI) of physical interaction with dynamically moving objects. Front Hum Neurosci 2016;10:306. [doi: 10.3389/fnhum.2016.00306] [Medline: 27445747]
7. Wu T, Sherman G, Giorgi S, et al. Smartphone sensor data estimate alcohol craving in a cohort of patients with alcohol-associated liver disease and alcohol use disorder. Hepatol Commun 2023 Dec 1;7(12):12. [doi: 10.1097/HC9.0000000000000329] [Medline: 38055637]
8. Coughlan G, Coutrot A, Khondoker M, Minihane AM, Spiers H, Hornberger M. Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. Proc Natl Acad Sci U S A 2019 May 7;116(19):9285-9292. [doi: 10.1073/pnas.1901600116] [Medline: 31015296]
9. Taylor JC, Heuer HW, Clark AL, et al. Feasibility and acceptability of remote smartphone cognitive testing in frontotemporal dementia research. Alzheimers Dement (Amst) 2023;15(2):e12423. [doi: 10.1002/dad2.12423] [Medline: 37180971]
10. Hackett K, Xu S, McKniff M, Paglia L, Barnett I, Giovannetti T. Mobility-based smartphone digital phenotypes for unobtrusively capturing everyday cognition, mood, and community life-space in older adults: feasibility, acceptability, and preliminary validity study. JMIR Hum Factors 2024 Nov 22;11:e59974. [doi: 10.2196/59974] [Medline: 39576984]
11. Zawada S, Acosta J, Collins C, et al. Real-world smartphone data predicts mood after ischemic stroke and transient ischemic attack symptoms and may constitute digital endpoints: a proof-of-concept study. Mayo Clin Proc Digit Health 2025 Sep;3(3):100240. [doi: 10.1016/j.mcpdig.2025.100240] [Medline: 40881107]
12. Degenhard J. Number of smartphone users worldwide from 2014 to 2029. Statista. URL: https://www.statista.com/forecasts/1143723/smartphone-users-in-the-world [accessed 2024-05-16]
13. Hodes RJ, Insel TR, Landis SC, NIH Blueprint for Neuroscience Research. The NIH toolbox: setting a standard for biomedical research. Neurology (ECronicon) 2013 Mar 12;80(11 Suppl 3):S1. [doi: 10.1212/WNL.0b013e3182872e90] [Medline: 23479536]
14. Fox RS, Zhang M, Amagai S, et al. Uses of the NIH Toolbox® in clinical samples. Neur Clin Pract 2022 Aug;12(4):307-319. [doi: 10.1212/CPJ.0000000000200060]
15. Siegler AJ, Knox J, Bauermeister JA, Golinkoff J, Hightow-Weidman L, Scott H. Mobile app development in health research: pitfalls and solutions. Mhealth 2021;7:32. [doi: 10.21037/mhealth-19-263] [Medline: 33898601]

XSL•FO
RenderX

16. Kothe C, Shirazi SY, Stenner T, et al. The lab streaming layer for synchronized multimodal recording. Neuroscience. Preprint posted online on Jul 14, 2024. [doi: 10.1101/2024.02.13.580071]
17. Blum S, Hölle D, Bleichner MG, Debener S. Pocketable labs for everyone: synchronized multi-sensor data streaming and recording on smartphones with the lab streaming layer. Sensors (Basel) 2021 Dec 5;21(23):8135. [doi: 10.3390/s21238135] [Medline: 34884139]
18. BRANY. URL: https://www.brany.com/ [accessed 2024-11-27]
19. McCarthy G, Luby M, Gore J, Goldman-Rakic P. Infrequent events transiently activate human prefrontal and parietal cortex as measured by functional MRI. J Neurophysiol 1997 Mar;77(3):1630-1634. [doi: 10.1152/jn.1997.77.3.1630] [Medline: 9084626]
20. Izzetoglu M, Bunce SC, Izzetoglu K, Onaral B, Pourrezaei AK. Functional brain imaging using near-infrared technology. IEEE Eng Med Biol Mag 2007;26(4):38-46. [doi: 10.1109/memb.2007.384094] [Medline: 17672230]
21. Patel SH, Azzam PN. Characterization of N200 and P300: selected studies of the event-related potential. Int J Med Sci 2005;2(4):147-154. [doi: 10.7150/ijms.2.147] [Medline: 16239953]
22. Ardekani BA, Choi SJ, Hossein-Zadeh GA, et al. Functional magnetic resonance imaging of brain activity in the visual oddball task. Cognitive Brain Research 2002 Nov;14(3):347-356. [doi: 10.1016/S0926-6410(02)00137-4]
23. Rizer W, Aday JS, Carlson JM. Changes in prefrontal cortex near infrared spectroscopy activity as a function of difficulty in a visual P300 paradigm. J Near Infrared Spectrosc 2018 Aug;26(4):222-228. [doi: 10.1177/0967033518791320]
24. Classon I. Migrating From Xamarin.Forms to .NET MAUI: A Comprehensive Guide: Apress; 2025. [doi: 10.1007/979-8-8688-1215-6]
25. Kay M, Rector K, Consolvo S, et al. PVT-touch: adapting a reaction time test for touchscreen devices. : IEEE Presented at: ICTs for improving Patients Rehabilitation Research Techniques; May 5-8, 2013; Venice, Italy. [doi: 10.4108/pervasivehealth.2013.252078]
26. Evans MS, Harborne D, Smith AP. Developing an objective indicator of fatigue: an alternative mobile version of the psychomotor vigilance task (m-PVT). In: Longo L, Leva MC, editors. Human Mental Workload: Models and Applications: Springer; 2019:49-71. [doi: 10.1007/978-3-030-14273-5_4]
27. Verbruggen F, Logan GD. Automatic and controlled response inhibition: associative learning in the go/no-go and stop-signal paradigms. J Exp Psychol Gen 2008 Nov;137(4):649-672. [doi: 10.1037/a0013170] [Medline: 18999358]
28. Kane MJ, Conway ARA, Miura TK, Colflesh GJH. Working memory, attention control, and the N-back task: a question of construct validity. J Exp Psychol Learn Mem Cogn 2007 May;33(3):615-622. [doi: 10.1037/0278-7393.33.3.615] [Medline: 17470009]
29. Owen AM, McMillan KM, Laird AR, Bullmore E. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum Brain Mapp 2005 May;25(1):46-59. [doi: 10.1002/hbm.20131] [Medline: 15846822]
30. Spanakis P, Jones A, Field M, Christiansen P. A Stroop in the hand is worth two on the laptop: superior reliability of a smartphone based alcohol Stroop in the real world. Subst Use Misuse 2019;54(4):692-698. [doi: 10.1080/10826084.2018.1536716] [Medline: 30572780]
31. Yeung MK, Lee TL, Chan AS. Neurocognitive development of Flanker and Stroop interference control: a near-infrared spectroscopy study. Brain Cogn 2020 Aug;143:105585. [doi: 10.1016/j.bandc.2020.105585] [Medline: 32535484]
32. Pronk T, Hirst RJ, Wiers RW, Murre JMJ. Can we measure individual differences in cognitive measures reliably via smartphones? A comparison of the flanker effect across device types and samples. Behav Res Methods 2023 Jun;55(4):1641-1652. [doi: 10.3758/s13428-022-01885-6] [Medline: 35710865]
33. Nyhus E, Barceló F. The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: a critical update. Brain Cogn 2009 Dec;71(3):437-451. [doi: 10.1016/j.bandc.2009.03.005] [Medline: 19375839]
34. Wang LM, Huang YH, Chou PH, Wang YM, Chen CM, Sun CW. Characteristics of brain connectivity during verbal fluency test: convolutional neural network for functional near-infrared spectroscopy analysis. J Biophotonics 2022 Jan;15(1):1. [doi: 10.1002/jbio.202100180] [Medline: 34553833]
35. Broster LS, Li J, Wagner B, et al. Spared behavioral repetition effects in Alzheimer's disease linked to an altered neural mechanism at posterior cortex. J Clin Exp Neuropsychol 2018 Sep 14;40(8):761-776. [doi: 10.1080/13803395.2018.1430230]
36. Nguyen T, Babawale O, Kim T, Jo HJ, Liu H, Kim JG. Exploring brain functional connectivity in rest and sleep states: a fNIRS study. Sci Rep 2018 Nov;8(1):16144. [doi: 10.1038/s41598-018-33439-2]
37. BioRender. URL: https://BioRender.com [accessed 2026-01-26]
38. Mobile EEG-fNIRS – LiveAmp with actiCAP electrodes and NIRx NIRSport2. Brain Products. 2020. URL: https://pressrelease.brainproducts.com/eeg-fnirs-setup/ [accessed 2024-01-19]
39. Concurrent fNIRS and EEG. NIRx Medical Technologies. URL: https://nirx.net/fnirs-eeg [accessed 2024-01-18]
40. Customized NIRScaps and probes. NIRx Medical Technologies. URL: https://nirx.net/nirscap [accessed 2025-05-05]
41. Kothe C. Labstreaminglayer/app-labrecorder. GitHub. 2024 May 9. URL: https://github.com/labstreaminglayer/App-LabRecorder [accessed 2024-05-15]
42. Woods D, Yund W, Pebler P, Grivich MI. The timing precision of iOS and Android apps. Neurobehavioral Systems, Inc. URL: https://cdn1.neurobs.com/misc/mobile_timing_poster.pdf [accessed 2026-01-19]
43. MATLAB version: 9.13.0 (r2022b). The MathWorks Inc. URL: https://www.mathworks.com [accessed 2025-05-15]

XSL•FO
RenderX

44. Huang R, Hong KS, Yang D, Huang G. Motion artifacts removal and evaluation techniques for functional near-infrared spectroscopy signals: a review. Front Neurosci 2022;16:878750. [doi: 10.3389/fnins.2022.878750] [Medline: 36263362]
45. Abdelnour AF, Huppert T. Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. Neuroimage 2009 May 15;46(1):133-143. [doi: 10.1016/j.neuroimage.2009.01.033] [Medline: 19457389]
46. Li R, Yang D, Fang F, Hong KS, Reiss AL, Zhang Y. Concurrent fNIRS and EEG for brain function investigation: a systematic, methodology-focused review. Sensors (Basel) 2022 Aug;22(15):5865. [doi: 10.3390/s22155865]
47. Habibzadeh Tonekabony Shad E, Molinas M, Ytterdal T. Impedance and noise of passive and active dry EEG electrodes: a review. IEEE Sensors J 2020 Dec;20(24):14565-14577. [doi: 10.1109/JSEN.2020.3012394]
48. Pollonini L, Bortfeld H, Oghalai JS. PHOEBE: a method for real time mapping of optodes-scalp coupling in functional near-infrared spectroscopy. Biomed Opt Express 2016 Dec 1;7(12):5104-5119. [doi: 10.1364/BOE.7.005104] [Medline: 28018728]
49. Luck SJ. An Introduction to the Event-Related Potential Technique, 2nd edition: MIT Press; 2014.
50. Delorme A, Makeig S. Extract data epochs. EEGLAB Wiki. URL: https://eeglab.org/tutorials/07_Extract_epochs/Extracting_Data_Epochs.html [accessed 2024-06-27]
51. Yücel MA, Selb J, Aasted CM, et al. Mayer waves reduce the accuracy of estimated hemodynamic response functions in functional near-infrared spectroscopy. Biomed Opt Express 2016;7(8):3078. [doi: 10.1364/BOE.7.003078]
52. Bunce SC, Izzetoglu M, Izzetoglu K, Onaral B, Pourrezaei K. Functional near-infrared spectroscopy. IEEE Eng Med Biol Mag 2006;25(4):54-62. [doi: 10.1109/memb.2006.1657788] [Medline: 16898659]
53. Uga M, Dan I, Sano T, Dan H, Watanabe E. Optimizing the general linear model for functional near-infrared spectroscopy: an adaptive hemodynamic response function approach. Neurophotonics 2014 Jul;1(1):015004. [doi: 10.1117/1.NPh.1.1.015004] [Medline: 26157973]
54. Polich J, Bondurant T. P300 sequence effects, probability, and interstimulus interval. Physiol Behav 1997 Jun;61(6):843-849. [doi: 10.1016/s0031-9384(96)00564-1] [Medline: 9177555]
55. Polich J. Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 2007 Oct;118(10):2128-2148. [doi: 10.1016/j.clinph.2007.04.019] [Medline: 17573239]
56. Muñoz V, Muñoz-Caracuel M, Angulo-Ruiz BY, Gómez CM. Neurovascular coupling during auditory stimulation: event-related potentials and fNIRS hemodynamic. Brain Struct Funct 2023 Nov;228(8):1943-1961. [doi: 10.1007/s00429-023-02698-9] [Medline: 37658858]
57. Poskotinova L, Khasanova N, Kharak A, Krivonogova O, Krivonogova E. Parameters of auditory evoked related potentials p300 in disorders of different cognitive function domains (visuospatial/executive and memory) in elderly hypertensive persons. Diagnostics (Basel) 2023 Apr 30;13(9):1598. [doi: 10.3390/diagnostics13091598] [Medline: 37174989]

Abbreviations

BRANY: Biomedical Research Alliance of New York
Cz: central midline
EEG: electroencephalogram
ERP: event-related potential
fNIRS: functional near-infrared spectroscopy
HbO: oxygenated hemoglobin
LSL: lab streaming layer
PFC: prefrontal cortex
Pz: parietal midline
ROI: region of interest
RT: response time
WearCAAT: Wearable Cognitive Assessment and Augmentation Toolkit

Edited by A Zampogna; submitted 28.May.2025; peer-reviewed by M Patera, S Zawada; revised version received 31.Oct.2025; accepted 11.Nov.2025; published 06.Feb.2026.

Please cite as:
Rokowski P, Izzetoglu M, Gomero L, Holtzer R
A Pocket Laboratory for Functional Neuroimaging Research Using Mobile Visual Oddball, Multimodal Electroencephalography, and Functional Near-Infrared Spectroscopy Imaging: Instrument Validation Study
JMIR Neurotech 2026;5:e78217
URL: https://neuro.jmir.org/2026/1/e78217
doi:10.2196/78217

XSL•FO
RenderX

© Peter Rokowski, Meltem Izzetoglu, Luis Gomero, Roee Holtzer. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 6.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Feasibility of Interleaving Computerized Cognitive Training With Repetitive Transcranial Magnetic Stimulation: Pilot Studies in Mild Cognitive Impairment Due to Alzheimer's Disease and Stroke

Laura M Campbell[1,2], PhD; Stephanie Aghamoosa[3], PhD; Olivia Horn[4], BS; Holly H Fleischmann[2,5], MS; James Lopez[2], BS; Katrina S Rbeiz[4], MS; Katrina Madden[4], MBA; Michael Antonucci[6], MD; Gonzalo Revuelta[4], MD; Lisa M McTeague[2], PhD; Andreana Benitez[4,6], PhD

[1]Department of Psychiatry, University of California, San Diego, La Jolla, CA, United States

[2]Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, Charleston, SC, United States

[3]Department of Health Sciences and Research, Medical University of South Carolina, Charleston, SC, United States

[4]Department of Neurology, Medical University of South Carolina, Charleston, SC, United States

[5]Department of Psychology, University of Georgia, Athens, Georgia, United States

[6]Department of Radiology and Radiological Science, Medical University of South Carolina, 96 Jonathan Lucas St, Charleston, SC, United States

Corresponding Author:
Andreana Benitez, PhD
Department of Neurology, Medical University of South Carolina, Charleston, SC, United States

Abstract

Background: Repetitive transcranial magnetic stimulation induces neural plasticity, which may be maximized via adjunctive interventions.

Objective: This study aimed to examine the feasibility and acceptability of interleaving computerized cognitive training (CCT) with accelerated intermittent theta burst (iTBS) repetitive transcranial magnetic stimulation in 2 open-label phase I trials of amnestic mild cognitive impairment (aMCI; NCT04503096) and poststroke mild cognitive impairment (psMCI; NCT04655963).

Methods: Participants (aMCI: n=21; and psMCI: n=14) received 24 sessions of accelerated iTBS over 3 days (8 sessions per day). During the 10 - to 15-minute intervals between iTBS sessions, participants engaged in CCT via the BrainHQ platform. During the 4-week follow-up period, participants were asked to complete 20 minutes daily of BrainHQ. Both studies offered identical compensation for daily practice, but the aMCI group received weekly reminders, whereas the psMCI group did not. Study-specific questionnaires assessed BrainHQ acceptability (administered post-treatment and at follow-up).

Results: During in-laboratory treatment, the groups completed similar amounts of BrainHQ (aMCI: median 107, IQR 92-121 min; and psMCI: median 111, IQR 93-137 min; *P*=.51). However, engagement differed during the unmonitored follow-up (aMCI: median 450, IQR 167-581 min,; and psMCI: median 142, IQR 0-477 min); practice time increased in the aMCI group but decreased in the psMCI group over the 4 weeks (*P*=.001). Both groups generally rated BrainHQ positively at both time points in terms of their experience with the platform, exercises, and perceived benefit for their mood and functioning.

Conclusions: Combined treatment with CCT and accelerated iTBS is feasible and acceptable in people with mild cognitive impairment. Standardized dosing of CCT is feasibly achieved when interleaved with iTBS sessions and completed in-laboratory, but ecological adherence to CCT in an unmonitored, self-directed context likely requires additional supports and may be influenced by patient characteristics.

Trial Registration: ClinicalTrials.gov NCT04503096; https://clinicaltrials.gov/study/NCT04503096 and NCT04655963; https://clinicaltrials.gov/study/NCT04655963

(JMIR Neurotech 2026;5:e81437) doi:10.2196/81437

KEYWORDS

cognitive remediation; cognitive training; mild cognitive impairment; Alzheimer's disease; stroke; neuromodulation; transcranial magnetic stimulation; mHealth; feasibility studies; older adults

XSL•FO
RenderX

Introduction

Repetitive transcranial magnetic stimulation (rTMS) of the left dorsolateral prefrontal cortex (l-dlPFC) is a non-invasive brain stimulation technique that can induce neuroplasticity [1]. While rTMS has primarily been studied within psychiatric disorders (eg, treatment-refractory depression), there is growing interest in its use as a treatment for cognitive impairments in neurological disorders, such as mild cognitive impairment (MCI) due to Alzheimer's disease (AD) [2] and stroke [3,4]. The evidence base in these neurological conditions is nascent but steadily growing. Early meta-analyses of sham-controlled trials of rTMS to enhance cognitive functions in these conditions have found medium-to-large effect sizes for MCI and AD dementia [2,4,5] and more modest but nonetheless positive effects in stroke [3,4,6,7]. This is particularly promising, given that many other therapies have limited efficacy in these populations [8-10].

While these results are encouraging, there remains a significant need to optimize the efficacy of rTMS to l-dlPFC for cognitive impairment. One approach for doing so is to incorporate an adjunctive treatment to enhance its effects. As rTMS can induce neuroplasticity in targeted cognitive circuitry [11], incorporating computerized cognitive training (CCT) as an adjunctive intervention may work synergistically to enhance cognitive gains. CCT uses a restorative approach to cognitive remediation through repeated practice of tasks or exercises in specific cognitive domains using a technological device (eg, a computer or tablet) [12].

CCT has been studied within the context of cognitive impairment due to AD and stroke. With adequate training and optimization of the CCT to fit the needs of the intended population, CCT in middle-aged and older adults with and without MCI is feasible [13-15]. However, the efficacy of these interventions for MCI is less clear, as some studies report improvements in cognition, albeit with small effect sizes, whereas other studies do not [16-20]. Likely sources of this variability in findings are the heterogeneity of research design features, such as CCT platform, exercises, and dose. Notably, a recent meta-analysis of randomized or quasi-randomized trials with healthy older adults demonstrated that 2 platforms, BrainHQ and CogniFit—in contrast to other commercially available options—have the highest quality evidence supporting their efficacy [21].

CCT that trains specific cognitive domains is presumed to enhance performance on similar tasks via neuroplasticity [16]. As rTMS has also been shown to induce neuroplasticity [22], it is therefore plausible that combining CCT with rTMS may yield greater cognitive gains than either rTMS or CCT alone. CCT is a particularly attractive complement to rTMS, given that it poses minimal clinical burden, is cost-effective and low risk, and can be self-administered in between rTMS sessions. However, to date, CCT and rTMS have mostly been studied in isolation, and there is only limited literature examining the combination of these interventions [23]. Moreover, the majority of studies have focused on once-daily rTMS over weeks, which can be quite burdensome, as it requires daily trips to a laboratory or clinic for treatment. Newer rTMS protocols, such as accelerated intermittent theta burst stimulation (iTBS), which delivers similar or higher doses than conventional rTMS in as little as 3 multi-session days [24], are better suited for combination with CCT, as it can be interleaved between same-day iTBS sessions. Critically, the feasibility and acceptability of combined accelerated iTBS and CCT have yet to be established, which is an important first step. For example, the common, albeit mild and transient, side effects of rTMS (eg, headache [25]) may impact engagement with CCT; therefore, it should not be assumed that the feasibility and acceptability of each intervention in isolation apply to their joint delivery.

To this end, we conducted 2 open-label phase I clinical trials of accelerated iTBS to the l-dlPFC with interleaved CCT via the empirically supported BrainHQ platform in adults with amnestic MCI (aMCI) due to possible AD (NCT04503096) or poststroke MCI (psMCI; NCT04655963). While the primary aim of these studies, with parallel designs, was to establish the safety of accelerated iTBS treatment in these groups, an exploratory aim was to examine the feasibility and acceptability of interleaved CCT. Accelerated iTBS was delivered with multiple same-day sessions separated by 10 - to 15-minute intersession intervals, making it ideally suited for combination with CCT. As such, we administered BrainHQ in a standard, supervised manner during the intervals between 8 accelerated iTBS sessions on each of 3 treatment days. At post-treatment, participants were asked to independently engage in BrainHQ for 20 minutes daily over a 4-week follow-up period. The purpose of including this follow-up phase was to assess ecological adherence to CCT outside of the monitored laboratory setting. This can inform the design of future trials that are likely to include CCT engagement outside of iTBS treatment visits, capitalizing on the post-iTBS window of neuroplasticity and enabling an increase in CCT dose to potentially efficacious ranges.

Here, we report data on feasibility (ie, retention and adherence), engagement (ie, usage patterns and types of programs and exercises completed), and acceptability ratings observed over the treatment phase and the 4-week follow-up period. Due to the standardized administration of BrainHQ during accelerated iTBS treatment, a post-treatment compensation schedule that incentivized use, modest daily use expectations, and a comparable degree of cognitive function between groups (ie, no dementia), we hypothesized that engagement in BrainHQ during and post-treatment would be comparable across both MCI groups. We similarly expected that interleaved BrainHQ would be rated as acceptable on questionnaires that assess participants' experiences.

Methods

Ethical Considerations

Both studies were approved by the Medical University of South Carolina's (MUSC) Institutional Review Board and were pre-registered in ClinicalTrials.gov (aMCI: NCT04503096; and psMCI: NCT04655963; IRB Pro00100536 and Pro00083136, respectively). All participants provided written informed consent

XSL•FO
RenderX

and were informed of their right to withdraw from the study at any time without penalty. Participant privacy and confidentiality were strictly maintained in accordance with IRB requirements, and no personally identifiable information is reported in this manuscript or supplementary materials. Participants were offered compensation at the rate of US $4.50 per completed 20-minute session per day, equating to up to US $126 total. Efforts to reduce potential sources of bias in these open-label trials included (1) minimizing selection bias by using rigorous and predefined eligibility criteria applied consistently to all participants; (2) minimizing confounding factors by constraining disease severity to MCI and excluding comorbidities likely to impact the study outcomes; and (3) avoiding reporting bias through trial pre-registration and a transparent and complete description of the study procedures, outcome measures, hypotheses, planned analyses, and results. The primary results from the aMCI trial focused on safety, feasibility, and acceptability of iTBS are reported in a study by Aghamoosa et al [26]. The description of the BrainHQ methods adheres to the Template for Intervention Description and Replication (TIDieR) checklist [27], as described below and summarized in Table S1 in Multimedia Appendix 1.

Participants

Overview

Trained coordinators screened potential participants for eligibility over the phone and using the participant's electronic health record, with appropriate permissions. The following 2 paragraphs describe the eligibility criteria that were unique to each study, and the last paragraph describes the criteria shared by both studies. Target sample sizes were selected to generate estimates of the trials' primary outcomes (ie, accelerated iTBS treatment feasibility, acceptability, and outcome variability) for use in planning subsequent trials.

aMCI Participants

Twenty-two participants enrolled in the study and initiated treatment, and 21 completed all procedures. Participants were recruited from MUSC Neurology and Neuropsychology outpatient clinics. Participants had to be aged between 60 and 85 years and diagnosed with MCI due to possible AD by a neurologist or neuropsychologist within the last 2 years, using National Institute on Aging and Alzheimer's Association (NIA-AA) criteria [28]. They also needed to meet actuarial neuropsychological criteria for aMCI upon clinical assessment, which requires impairment (≤16th percentile using demographically corrected norms) in ≥2 scores within one cognitive domain or ≥1 score in ≥3 domains, one of which being memory [29]. The primary suspected cause had to be possible or probable AD, with other primary, competing etiologies (eg, psychiatric, movement disorder, stroke, and reversible causes) ruled out.

psMCI Participants

Twenty-one participants enrolled in the study; 15 (71%) initiated treatment, and 14 (67%) completed all procedures. Participants were recruited from the National Institutes of Health (NIH)–funded COBRE for Stroke Recovery Registry at MUSC and from the community (eg, advertisements, community events, and social media). All participants had to be aged ≥40 years and have a history of ischemic or hemorrhagic stroke (with intact cortex under the transcranial magnetic stimulation coil) with at least 6-month chronicity. Acute or subacute hemorrhage or any age bilateral hemorrhage or hemorrhage with mass effect, confirmed with scanning, was exclusionary. Participants had to retain the ability to perform the cognitive tasks (eg, no moderate to severe global aphasia or visual impairment). Although not required for enrollment, we retrospectively confirmed that all psMCI participants met actuarial neuropsychological criteria for MCI [29] on the neuropsychological testing completed during the study.

For both groups, contraindications to magnetic resonance imaging (MRI) or rTMS were exclusionary, including electrically, magnetically, or mechanically activated metal or nonmetal implants; pregnancy; claustrophobia; history of a seizure disorder; or preexisting scalp lesion, wound, bone defect, or hemicraniectomy. Additional exclusionary criteria included a clinical diagnosis of major neurocognitive disorder or dementia, history of significant or unstable medical conditions that could have affected cognition (eg, lifetime substance or alcohol use disorder, severe mental illness, other neurological disorders, severe cardiometabolic disorders, and developmental disorders), no detectable active or passive motor threshold in the paretic or nonparetic right hand, and current enrollment in a clinical trial that involved any investigational treatment. Participants were excluded if they endorsed daily or weekly use of anticholinergics, neuroleptics, sedative-hypnotics, bupropion, or stimulants (unless stimulant use was deemed low risk by the coinvestigator physician safety monitor). Use of cholinesterase inhibitors, N-methyl-D-aspartate (NMDA) receptor antagonists, and antidepressants was permitted if the participant was on a stable regimen for at least 4 weeks prior to enrollment. Participants were required to have English as their first or primary language.

Procedures

Study Design Overview

These 2 pilot studies were designed to have parallel procedures, although they were not conducted simultaneously and were not identical. There were minor differences in some study procedures (eg, experience questionnaire and delivery of BrainHQ post-treatment) arising from differences in study aims and refinements to the aMCI study based on lessons learned from the psMCI study, which started first. These trials were open-label phase I studies (see a study by Aghamoosa et al [26] for details). Potential participants underwent screening to determine eligibility and provided written informed consent. All in-person procedures were conducted in research-designated space at an academic medical center. Participants completed the testing of resting motor threshold at pretreatment. Pre- and posttreatment assessments (including self-reported demographic characteristics and medical history, neuropsychological testing, neuropsychiatric interviews and questionnaires, and brain MRI) were completed within 1 week of treatment initiation and termination. Participants completed 3 accelerated iTBS treatment days with interleaved CCT. In the interest of feasibility and acceptability, participants chose which 3 days within an 8-day

XSL•FO
RenderX

span to receive treatment. For the 4 weeks after treatment, participants were asked to complete BrainHQ for 20 minutes per day at home using their own device and completed weekly online questionnaires. Study iPad Minis were available if participants did not have their own device, but no participants chose this option. At the end of the 4-week monitoring period, they completed a telehealth follow-up assessment (questionnaires and a global cognitive screening measure). Participants were compensated for completion of each study visit and engagement in BrainHQ during the follow-up period (as described in the following sections). The aMCI participants completed the study between 2021 and 2022. All psMCI participants completed the study in 2021.

Accelerated iTBS Treatment

Both groups received identical accelerated iTBS treatment. Each participant completed 3 accelerated iTBS treatment days (~2 h per day), each consisting of 8 iTBS sessions with 10‐ to 15-minute inter-session intervals (24 total sessions). Treatment was delivered using a MagVenture MagPro TMS System with a figure-8 coil. We administered 600 pulses per session (14,400 total pulses) in 50 Hz triplets at 120% of resting motor threshold (2-s trains repeated every 10 s with 8-s inter-train intervals for 190 s). Stimulation was targeted to the l-dlPFC at F3 using scalp measurements from the 10 to 20 electroencephalogram (EEG) system [30]. The Brainsight NeuroNavigation System (Rogue Research Inc) was used to ensure precise and consistent coil placement across sessions.

CCT via BrainHQ

Participants engaged in Posit Science's online BrainHQ exercises (version 7.13; Posit Science Corporation). During the in-laboratory treatment phase, BrainHQ was delivered in 10‐ to 15-minute intervals between each of the 8 accelerated iTBS sessions (ie, 7 BrainHQ sessions per treatment day for 210‐315 total minutes) on a study-provided iPad (10.2-inch iPad, 8th generation, iOS 16.4; Apple Inc). After each accelerated iTBS session, study staff handed the iPad to participants to complete the BrainHQ exercises and remained present to answer any questions. During the 4-week follow-up period, participants were asked to independently complete at least 20 minutes of BrainHQ exercises per day (7 sessions per week, 28 total sessions, and 560 total minutes) on their own device (ie, computer, tablet, or smartphone) in their preferred location.

The BrainHQ platform is commercially available and includes exercises in 6 BrainHQ-defined categories: attention, brain speed, memory, people skills, intelligence, and navigation. Most exercises require an upper extremity motor response (eg, touch screen or mouse click), and many require speeded responses even if they are not in the "brain speed" domain; hence, there is some overlap in the cognitive processes being trained across domains. Participants are provided feedback on their performance on each exercise, and the exercises are adaptive, becoming more difficult if performance improves. BrainHQ records the number of seconds engaged in each exercise. A detailed description of each exercise and additional information about the BrainHQ platform can be found on the website provided earlier.

Participants were shown how to use the BrainHQ platform by study staff, who were available to answer questions as needed. Additionally, the BrainHQ platform includes instructions and a sample item for each exercise. During the treatment phase, participants were instructed to use the adaptive "personal trainer" option that selects exercises and difficulty levels based on past performance. During the 4-week follow-up phase, aMCI participants were instructed that they could use either the personal trainer or the "a la carte" option, in which participants select which exercise they would like to complete. psMCI participants were instructed to continue to use the personal trainer during the follow-up phase, but they did have access to the à la carte option. CCT during the follow-up phase was self-administered at the participants' preferred location and was not actively monitored but was recorded via the BrainHQ platform. In the aMCI study, participants were reminded to complete the BrainHQ activities via an automated weekly email. The psMCI study did not send BrainHQ reminders. BrainHQ was not tailored beyond the built-in adaptivity, and no modifications were made to the BrainHQ protocol during the study.

Measures

Sample Characterization

The NINDS (National Institute of Neurological Disorders and Stroke) and Canadian Stroke Network neuropsychological test battery [31] was used to characterize baseline cognitive function. This battery includes the Hopkins Verbal Learning Test-Revised, Digit Span Forward and Backward, Digit Symbol Coding, Trail Making Tests A & B, semantic (animal) fluency, and phonemic (letter) fluency. Age-adjusted norms provided in the test's manual were used to correct Hopkins Verbal Learning Test-Revised scores (immediate recall, delayed recall, retention, and recognition), while other scores were corrected for age, sex, and years of education using normative calculators published in the literature (ie, Digit Span and Digit Symbol Coding scores [32] and Trail Making Tests and verbal fluencies [33]). We report normative percentile scores for each test.

Feasibility

We assessed the feasibility of combining BrainHQ with accelerated iTBS by evaluating study retention and adherence. Retention was measured as the percentage of participants who initiated treatment and completed all study procedures. Adherence to BrainHQ was assessed separately for each MCI group and for the in-laboratory treatment phase and 4-week follow-up phase, the latter being referred to as ecological adherence. For the treatment phase, adherence was measured as the percentage of prescribed minutes completed out of the total 210 minutes across the 3 treatment days. We could not compute adherence to the number of same-day sessions during the treatment phase as the data were labeled by treatment day, not individual sessions, precluding the enumeration of completed sessions. For the follow-up phase, ecological adherence was measured as the percentage of target minutes completed out of the target total of 560 minutes. We also report the percentage of participants completing all 28 daily sessions and the number of completed sessions (ie, how many days trained) over the 4-week follow-up period.

XSL•FO
RenderX

Engagement

To explore how participants interacted with BrainHQ, we evaluated several aspects of engagement across the 2 MCI groups. These included the overall patterns of practice time, the types of exercises completed, and the devices used to complete practice during the treatment and follow-up phases.

Acceptability

To evaluate the acceptability of BrainHQ, participants were asked to complete study-specific questionnaires that solicited feedback on their experience. Each questionnaire was administered immediately after the treatment phase and again at the conclusion of the 4-week follow-up period. The questionnaire used in the psMCI study was developed first and then modified for the aMCI study, considering feedback from the initial questionnaire and differences in the type of experience data sought by each study. Both questionnaires inquired about the impact of BrainHQ on functioning, participants' experiences (such as enjoyment and frustration) with the exercises, and the overall usability experience with the BrainHQ platform.

Statistical Analyses

Differences in demographic and clinical characteristics between groups were examined using Wilcoxon rank-sum tests and chi-squared or Fisher exact tests, as appropriate. We present descriptive statistics to summarize adherence, engagement, and acceptability for each group. Wilcoxon rank-sum tests were used to compare the total number of minutes of engagement with the BrainHQ exercises between the aMCI and psMCI groups during the supervised treatment phase. During the unsupervised follow-up phase, a rank-based nonparametric alternative to a repeated-measures ANOVA was performed using the *nparLD 2.2* package in R (4.4.2), which produces ANOVA-type statistics (ATS) [34]. This analysis tested whether there was a difference in the number of minutes of engagement in BrainHQ by the MCI group across all weeks as well as whether engagement changed across the weeks by group (ie, a group-by-week interaction).

Results

Participants

Demographic and cognitive characteristics are summarized in Table 1. Consistent with the differing age range criteria in the 2 studies, participants in the aMCI group were older than those in the psMCI group. The groups did not differ significantly by sex, and both groups were predominantly non-Hispanic White. Consistent with their distinct underlying etiologies, the 2 MCI groups demonstrated differences in cognitive functioning (ie, poorer memory and semantic fluency in the aMCI group and poorer processing speed in the psMCI group).

XSL•FO
RenderX

Table . Demographic and cognitive characteristics.

Demographic or cognitive characteristics	aMCI[a] (n=21)	psMCI[b] (n=14)	*P* value
Age (y), median (IQR)	72 (69.5 - 78.5)	65 (61.3 - 72)	.003[c]
Sex n (%)			.16
Male	11 (52%)	11 (79%)	
Female	10 (48%)	3 (21%)	
Education, median (IQR)	16 (14-17)	14.5 (12-16)	.19
Race or ethnicity, n (%)			.07
Non-Hispanic White	20 (95)	10 (71)	—[d]
Non-Hispanic Black or African American	1 (5)	4 (29)	—
Neuropsychological battery, median percentile (IQR)			
HVLT-R[e] immediate recall	3 (1-14)	14 (1.75 - 32.75)	.08
HVLT-R delayed recall	0 (0 - 0)	1 (0.75 - 10.5)	.001[c]
HVLT-R retention	0 (0 - 0)	5 (0 - 7)	.003[c]
HVLT-R recognition	0 (0 - 3.5)	6.5 (0 - 15)	.07
Digits forward	22 (8-37)	39.5 (12.75 - 53)	.13
Digits backward	24 (14 - 43.5)	19.5 (7 - 47.75)	.39
Trails A	12 (0 - 56)	2.5 (0 - 6.25)	.10
Trails B	2 (0 - 42.5)	10 (0 - 28.25)	≥.99
Digit symbol coding	31 (11-41)	3.5 (1 - 22.25)	.03[c]
Animal fluency	8 (2.5 - 19.5)	27.5 (12.75 - 34.25)	.006[c]
Letter fluency	28 (11.5 - 52)	38.5 (12.5 - 68.25)	.34

[a]aMCI: amnestic mild cognitive impairment.

[b]psMCI: post-stroke mild cognitive impairment.

[c]$P<.05$.

[d]Not applicable.

[e]HVLT-R: Hopkins Verbal Learning Test-Revised.

BrainHQ Feasibility

Retention was high in both studies. One treatment initiator withdrew from the aMCI group (ie, 21/22, 95.5% completed all study procedures). All participants in the psMCI group who initiated treatment completed study procedures (ie, 14/14, 100%).

Adherence to BrainHQ during the in-laboratory treatment phase was similar in the 2 groups, with both groups completing approximately half of the prescribed 210 minutes: aMCI group median: 51.1% (IQR 43.6% - 57.6%; range 12.9% - 79.5%) of minutes and psMCI group median: 52.9% (IQR 44.4% - 65.3%; range 44.1% - 78.4%; see see Table 2 for the median number of minutes). No participant in either group completed 210 minutes during the treatment phase. Ecological adherence to the target 560 minutes of BrainHQ during the 4-week follow-up phase was higher in the aMCI group (median 80.2%, IQR 29.9% - 103.7% of minutes; range: 0% - 172.4%) than in the psMCI group (median 25.3%, IQR 0% - 85.3% of minutes; range: 0% - 192.4%). The percentage of participants who completed the target 560 minutes over the 4 weeks was 33.3% in the aMCI group and 21.4% in the psMCI group. The percentage of participants who met weekly goals (ie, target 140 min/wk) was variable in the aMCI group: week 1, 2 (10%) participants; week 2, 12 (57%) participants; week 3, 5 (24%) participants; and week 4, 7 (33%) participants. The percentage of psMCI participants who completed the target 140 minutes each week was lower and relatively stable: week 1, 3 (21%) participants; week 2, 5 (36%) participants; week 3, 3 (21%) participants; and week 4, 2 (14%) participants. In both groups, no participant (ie, 0%) completed the target 28 sessions of daily BrainHQ. The aMCI group median number of sessions was 16 (IQR 9 - 21), and the median for the psMCI group was 4 (IQR 1 - 14).

XSL•FO
RenderX

Table . Median minutes of BrainHQ during repetitive transcranial magnetic stimulation (rTMS) treatment and in 4-week follow-up.

	aMCI[a] (n=21)			psMCI[b] (n=14)		
	Minutes, median (IQR)	Range	Participants who did not do Brain-HQ, n (%)	Minutes, median (IQR)	Range	Participants who did not do Brain-HQ, n (%)
rTMS treatment phase						
Treatment day 1	25.7 (19.6 - 33.9)	6.25 - 47.4	1 (5)	42.4 (32.2 - 49.7)	22.8 - 59.1	2 (14)
Treatment day 2	38.5 (32.4 - 45.9)	13.9 - 59.6	0 (0)	40.1 (27.6 - 51)	17.7 - 60.4	1 (7)
Treatment day 3	42.8 (32.2 - 52.8)	10 - 70.5	0 (0)	40.2 (27.8 - 56.3)	19.8 - 73	0 (0)
Total	107.2 (91.6–121.0)	27.2-167.0	0 (0)	111.0 (93.2-137.2)	44.1–164.7	0 (0)
Follow-up phase						
Week 1	34.3 (0 - 97.6)	0 - 178.6	7 (33)	28.6 (0 - 109.8)	0 - 282.9	5 (36)
Week 2	152.4 (71.3 - 177.5)	0 - 392.7	3 (14)	27.5 (0 - 214.9)	0 - 350.8	7 (50)
Week 3	116.4 (49.2 - 142.6)	0 - 241.9	4 (19)	0 (0 - 94.4)	0 - 269	8 (57)
Week 4	124.9 (63.2 - 163.4)	0 - 287.9	2 (10)	0 (0 - 56.2)	0 - 281.1	9 (64)
Total	449.5 (167.3-580.5)	0-965.4	1 (5)	142.0 (0-477.4)	0-1077.7	5 (36)

[a]aMCI: amnestic mild cognitive impairment.

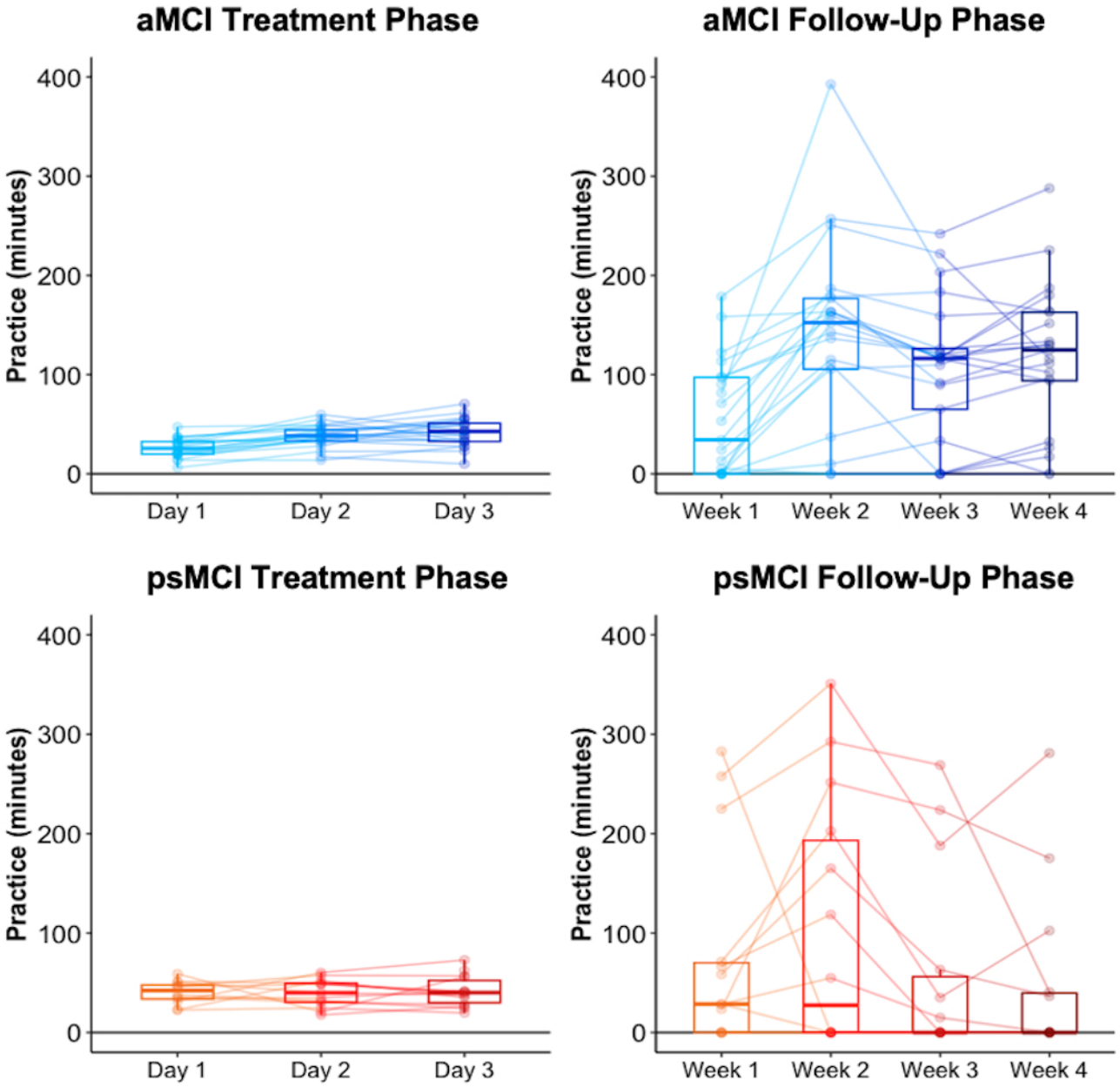
[b]psMCI: post-stroke mild cognitive impairment.

BrainHQ Engagement

The median and IQR minutes of BrainHQ exercises completed on each treatment day and during the 4-week follow-up period are included in Table 2 and depicted in Figure 1. The mean minutes spent engaging in exercise domain on each treatment day and during each week of the 4-week follow-up period are depicted in Figure 2.

XSL•FO
RenderX

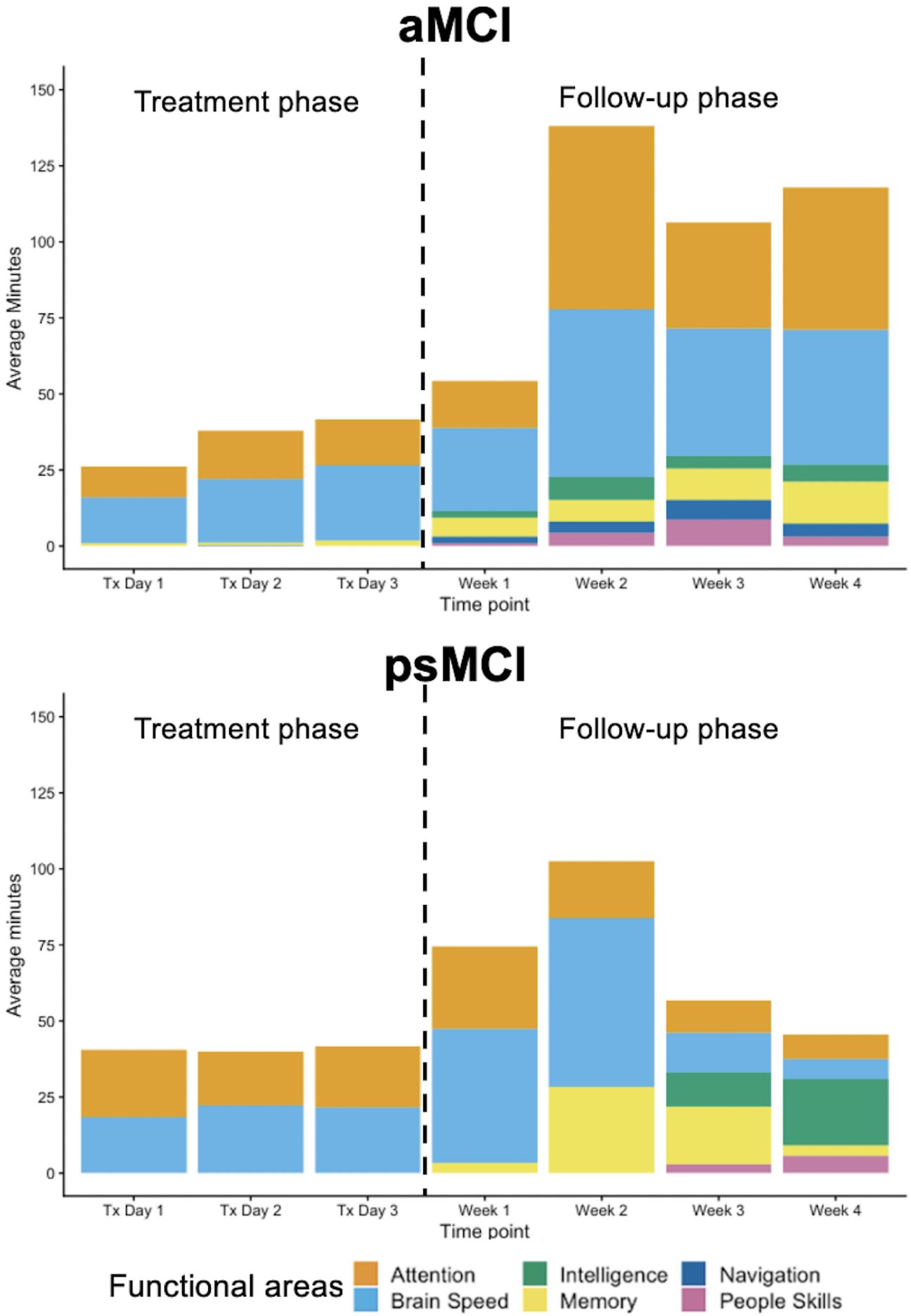
Figure 1. Median (IQR) minutes of BrainHQ engagement during repetitive transcranial magnetic stimulation treatment and in 4-week follow-up. Missing datapoints were excluded from the treatment phase figures. aMCI: amnestic mild cognitive impairment; psMCI: post-stroke mild cognitive impairment.



XSL•FO
RenderX

Figure 2. Mean minutes of BrainHQ engagement by exercise domain. aMCI: amnestic mild cognitive impairment; psMCI: post-stroke mild cognitive impairment; Tx: treatment.

XSL•FO
RenderX



XSL•FO
RenderX

Treatment Phase

Participants in the aMCI group and the psMCI group completed a comparable number of minutes of BrainHQ exercises during the treatment phase. The groups also engaged in comparable numbers of exercises per cognitive domain during the treatment phase (ie, primarily the brain speed and attention domains), but the aMCI group engaged in twice as many different exercise types. In the aMCI group, most of the completed exercises were in the domains of brain speed (59.7%) and attention (35.4%), with some engagement in the memory (4.8%) and navigation (0.1%) domains. Ten different exercises were completed. In the psMCI group, all completed exercises were in the domains of brain speed (45.6%) and attention (54.4%) and included 5 exercises. Specific exercises that comprised >5% of the exercises by group are reported in Table 3.

Table . BrainHQ exercises by domain[a].

BrainHQ domain	aMCI[b] group BrainHQ exercises	psMCI[c] group BrainHQ exercises
Treatment phase		
Brain Speed	• Eye for Detail: 20.2% • Hawk Eye: 15% • Visual Sweeps: 13.1% • Sound Sweeps: 11.4%	• Eye for Detail: 18.7% • Hawk Eye: 18.4% • Visual Sweeps: 17.2%
Attention	• Target Tracker: 19.2% • Double Decision: 15.7%	• Double Decision: 24.5% • Target Tracker: 21.1%
Follow-up phase		
Brain Speed	• Eye for Detail: 17.6% • Hawk Eye: 8.3% • Sound Sweeps: 6.6% • Visual Sweeps: 6.6%	• Eye for Detail: 12.9% • Sound Sweeps: 10.2% • Visual Sweeps: 7.9% • Fine Tuning: 6.7%
Attention	• Target Tracker: 13.2% • Double Decision: 10.4% • Divided Attention: 8.8%	• Target Tracker: 9.5% • Double Decision: 8.7%
Memory	—[d]	• Memory Grid: 7.7% • Syllable Stacks: 7.4%
Intelligence	—	• Mind Bender: 7.5% • Juggle Factor: 6.2%

[a]Exercises that comprised ≤5 of the exercises are not included.

[b]aMCI: amnestic mild cognitive impairment.

[c]psMCI: post-stroke cognitive impairment.

[d]Not applicable.

Follow-Up Phase

Twenty (95%) of 21 participants in the aMCI group completed BrainHQ exercises in the follow-up phase. In contrast, in the psMCI group, only 9 (64%) of 14 participants completed any BrainHQ exercise in the follow-up phase. The proportion of those who engaged in any BrainHQ exercise significantly differed for weeks 3 and 4 (Fisher exact $P<.03$) as well as across all weeks combined (Fisher exact $P=.03$). Nonparametric longitudinal analyses showed that the number of minutes of engagement with BrainHQ exercises did not significantly differ between the 2 groups (ATS=3.06; $P=.08$). However, these

XSL•FO
RenderX

analyses also showed that the number of minutes that aMCI participants completed increased over time compared to that of the psMCI group (ATS=6.72; *P*=.001), which decreased over the follow-up period. See Table 2 for the median number of minutes per week.

The variety of exercise types completed during follow-up was similar to the treatment phase. In the aMCI group, most exercises were in the brain speed (39.7%) and attention (36.9%) domains, followed by memory (8.5%), intelligence (5.8%), people skills (5.4%), and navigation (3.7%). A total of 28 different exercises were completed during the follow-up phase (Table 3). In the psMCI group, 38.2% of exercises during follow-up were in the domain of brain speed, followed by attention (24.5%), memory (19.7%), intelligence (13.7%), and people skills (3.9%). Fifteen different exercises were completed during the follow-up phase.

The aMCI group completed 2915 total exercises in the follow-up period. Of those, 62.5% were completed using the personal trainer option, and 48.5% were completed via the *a la carte* option in which participants chose the exercise, with 11 (55%) of 20 participants using the *à la carte* option for at least some follow-up-phase exercises. In comparison, of the 1224 exercises in the follow-up period that the psMCI group completed, only 1.7% of the exercises were completed via the *à la carte* option; only 1 (11%) of 9 psMCI participants who completed BrainHQ exercises during follow-up used the *à la carte option*.

In terms of device used, in the aMCI group, most exercises (60.5%) were completed on an iPad, followed by 22.6% completed via a web browser, 11.8% completed on an iPhone, and 5.1% completed on an Android device. Similarly, in the psMCI group, most exercises were completed using an iPad (41.3%), followed by 33.7% completed via a web browser, 24.2% via an Android device, and 0.8% via an iPhone.

BrainHQ Acceptability

Medians and interquartile ranges for the items from the study-specific experience questionnaires are reported in Table 4 for the aMCI group and Table 5 for the psMCI group. Of note, although not all participants engaged with BrainHQ during the follow-up phase, all participants provided ratings of their experience immediately posttreatment. Thus, in Table 4, we present ratings separately for all participants post-treatment, all participants post-study, and the subset of participants who engaged in any BrainHQ during follow-up.

XSL•FO
RenderX

Table . aMCI[a] group (n=21) BrainHQ experience questionnaire data.

Question	Post-TMS[b] (n=21), median (IQR)	Post-study (n=21), median (IQR)	Post-study BrainHQ Engagers (n=20), median (IQR)
Rating scale (1=made things a lot worse, 2=made things somewhat worse, 3=made no difference, 4=made things somewhat better, 5=made things a lot better)			
How much did this treatment help with your mood?	4 (3-5)	4 (3-5)	4 (3-4)
How much did this treatment help with your cognitive functioning?	4 (3-5)	4 (3-4)	4 (3-5)
Rating scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree)			
I became tired of doing BrainHQ exercises	2 (1-3)	2 (1-3)	2 (1-3)
I got frustrated using BrainHQ	2 (1-3)	3 (1-4)	3 (1-4)
I feel the BrainHQ exercises were appropriately challenging	5 (4-5)	5 (4-5)	5 (4-5)
I tried my best on all the BrainHQ exercises	5 (5-5)	5 (5-5)	5 (4.25-5)
I thought the BrainHQ exercise were fun	4 (4-5)	4 (3-4)	4 (3-4)
Rating scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree)			
The research coordinators followed up with me to make sure I was completing my BrainHQ exercises	5 (5-5)	5 (4-5)	5 (4-5)
I understood what to do for each BrainHQ exercise	4 (4-5)	4 (4-5)	4 (3-5)
The research coordinators adequately explained how to use BrainHQ	5 (5-5)	5 (4-5)	5 (4-5)
I feel the BrainHQ website was easy to navigate	4 (4-5)	4 (4-5)	4 (4-5)
I was able to access the BrainHQ exercises at home without any problems	__[c]	5 (4-5)	5 (4-5)
I was satisfied with the feedback I got from BrainHQ about my performance on the exercises	5 (4-5)	4 (4-5)	4 (4-5)

[a]aMCI: amnestic mild cognitive impairment.

[b]TMS: transcranial magnetic stimulation.

[c]Not applicable.

XSL•FO
RenderX

Table . psMCI[a] group (n=14) BrainHQ experience questionnaire data.

Question	Post-TMS[b] (n=14), median (IQR)	Post-study (n=14), median (IQR)	Post-study BrainHQ engagers (n=9), median (IQR)
Rating scale (1=not at all, 2, 3=somewhat, 4, 5=very much so)			
The BrainHQ exercises helped improve my day-to-day activity levels and overall functioning.	3 (2-4)	3 (3-5)	3 (3-5)
The BrainHQ exercises helped improve mental abilities (eg, concentration, remembering, thinking, or making decisions).	3.5 (3-5)	4 (3-5)[c]	4 (3-4.75)[d]
Rating scale (1=not at all, 2, 3=somewhat, 4, 5=very much so)			
The BrainHQ sessions were overly time consuming and interfered with my daily life/responsibilities.	1 (1-1)	1 (1-1)	1 (1-1.5)
The BrainHQ sessions were overall difficult and burdensome.	1 (1-2)	2 (1-2)	2 (1-3)
I became bored with the BrainHQ training exercises.	1 (1-1)	1 (1-2)	1 (1-2.5)
I was engaged and tried my best to do well on the BrainHQ training exercises.	5 (5-5)	5 (5-5)	5 (5-5)
I was motivated to engage with and complete the BrainHQ training exercises.	5 (5-5)	5 (5-5)[c]	5 (4-5)
I enjoyed the BrainHQ training exercises.	5 (4-5)	4.5 (4-5)	4 (4-5)
Rating scale (1=not at all, 2, 3=somewhat, 4, 5=very much so)			
I would participate in studies that involve completing tasks similar to the BrainHQ training exercises.	5 (4.25-5)	5 (4.25-5)[c]	5 (4.5-5)
I am interested to learn what the researchers discover about the effects of the BrainHQ exercises on problems similar to mine.	5 (5-5)	5 (5-5)[c]	5 (5-5)
I am interested to learn more about the BrainHQ training exercises.	5 (5-5)	5 (4-5)	5 (4.5-5)
I would be interested in continuing BrainHQ exercises.	5 (4-5)	5 (4-5)	5 (4.5-5)
It would be beneficial if the BrainHQ training exercises were offered to me by my regular treatment providers.	4 (3-5)	4 (3-5)	5 (3-5)
I would recommend the BrainHQ exercises to a friend who experiences similar problems.	5 (5-5)	5 (4.25-5)	5 (4.5-5)
The instructions for using the BrainHQ website were explained clearly and were easy to understand.	5 (4-5)	5 (5-5)[c]	5 (4.5-5)

[a]psMCI: post-stroke mild cognitive impairment.

[b]TMS: transcranial magnetic stimulation.

[c]n=13.

XSL•FO
RenderX

[d]n=8.

Participants in the aMCI group, on average, rated the BrainHQ exercises as making their cognitive functioning and mood "somewhat better" at both post-treatment and 4-week follow-up. The aMCI participants generally rated their experience with BrainHQ positively at both time points; they found the exercises fun, appropriately challenging, not tiresome or frustrating, and they indicated that they tried their best. Finally, aMCI participants rated their experience with the BrainHQ platform as positive, finding it easy to navigate and access and adequately explained, understanding what to do, and being generally satisfied with the feedback from the BrainHQ platform at both time points.

Participants in the psMCI group generally rated the BrainHQ exercises as "*somewhat*" to "*very much so*" helpful in improving mental abilities and everyday functioning both post-treatment and at 4-week follow-up. psMCI participants also rated their experience with BrainHQ positively in terms of their motivation, trying their best, and enjoyment, and overall did not find BrainHQ exercises time-consuming, difficult or burdensome, or boring at both time points. On average, psMCI participants rated the platform positively, finding that it was easy to understand and that they had interest in continuing, interest in learning more about it or participating in similar studies, and that they would recommend the platform. Ratings of BrainHQ were similar posttreatment and post 4-week follow-up, despite somewhat limited engagement with BrainHQ in the follow-up period.

Discussion

Principal Findings

These 2 pilot studies provide preliminary support for the feasibility and acceptability of interleaving CCT with accelerated iTBS for people with MCI, given that there was comparably high engagement in the BrainHQ exercises during the in-laboratory treatment phase across both groups. This was consistent with our hypothesis, given the structured nature of administration after each accelerated iTBS session. Although, on average, participants completed only approximately half of the possible 210 training minutes, this finding is somewhat expected. The 10‐ to 15-minute inter-session intervals used for delivering CCT also had to include time spent navigating the platform, transitioning in and out of BrainHQ from iTBS sessions, and taking breaks (eg, resting in between sessions and using the restroom). Future studies using intermittent CCT will need to account for these timing factors when planning training time.

The groups differentially engaged with BrainHQ during the 4-week unsupervised follow-up phase. The engagement rate increased in the aMCI group over the 4 weeks, with most participants independently engaging with BrainHQ in the follow-up phase. Engagement in the psMCI group was numerically lower, although not statistically significant, and decreased over time, with the majority of participants not engaging with BrainHQ in follow-up weeks 3 and 4. Generally, this indicates that ecological adherence to CCT following iTBS may be feasible for many persons with MCI, although some patients may require more support (eg, regular reminders, in clinic, with care partner) to reach CCT engagement goals following iTBS. Notably, while several in the aMCI group and some in the psMCI group met the prescribed number of minutes in the follow-up phase, no participant in either group completed daily BrainHQ practice (ie, all 28 d). This may suggest that a minute-based target may be more attainable than a day-based one or that daily training may not be feasible in these populations. For both groups, the most commonly completed exercises were in the domains of attention and brain speed in both the treatment and follow-up phases. During the follow-up phase, the psMCI group engaged in somewhat more memory and intelligence exercises than those with aMCI; however, differences in exercise choice are difficult to interpret, given the lower engagement in this group. Finally, despite this differential engagement, both groups rated the acceptability of the BrainHQ exercises positively.

One of the most salient findings was that during the self-administered follow-up phase, the aMCI group, on average, completed closer to the recommended BrainHQ engagement time of 140 minutes per week, whereas the psMCI group, on average, did not complete the prescribed number of minutes, and engagement decreased over time. Likely contributors to this observed difference in ecological adherence are differences in study design. Notably, the aMCI study provided weekly email reminders to complete the daily CCT, whereas the psMCI study did not. These findings suggest that in cognitively impaired samples, the use of regular reminders and other remote supports may be necessary to support completion of CCT outside of the laboratory setting. Another potential factor is that the aMCI group was offered the à la carte option during follow-up, whereas the psMCI group was instructed to continue using the personal trainer. Allowing aMCI participants to self-select exercises (via à la carte) may have invoked greater autonomy over their training experience, leading to increased adherence.

It is also possible that disease-specific factors could have contributed to discrepant engagement in CCT during the self-administered follow-up phase. The aMCI group may have had greater motivation, perceived benefit, or expectancy of benefit from CCT, given that these participants knew that, given the neurodegenerative nature of AD, they would likely progressively decline. In comparison, the psMCI group, while still experiencing objective impairments, would have already experienced significant spontaneous recovery from their initial deficits due to the stroke, with a prognosis of stability of functioning if other vascular events are avoided. The type of cognitive deficits experienced by the 2 groups may have also contributed. As expected, the aMCI group had primarily amnestic impairment along with set-shifting and semantic fluency difficulties. In comparison, the psMCI group performed significantly worse on a test that involves cognitive processing speed and visuo-motor organization (ie, coding), which are processes required for completion of all BrainHQ exercises. The psMCI group also had more variable deficits, which is expected given that cognitive sequelae poststroke are

XSL•FO
RenderX

heterogeneous. Additionally, while significant deficits that would affect the ability to complete cognitive tasks were exclusionary, mild visual and motor deficits are not uncommon post-stroke [35]. Most BrainHQ exercises, and particularly those in the domains of attention and processing speed, which were most often completed in this study, are visual in nature and require a fast response with the mouse or a finger. Therefore, the psMCI compared to the aMCI group may have found these exercises more challenging as a function of sensorimotor impairments that compounded the cognitive demands of the exercises, thus impacting their engagement.

This is the first study to our knowledge to elicit feedback on participants' experience with the BrainHQ platform interleaved with accelerated iTBS. Although the study-specific surveys differed, limiting direct comparisons, results from both surveys generally indicated that participants in both groups found it relevant and potentially beneficial to their functioning, enjoyable and not excessively burdensome, and feasible for use during both treatment and post-treatment periods. The psMCI group rated BrainHQ favorably despite limited engagement in the follow-up post-treatment phase, so this questionnaire did not aid in identifying why engagement was lower compared to that in the aMCI group. Future research using qualitative and mixed methods may elucidate the factors driving differential engagement, which could inform CCT delivery in these populations. Despite the small-scale nature of these pilot studies, our findings highlight the importance of individualizing interventions to the specific population and underscore the need to avoid treating all cases of MCI as homogeneous syndromes when planning treatment approaches.

Comparisons to Prior Work

Prior work investigating combined neuromodulation (eg, rTMS or transcranial direct current stimulation) and cognitive training has provided some additional insights into these combined interventions in the context of AD [36-38] and poststroke [39]. Although a growing literature supports the potential additive benefit of combining noninvasive brain stimulation (including rTMS) with cognitive training (including CCT; [40,41]), notably missing from the literature are studies refining methodologies for such combined interventions. Systematic reviews of the CCT literature demonstrate that CCT methodology is quite heterogeneous in terms of platform, training exercises, dose, and setting of administration [12,42]. While there is some evidence of the benefit of CCT to the same cognitive construct, there is limited support of benefits to different cognitive constructs and everyday activities [16]. However, many of these studies may not have been dosed for efficacy and used platforms with limited empirical support [21]. CCT was not explicitly dosed for efficacy in the current feasibility studies; however, a "gold standard" dose to elicit change has not been established [21]. Overall, the mixed literature and varied methodology suggest that there needs to be more research to establish which CCT methodology is most efficacious as well as how to deploy it in a way to elicit continued engagement (eg, increasing motivation or incentives and improved gamification). Moreover, more work is needed to identify the optimal method for combined delivery of CCT with other interventions, such as rTMS, particularly given that some methods (eg, required dose and most beneficial timing) may differ from CCT alone.

Limitations

The studies presented here have some limitations and reveal several opportunities for future work. First, given that these were pilot studies focused on the safety, feasibility, and acceptability of combined accelerated iTBS and CCT, the intervention components were not explicitly dosed for efficacy. However, as higher doses (ie, increased treatment demands or durations) may impact participants' engagement, continued consideration of feasibility and acceptability in future efficacy trials is necessary. Relatedly, the design of these studies did not include sham conditions, randomization, or blinding of participants or experimenters for either accelerated iTBS or CCT. Future randomized controlled trials that assess additive or synergistic effects of rTMS (accelerated iTBS or otherwise) and CCT are therefore needed. Second, both groups were predominantly non-Hispanic White with high levels of educational attainment. To ensure that research is representative of the general population, it is imperative that future research include participants from diverse sociodemographic backgrounds with varying levels of technology familiarity (which was not explicitly assessed in these studies). Third, people with aMCI and a history of stroke often experience neuropsychiatric symptoms, most commonly depression [43,44]; however, participants in our studies did not have clinically significant neuropsychiatric symptoms. Given that depressive symptoms are associated with poorer adherence to multidomain behavioral interventions (including cognitive training) in older adults [45], feasibility and acceptability of combined rTMS and CCT may be lower in people with MCI and comorbid neuropsychiatric symptoms. Alternatively, as rTMS to the l-dlPFC is US Food and Drug Administration–approved to treat depression, and both rTMS [46] and CCT [47] may improve depression or apathy in MCI, engagement in combined rTMS and CCT may instead be high in those with MCI and depression if they experience such benefits from the treatment. Therefore, our findings warrant replication in people with MCI and comorbid neuropsychiatric conditions.

Conclusions

In conclusion, this study provides initial evidence that accelerated iTBS to the l-dlPFC interleaved with CCT during treatment is feasible and acceptable to many patients with MCI. However, more limited engagement in the follow-up phase suggests that the feasibility of self-directed CCT may vary across people with MCI and by CCT delivery format (eg, personal trainer vs. à *la carte)*, and indicates that regular reminders may be beneficial. These findings suggest that personalized treatment approaches and methods to reinforce CCT may enhance feasibility and adherence. Our results can directly inform future trials aimed at optimizing treatments that combine neuromodulatory and CCT interventions. Subsequent randomized controlled trials that evaluate whether combined rTMS and CCT yields a synergistic or additive impact on MCI symptoms are warranted.

XSL•FO
RenderX

Acknowledgments

We thank Dr. E. Scott, Dr. T. Turner, Dr. M. Sugarman, Dr. M. Wagner, Dr. D. Szeles, Dr. F. Rodriguez-Porcel, Dr. N. Milano, and A. Swanson for their help with participant recruitment in support of this project. For their expertise and oversight, we are grateful to Dr. M. George, Dr. S. Kerns, Dr. P. Nietert, and Dr. F. Rodriguez-Porcel (for trial NCT04503096), and Dr. L. Bonilha, Dr. N. Rowland, Dr. R. Adams, Dr. X. Li, and Dr. M. Hoffman (for trial NCT04655963). Additionally, we gratefully acknowledge all our participants whose enthusiasm, trust, and dedication to research make this work possible. No data are available.

Funding

This work was supported by the National Institutes of Health National Center of Neuromodulation for Rehabilitation (NM4R; grant P2CHD086844 pilot project 29 to AB and pilot project 17 to LMM); the National Institute of General Medical Sciences of the National Institutes of Health (P30GM154630); the National Institute of Child Health and Human Development (P50HD118628); the COBRE Restoration of Neural-Based Function (P30GM154630), and the Neuro-PRECISE center (P50HD118628); the New Vision Research Foundation (grant NVR2020-001-3 to AB and LMM); the Alzheimer's Association (grant AARF-21-850073 to SA); the National Institute on Aging at the National Institutes of Health (grant R01AG081237 to AB); the National Institute of General Medical Sciences and the National Institutes of Health (grant P20GM109040 project 6134 to LMM); the National Institute of General Medical Sciences of the National Institutes of Health (P30GM154630); and the National Institute of Mental Health at the National Institutes of Health (grant T32MH019934 to LMC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data used in this study are not publicly available as they contain confidential clinical information, and participants were not asked to consent to data sharing.

Authors' Contributions

Conceptualization: LMC, SA, LMM, AB
Data curation: SA, OH, HHF, JL, KR
Formal analysis: LMC, SA
Funding acquisition: LMM, AB
Investigation: OH, HHF, JL, KR, MA
Methodology: LMM, AB
Project administration: SA, KM, LMM, AB
Resources: LMM, AB
Supervision: SA, GR, LMM, AB
Visualization: LMC, SA
Writing – original draft: LMC, SA
Writing – review & editing: OH, HHF, JL, KR, KM, MA, GR, LMM, AB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of the Template for Intervention Description and Replication (TIDieR) checklist.
[DOCX File, 16 KB - neuro_v5i1e81437_app1.docx]

References

1. Fitzsimmons SMDD, Oostra E, Postma TS, van der Werf YD, van den Heuvel OA. Repetitive transcranial magnetic stimulation-induced neuroplasticity and the treatment of psychiatric disorders: state of the evidence and future opportunities. Biol Psychiatry 2024 Mar 15;95(6):592-600. [doi: 10.1016/j.biopsych.2023.11.016] [Medline: 38040046]
2. Chou YH, Ton That V, Sundman M. A systematic review and meta-analysis of rTMS effects on cognitive enhancement in mild cognitive impairment and Alzheimer's disease. Neurobiol Aging 2020 Feb;86:1-10. [doi: 10.1016/j.neurobiolaging.2019.08.020] [Medline: 31783330]
3. Ding X, Zhang S, Huang W, et al. Comparative efficacy of non-invasive brain stimulation for post-stroke aphasia: a network meta-analysis and meta-regression of moderators. Neurosci Biobehav Rev 2022 Sep;140:104804. [doi: 10.1016/j.neubiorev.2022.104804] [Medline: 35926728]
4. Liu X, Li H, Yang S, et al. Efficacy of repetitive transcranial magnetic stimulation on post-stroke cognitive impairment: a systematic and a network meta-analysis. Int J Geriatr Psychiatry 2024 Jul;39(7):e6117. [doi: 10.1002/gps.6117] [Medline: 38925887]

XSL•FO
RenderX

5. Sharbafshaaer M, Gigi I, Lavorgna L, et al. Repetitive transcranial magnetic stimulation (rTMS) in mild cognitive impairment: effects on cognitive functions-a systematic review. J Clin Med 2023 Sep 25;12(19):6190. [doi: 10.3390/jcm12196190] [Medline: 37834834]
6. Chen X, Liu F, Lyu Z, Xiu H, Hou Y, Tu S. High-frequency repetitive transcranial magnetic stimulation (HF-rTMS) impacts activities of daily living of patients with post-stroke cognitive impairment: a systematic review and meta-analysis. Neurol Sci 2023 Aug;44(8):2699-2713. [doi: 10.1007/s10072-023-06779-9]
7. Li KP, Sun J, Wu CQ, et al. Effects of repetitive transcranial magnetic stimulation on post-stroke patients with cognitive impairment: a systematic review and meta-analysis. Behav Brain Res 2023 Feb;439:114229. [doi: 10.1016/j.bbr.2022.114229]
8. Farooq MU, Min J, Goshgarian C, Gorelick PB. Pharmacotherapy for vascular cognitive impairment. CNS Drugs 2017 Sep;31(9):759-776. [doi: 10.1007/s40263-017-0459-3] [Medline: 28786085]
9. Pleen J, Townley R. Alzheimer's disease clinical trial update 2019–2021. J Neurol 2022 Feb;269(2):1038-1051. [doi: 10.1007/s00415-021-10790-5]
10. Mijajlović MD, Pavlović A, Brainin M, et al. Post-stroke dementia - a comprehensive review. BMC Med 2017 Jan 18;15(1):11. [doi: 10.1186/s12916-017-0779-7] [Medline: 28095900]
11. Tik M, Hoffmann A, Sladky R, et al. Towards understanding rTMS mechanism of action: stimulation of the DLPFC causes network-specific increase in functional connectivity. Neuroimage 2017 Nov 15;162:289-296. [doi: 10.1016/j.neuroimage.2017.09.022] [Medline: 28912081]
12. Nguyen L, Murphy K, Andrews G. A game a day keeps cognitive decline away? A systematic review and meta-analysis of commercially-available brain training programs in healthy and cognitively impaired older adults. Neuropsychol Rev 2022 Sep;32(3):601-630. [doi: 10.1007/s11065-021-09515-2] [Medline: 34251578]
13. Irazoki E, Contreras-Somoza LM, Toribio-Guzmán JM, Jenaro-Río C, van der Roest H, Franco-Martín MA. Technologies for cognitive training and cognitive rehabilitation for people with mild cognitive impairment and dementia. A systematic review. Front Psychol 2020;11:648. [doi: 10.3389/fpsyg.2020.00648] [Medline: 32373018]
14. Gao M, Huang L, Yi J, et al. The effectiveness of computerized cognitive training in patients with poststroke cognitive impairment: systematic review and meta-analysis. J Med Internet Res 2025 Jun 12;27:e73140. [doi: 10.2196/73140] [Medline: 40503808]
15. Mueller KD. A review of computer-based cognitive training for individuals with mild cognitive impairment and Alzheimer's disease. Perspect ASHA SIGs 2016 Mar 31;1(2):47-61. [doi: 10.1044/persp1.SIG2.47]
16. Nguyen L, Murphy K, Andrews G. Cognitive and neural plasticity in old age: a systematic review of evidence from executive functions cognitive training. Ageing Res Rev 2019 Aug;53:100912. [doi: 10.1016/j.arr.2019.100912] [Medline: 31154013]
17. Elliott M, Parente F. Efficacy of memory rehabilitation therapy: a meta-analysis of TBI and stroke cognitive rehabilitation literature. Brain Inj 2014;28(12):1610-1616. [doi: 10.3109/02699052.2014.934921] [Medline: 25058353]
18. Boudreaux EO, Cherry KE, Elliott EM, Hicks JL. Effects of distraction and pictorial illustration on memory for countries in older adults with probable Alzheimer's disease. Exp Aging Res 2011 May;37(3):293-309. [doi: 10.1080/0361073X.2011.568816] [Medline: 21534030]
19. Mingming Y, Bolun Z, Zhijian L, Yingli W, Lanshu Z. Effectiveness of computer-based training on post-stroke cognitive rehabilitation: a systematic review and meta-analysis. Neuropsychol Rehabil 2022 Apr;32(3):481-497. [doi: 10.1080/09602011.2020.1831555] [Medline: 33092475]
20. Ge S, Zhu Z, Wu B, McConnell ES. Technology-based cognitive training and rehabilitation interventions for individuals with mild cognitive impairment: a systematic review. BMC Geriatr 2018 Sep 15;18(1):213. [doi: 10.1186/s12877-018-0893-1] [Medline: 30219036]
21. Shah TM, Weinborn M, Verdile G, Sohrabi HR, Martins RN. Enhancing cognitive functioning in healthly older adults: a systematic review of the clinical significance of commercially available computerized cognitive training in preventing cognitive decline. Neuropsychol Rev 2017 Mar;27(1):62-80. [doi: 10.1007/s11065-016-9338-9] [Medline: 28092015]
22. Pascual-Leone A, Freitas C, Oberman L, et al. Characterizing brain cortical plasticity and network dynamics across the age-span in health and disease with TMS-EEG and TMS-fMRI. Brain Topogr 2011 Oct;24(3-4):302-315. [doi: 10.1007/s10548-011-0196-8] [Medline: 21842407]
23. Yang T, Liu W, He J, et al. The cognitive effect of non-invasive brain stimulation combined with cognitive training in Alzheimer's disease and mild cognitive impairment: a systematic review and meta-analysis. Alzheimers Res Ther 2024 Jun 27;16(1):140. [doi: 10.1186/s13195-024-01505-9] [Medline: 38937842]
24. Caulfield KA, Fleischmann HH, George MS, McTeague LM. A transdiagnostic review of safety, efficacy, and parameter space in accelerated transcranial magnetic stimulation. J Psychiatr Res 2022 Aug;152:384-396. [doi: 10.1016/j.jpsychires.2022.06.038] [Medline: 35816982]
25. Rossi S, Antal A, Bestmann S, et al. Safety and recommendations for TMS use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert Guidelines. Clin Neurophysiol 2021 Jan;132(1):269-306. [doi: 10.1016/j.clinph.2020.10.003] [Medline: 33243615]
26. Aghamoosa S, Lopez J, Rbeiz K, et al. A phase I trial of accelerated intermittent theta burst rTMS for amnestic MCI. J Neurol Neurosurg Psychiatry 2024 Oct 16;95(11):1036-1045. [doi: 10.1136/jnnp-2023-332680] [Medline: 38719432]

XSL•FO
RenderX

27. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
28. Albert MS, DeKosky ST, Dickson D, 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. Alzheimers Dement 2011 May;7(3):270-279. [doi: 10.1016/j.jalz.2011.03.008] [Medline: 21514249]
29. Bondi MW, Edmonds EC, Jak AJ, et al. Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. J Alzheimers Dis 2014;42(1):275-289. [doi: 10.3233/JAD-140276] [Medline: 24844687]
30. Beam W, Borckardt JJ, Reeves ST, George MS. An efficient and accurate new method for locating the F3 position for prefrontal TMS applications. Brain Stimul 2009 Jan;2(1):50-54. [doi: 10.1016/j.brs.2008.09.006] [Medline: 20539835]
31. Hachinski V, Iadecola C, Petersen RC, et al. National Institute of Neurological Disorders and Stroke-Canadian Stroke Network vascular cognitive impairment harmonization standards. Stroke 2006 Sep;37(9):2220-2241. [doi: 10.1161/01.STR.0000237236.88823.47] [Medline: 16917086]
32. Shirk SD, Mitchell MB, Shaughnessy LW, et al. A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. Alzheimers Res Ther 2011 Nov 11;3(6):32. [doi: 10.1186/alzrt94] [Medline: 22078663]
33. Weintraub S, Besser L, Dodge HH, et al. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). Alzheimer Dis Assoc Disord 2018;32(1):10-17. [doi: 10.1097/WAD.0000000000000223] [Medline: 29240561]
34. Noguchi K, Gel YR, Brunner E, Konietschke F. NparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. J Stat Softw 2012;50:1-23. [doi: 10.18637/jss.v050.i12]
35. Shim H. Vascular cognitive impairment and post-stroke cognitive deficits. Curr Neurol Neurosci Rep 2014 Jan;14(1):418. [doi: 10.1007/s11910-013-0418-4] [Medline: 24272272]
36. Brem AK, Di Iorio R, Fried PJ, et al. Corticomotor plasticity predicts clinical efficacy of combined neuromodulation and cognitive training in Alzheimer's disease. Front Aging Neurosci 2020;12:200. [doi: 10.3389/fnagi.2020.00200] [Medline: 32733232]
37. Bagattini C, Zanni M, Barocco F, et al. Enhancing cognitive training effects in Alzheimer's disease: rTMS as an add-on treatment. Brain Stimul 2020;13(6):1655-1664. [doi: 10.1016/j.brs.2020.09.010] [Medline: 33002645]
38. Rabey JM, Dobronevsky E. Repetitive transcranial magnetic stimulation (rTMS) combined with cognitive training is a safe and effective modality for the treatment of Alzheimer's disease: clinical experience. J Neural Transm (Vienna) 2016 Dec;123(12):1449-1455. [doi: 10.1007/s00702-016-1606-6] [Medline: 27631152]
39. Chew E, Teo WP, Tang N, et al. Using transcranial direct current stimulation to augment the effect of motor imagery-assisted brain-computer interface training in chronic stroke patients-cortical reorganization considerations. Front Neurol 2020;11:948. [doi: 10.3389/fneur.2020.00948] [Medline: 32973672]
40. Poppe A, Ritter FDE, Bais L, et al. The efficacy of combining cognitive training and noninvasive brain stimulation: a transdiagnostic systematic review and meta-analysis. Psychol Bull 2024 Feb;150(2):192-213. [doi: 10.1037/bul0000406] [Medline: 37956054]
41. Gao Y, Qiu Y, Yang Q, et al. Repetitive transcranial magnetic stimulation combined with cognitive training for cognitive function and activities of daily living in patients with post-stroke cognitive impairment: a systematic review and meta-analysis. Ageing Res Rev 2023 Jun;87:101919. [doi: 10.1016/j.arr.2023.101919]
42. Hu M, Wu X, Shu X, et al. Effects of computerised cognitive training on cognitive impairment: a meta-analysis. J Neurol 2021 May;268(5):1680-1688. [doi: 10.1007/s00415-019-09522-7] [Medline: 31650255]
43. Köhler CA, Magalhaes TF, Oliveira JMMP, et al. Neuropsychiatric disturbances in mild cognitive impairment (MCI): a systematic review of population-based studies. Curr Alzheimer Res 2016;13(10):1066-1082. [doi: 10.2174/1567205013666160502123129] [Medline: 27137220]
44. Mitchell AJ, Sheth B, Gill J, et al. Prevalence and predictors of post-stroke mood disorders: a meta-analysis and meta-regression of depression, anxiety and adjustment disorder. Gen Hosp Psychiatry 2017 Jul;47:48-60. [doi: 10.1016/j.genhosppsych.2017.04.001] [Medline: 28807138]
45. Coley N, Ngandu T, Lehtisalo J, et al. Adherence to multidomain interventions for dementia prevention: data from the FINGER and MAPT trials. Alzheimers Dement 2019 Jun;15(6):729-741. [doi: 10.1016/j.jalz.2019.03.005] [Medline: 31047857]
46. Padala PR, Padala KP, Lensing SY, et al. Repetitive transcranial magnetic stimulation for apathy in mild cognitive impairment: a double-blind, randomized, sham-controlled, cross-over pilot study. Psychiatry Res 2018 Mar;261:312-318. [doi: 10.1016/j.psychres.2017.12.063] [Medline: 29331848]
47. Hill NTM, Mowszowski L, Naismith SL, Chadwick VL, Valenzuela M, Lampit A. Computerized cognitive training in older adults with mild cognitive impairment or dementia: a systematic review and meta-analysis. Am J Psychiatry 2017 Apr 1;174(4):329-340. [doi: 10.1176/appi.ajp.2016.16030360] [Medline: 27838936]

XSL•FO
RenderX

Abbreviations:

AD: Alzheimer's disease
aMCI: amnestic mild cognitive impairment
ATS: ANOVA-type statistic
CCT: computerized cognitive training
iTBS: intermittent theta burst
l-dlPFC: left dorsolateral prefrontal cortex
MCI: mild cognitive impairment
MUSC: Medical University of South Carolina
NIA-AA: National Institute on Aging and Alzheimer's Association
NIH: National Institutes of Health
NINDS: National Institute of Neurological Disorders and Stroke
NMDA: N-methyl-d-aspartate
PROMIS: Patient-Reported Outcomes Measurement Information System
psMCI: post-stroke mild cognitive impairment
rTMS: repetitive transcranial magnetic stimulation
TIDieR: Template for Intervention Description and Replication

Edited by A Zampogna; submitted 28.Jul.2025; peer-reviewed by C Pauletti, S Datla; revised version received 28.Apr.2026; accepted 15.May.2026; published 15.Jul.2026.

Please cite as:
Campbell LM, Aghamoosa S, Horn O, Fleischmann HH, Lopez J, Rbeiz KS, Madden K, Antonucci M, Revuelta G, McTeague LM, Benitez A
Feasibility of Interleaving Computerized Cognitive Training With Repetitive Transcranial Magnetic Stimulation: Pilot Studies in Mild Cognitive Impairment Due to Alzheimer's Disease and Stroke
JMIR Neurotech 2026;5:e81437
URL: https://neuro.jmir.org/2026/1/e81437
doi:10.2196/81437

© Laura M Campbell, Stephanie Aghamoosa, Olivia Horn, Holly H Fleischmann, James Lopez, Katrina S Rbeiz, Katrina Madden, Michael Antonucci, Gonzalo Revuelta, Lisa M McTeague, Andreana Benitez. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 15.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Diagnostic Accuracy of GPT-4 With Vision in Neuroradiology Board-Style Exam Questions: Cross-Sectional Case-Based Study

Tom T Sussan[1,2*], MS; Rebekah R Brawley[3*], BS; Joshua Eckroth[2*], PhD; James E Mossell[4*], MD; Tao Weitao[1*], MD, PhD

[1]Lake Erie College of Medicine, 5000 Lakewood Ranch Blvd., Bradenton, FL, United States

[2]Department of Computer Science, Stetson University, Deland, FL, United States

[3]College of Medicine, University of Central Florida, Orlando, FL, United States

[4]Department of Radiology, University of Florida College of Medicine, Gainesville, FL, United States

[*]all authors contributed equally

Corresponding Author:
Tao Weitao, MD, PhD
Lake Erie College of Medicine, 5000 Lakewood Ranch Blvd., Bradenton, FL, United States

Abstract

Background: Multimodal artificial intelligence systems combining text and image analysis represent a paradigm shift in clinical decision support. While GPT-4 with Vision (GPT-4V) has shown promise in medical imaging interpretation, existing studies report inconsistent performance (16% - 80% accuracy) across radiological subspecialties. Critical knowledge gaps persist regarding GPT-4V's capability to integrate clinical history with imaging findings in complex neuroradiology scenarios, and fundamental questions remain about whether the model appropriately balances visual and textual information sources when formulating diagnoses. Furthermore, documented artificial intelligence hallucination rates of 35.5% to 63% in radiology applications raise urgent safety concerns, yet the relationship between modality utilization patterns and diagnostic accuracy remains unexplored.

Objective: This study aims to evaluate GPT-4V's diagnostic accuracy on expert-validated neuroradiology board-style examination questions and to examine the model's self-reported reliance on imaging versus clinical text data when making diagnostic decisions. A secondary objective was to examine whether self-characterized modality utilization patterns differed systematically between correct and incorrect diagnoses, potentially identifying specific failure modes requiring targeted mitigation strategies.

Methods: This cross-sectional study evaluated GPT-4V using 29 neuroradiology cases from the RSNA (Radiological Society of North America) Case Collection, covering adult brain and central nervous system pathologies imaged via computed tomography or magnetic resonance imaging. The cases were authored by board-certified radiologists. GPT-4V was accessed via ChatGPT Plus (July 2024) with standardized prompts selecting 1 answer from 4 options, providing diagnostic rationale, and quantifying the percentage contributions of image versus text data. Binary scoring assessed diagnostic performance (correct=1, incorrect=0). Statistical analysis included Wilson score CIs, a binomial test comparing accuracy to chance, and a 2-tailed *t* test comparing self-reported modality reliance between correct and incorrect diagnoses (α=.05, Cohen *d* calculated).

Results: GPT-4V correctly diagnosed 22 of 29 cases (76% accuracy, 95% CI 57.9%-87.8%), significantly exceeding the chance performance of 25% (z=6.33; P<.001). The model self-reported mean contributions of 66.1% from imaging (95% CI 63.5% - 68.8%) and 33.9% from text (95% CI 31.2% - 36.5%). Correct diagnoses (n=22) showed significantly lower self-reported image reliance (62.8%, 95% CI 61.3% - 64.3%) compared to incorrect diagnoses (n=7; 76.7%, 95% CI 73.5% - 80.0%), with a mean difference of 13.9 percentage points (95% CI 10.6 - 17.3; P<.001; Cohen d=4.08, 95% CI 2.73 - 5.43). All 7 incorrect diagnoses demonstrated image-dominant attribution ≥70% (Fisher exact test P<.001), suggesting that excessive visual reliance may indicate diagnostic risk.

Conclusions: The 76% accuracy substantially exceeds prior GPT-4V radiology studies (43%), demonstrating that focused domain application with structured prompting enhances performance. Incorrect diagnoses are associated with higher self-reported visual reliance, suggesting a potential failure mode warranting experimental validation. This pattern identifies a potentially actionable signal for quality assurance systems. Clinical deployment should remain restricted to supervised educational applications with mandatory radiologist oversight until balanced context-aware integration is validated.

(JMIR Neurotech 2026;5:e69708) doi:10.2196/69708

KEYWORDS

neuroradiology; GPT-4 with Vision; GPT-4V; artificial intelligence; diagnostic accuracy; multimodal AI; medical imaging; clinical decision-making; text-image integration; board exam questions; RSNA Case Collection

XSL•FO
RenderX

Introduction

Background

The advent of multimodal artificial intelligence (AI) systems represents a transformative shift in medical diagnostics, particularly in radiology, where clinical decision-making requires integrated analysis of imaging findings and clinical context. Multimodal AI models combine diverse data modalities, such as imaging, text, structured records, and physiological signals, into unified analytical frameworks [1-3]. Recent advancements in transformer architectures and foundation models have enabled unprecedented capabilities in processing heterogeneous medical data [4-6], with AI adoption in radiology accelerating rapidly in recent years [7].

OpenAI's GPT-4 with Vision (GPT-4V), released in 2023, exemplifies this multimodal paradigm by enabling the simultaneous interpretation of text and images. Large language models (LLMs) have demonstrated utility in radiology report generation, board exam preparation, and clinical decision support [8-11], with studies reporting significant improvements in efficiency and consistency [7]. However, the addition of visual integration has yielded contradictory performance patterns across radiological subspecialties, raising fundamental questions about how these models process and integrate information from different modalities.

Current Evidence and Critical Knowledge Gaps

Empirical evaluations reveal substantial heterogeneity in GPT-4V diagnostic accuracy. Huppertz et al [12] demonstrated that diagnostic accuracy improved from 8.3% with images alone to 29.1% with contextualized prompts, though the model exhibited pronounced context bias and frequent fabricated findings, with similar concerns documented in other multimodal evaluations [13]. Studies report GPT-4V accuracy ranging from 16% to 49% in challenging radiology cases (characterized by rare pathologies, subtle findings, or complex differentials), consistently below trained radiologists' performance [14-16]. Albaqshi et al [17] found that among 6 LLMs evaluated on 56 neuroradiology cases, Claude 3.5 achieved the highest accuracy (80.4%), with LLMs performing comparably to first-year fellows while showing high consistency across repeated queries. Systematic reviews confirm variable results (16% - 80% accuracy) depending on case difficulty, prompt engineering, and domain specificity [18,19].

Fundamental questions persist about how GPT-4V integrates visual and textual information. Multiple studies document limited visual interpretation capabilities: Schramm et al [20] identified textual descriptions as the strongest contributor to performance, while Albaqshi et al [17] demonstrated that image-only accuracy plummeted to 21.5% to 63.1% compared to 62.5% to 76.8% with combined inputs. Conversely, some studies show GPT-4V's superiority over text-only approaches [21], while others report text-only models outperforming multimodal implementations [16,22,23]. These findings suggest a critical paradox: adding visual capabilities may not enhance but can potentially degrade diagnostic performance when multimodal integration is suboptimal.

A critical barrier to clinical deployment is AI hallucinations, the plausible but incorrect information that appears factually grounded [7,24-26]. LLM hallucinations in medical contexts remain a critical concern [7], manifesting as fabricated findings or misidentified modalities [12,26,27]. Jin et al [24] documented "hidden flaws behind expert-level accuracy," revealing systematic errors obscured by superficially correct outputs. Current literature lacks systematic investigation of whether diagnostic failures correlate with specific modality utilization patterns. Understanding these patterns is essential for safe deployment, as systematic overreliance on either modality could lead to predictable failure modes requiring targeted mitigation.

Current multimodal foundation models exhibit limitations precluding autonomous diagnostic use: inconsistent results across identical inputs, tendency toward confabulation [26,27], sensitivity to prompt engineering [28], lack of transparency [12,25], and variable performance across modalities and anatomical regions [12,29]. Recent position statements emphasize that AI integration must prioritize human-AI collaboration frameworks, transparent uncertainty quantification, and mandatory expert oversight [7,30]. Despite growing literature on diagnostic accuracy, critical gaps remain regarding (1) how multimodal AI characterizes its reliance on visual versus textual inputs, (2) whether modality attribution patterns differ between correct and incorrect diagnoses, and (3) whether self-reported information utilization reflects actual processing versus post hoc rationalization [17].

Study Objectives

This study addressed two objectives: (1) to evaluate GPT-4V's diagnostic accuracy on expert-validated neuroradiology board-style questions, providing benchmark performance data under standardized conditions and (2) as an exploratory analysis to document GPT-4V's self-reported reliance on imaging versus clinical text and examine whether self-characterized modality utilization patterns differ between correct and incorrect diagnoses. We acknowledge that determining whether these self-assessments reflect actual information processing versus post hoc rationalization requires rigorous experimental validation through controlled text-only and image-only conditions.

Methods

Study Design and Data Source

This cross-sectional study, reported according to JARS-Quant guidelines [31], evaluated GPT-4V's diagnostic accuracy using 29 neuroradiology cases from the RSNA (Radiological Society of North America) Case Collection. The cases included adult brain and central nervous system pathologies imaged via computed tomography (CT) or magnetic resonance imaging (MRI; Table S1 in Multimedia Appendix 1, Figure S5.1 and Table S2.1 in Multimedia Appendix 2, and Table S6.1 in Multimedia Appendix 3). Each case included a clinical vignette and diagnostic-quality imaging studies (Figures S2 in Multimedia Appendix 4 and Figure S3 in Multimedia Appendix 5, respectively). The inclusion criteria required expert-verified diagnoses in multiple-choice format; the cases were authored by board-certified radiologists and underwent editorial review

XSL•FO
RenderX

following established quality standards for educational radiology assessments [32], analogous to standardized board examination validation. Cases were excluded if they involved pediatric patients, lacked diagnostic images, or had no definitive correct answer.

While the RSNA Case Collection's restricted membership access reduces the likelihood of training data contamination, we acknowledge that with closed-source models, data leakage cannot be definitively ruled out and could artificially inflate performance estimates. All case materials were deidentified and used with permission. The cases were accessed in July 2024.

Assessment and Scoring Methodology

Complete prompt structure, standardized instructions, and example responses are documented in Multimedia Appendix 4 (Parts A-E), ensuring reproducibility. Binary scoring evaluated diagnostic performance: correct (score=1) if the model's answer matched the peer-reviewed correct diagnosis from RSNA documentation; incorrect (score=0) otherwise (Table S6.1 in Multimedia Appendix 3 and Table S7.1 and Section 8.1 in Multimedia Appendix 6). All cases underwent peer review and editorial vetting by the RSNA's editorial board prior to publication (Multimedia Appendix 5). No partial credit was given. Overall accuracy was calculated as percentage correct out of 29 cases. As exploratory measures, we recorded self-assessed percentage influence of image versus text for each case, emphasizing that these represent subjective self-reports rather than validated measurements of actual information contribution (Table S7.2 and Section 8.1 in Multimedia Appendix 6).

Ethical Considerations

This study did not constitute human subjects research as defined by US Department of Health and Human Services regulations at 45 CFR 46.102(e) and (l) [33]. The study involved secondary analysis of fully deidentified educational case materials from the RSNA Case Collection, accessed through authorized membership. No living individuals were contacted, and no identifiable private information was obtained, used, or generated. The RSNA Case Collection requires authors to remove all patient identifiers prior to submission (Multimedia Appendix 5). All figures and multimedia appendices in this study contain only fully deidentified radiological images and clinical information from the RSNA Case Collection, with no possibility of individual identification.

Missing Data Analysis

All 29 cases included in the analysis had complete data for the primary outcome (diagnostic accuracy; Table S5.1 in Multimedia Appendix 7). Each case successfully elicited a diagnostic response from GPT-4V, with the model selecting 1 of 4 answer options in all instances. For the exploratory modality attribution measures, the model provided self-reported percentage contributions (image vs text) for all 29 cases, resulting in zero missing data for both primary and exploratory outcome measures (Figure S5.1 in Multimedia Appendix 7 and Part C in Multimedia Appendix 4). Therefore, no imputation procedures or missing data analyses were necessary. The completeness of data reflects the controlled nature of the study design, where GPT-4V was systematically prompted to provide both diagnostic answers and modality attribution percentages for each case. One trial per case was conducted with standardized prompts designed to elicit complete responses. The single-trial design means that response variability across multiple trials was not assessed. GPT-4V's temperature setting and stochastic sampling could produce different responses on repeated trials; this variability is addressed in the *Limitations* section. The study protocol specified that any case with incomplete model responses would be excluded and reported as a protocol deviation. This scenario did not occur.

Statistical Considerations

The sample size (N=29) was determined by available cases meeting the inclusion criteria (Tables S3.1 and S3.2 in Multimedia Appendix 8 document post hoc power >99.9% for primary analysis, 97% for exploratory analysis, and ±16.3 percentage point margin of error). Statistical significance was assessed using a 2-sided alpha level of .05 for all hypothesis tests. Statistical analysis was primarily descriptive (Tables S6.1-S6.2 in Multimedia Appendix 3). Statistical analysis used a 2-sample 2-tailed *t* test with standard error–based CIs for continuous variables; Wilson score method was applied separately for the diagnostic accuracy proportion (Table S4.2 in Multimedia Appendix 9 presents complete 2-tailed *t* test results: t_{27}=9.40; P<.001, Cohen d=4.08; Tables S5.3-S5.5 in Multimedia Appendix 7 verify assumptions, including normality and equal variances, and provide comprehensive descriptive statistics). We compared the findings qualitatively to previous studies [20-23,30].

A 1-sample binomial test assessed whether diagnostic accuracy exceeded random guessing (25% for 4-option questions, z=6.33, P<.001; Table S4.1 in Multimedia Appendix 9). CIs for proportions were calculated using the Wilson score method. For modality weighting, 95% CIs were calculated using the *t* distribution. A 2-sample 2-tailed *t* test compared self-reported image reliance between correct and incorrect cases. Effect sizes (Cohen *d*) with 95% CIs are reported to allow readers to interpret clinical and statistical significance (Table S4.3 in Multimedia Appendix 9 shows all incorrect diagnoses demonstrated image-dominant attribution ≥70%, Fisher exact test P<.001; Table S5.2 in Multimedia Appendix 7 confirms no statistical outliers; Tables S7.1 and S7.2 and Section S8.2 in Multimedia Appendix 6 document variable definitions and derived measures). Complete statistical methods are detailed in Multimedia Appendix 9.

Figure 1 is the systematic methodology for evaluating GPT-4V diagnostic performance on 29 neuroradiology cases from the RSNA Case Collection. The workflow includes (1) data sources—cases containing CT or MRI scans of adult brain and central nervous system pathologies, clinical vignettes, and peer-reviewed multiple-choice questions; (2) standardized prompt structure—a consistent template instructing GPT-4V to review all radiographic imaging, select 1 diagnostic answer from 4 options, provide diagnostic rationale, and quantify the percentage contribution of visual versus textual information to its diagnostic decision (exploratory outcome measure); and (3) ChatGPT Plus implementation—prompt delivery via ChatGPT

XSL•FO
RenderX

Plus web interface. This structured methodology ensures standardized evaluation while systematically capturing the model's self-reported reliance on visual versus textual information sources. One trial per case was conducted without iterative prompting to simulate real-world clinical conditions where single diagnostic assessments are typical.

Figure 1. Prompt engineering workflow for GPT-4 with Vision (GPT-4V) evaluation in neuroradiology board-style questions: a cross-sectional study of 29 adult brain and central nervous system pathologies from the RSNA Case Collection (July 2024). API: application programming interface; CT: computed tomography; MRI: magnetic resonance imaging.

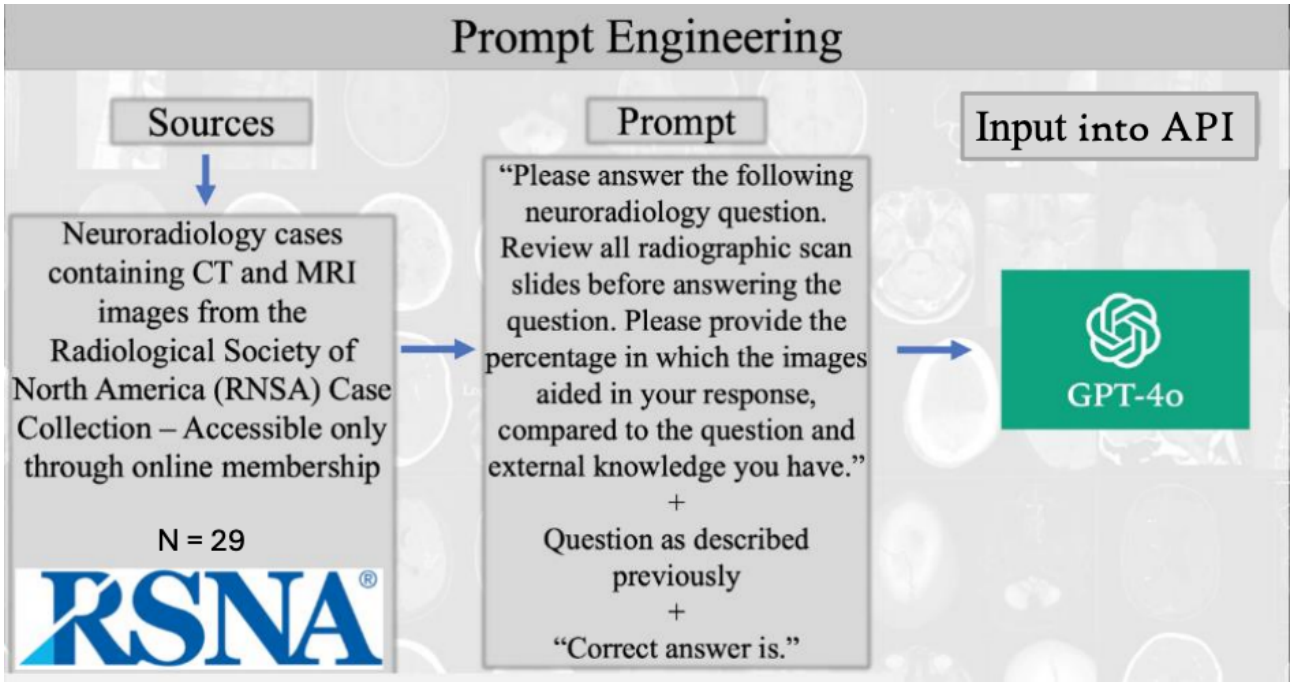


Figure 2 is a representative infectious disease case (Table S6.1 in Multimedia Appendix 3, Case #14: Neurocysticercosis) from the RSNA Case Collection, 1 of 29 adult brain and central nervous system pathology cases (Figure S5.1 in Multimedia Appendix 2, Table S6.1 in Multimedia Appendix 3) used in this July 2024 cross-sectional evaluation (Figure S5.1 in Multimedia Appendix 2) of GPT-4V diagnostic accuracy. This case features a 32-year-old male presenting with first-time seizure (complete case vignette in Multimedia Appendix 5) and includes (1) a clinical vignette containing patient demographics, symptoms, and history (textual information) and (2) diagnostic-quality neuroimaging studies in PNG format (image format not documented in appendices) obtained via CT or MRI (visual information; Multimedia Appendix 5 and Figure S5.1 in Multimedia Appendix 2). GPT-4V was required to integrate both clinical context and imaging findings (Part A: standardized prompt structure in Multimedia Appendix 4) to select the correct diagnosis from 4 multiple-choice options (Part A in Multimedia Appendix 4 and Answer Choices A-D in Multimedia Appendix 5), with self-reported percentage contributions from each modality recorded as an exploratory measure (Parts A-B in Multimedia Appendix 4; Table S6.1 in Multimedia Appendix 3 confirms 65% image, 35% text attribution for this case).

XSL•FO
RenderX

Figure 2. Representative infectious disease neuroradiology case from the RSNA Case Collection demonstrating clinical context integration in GPT-4 with Vision (GPT-4V) multimodal diagnostic decision-making.

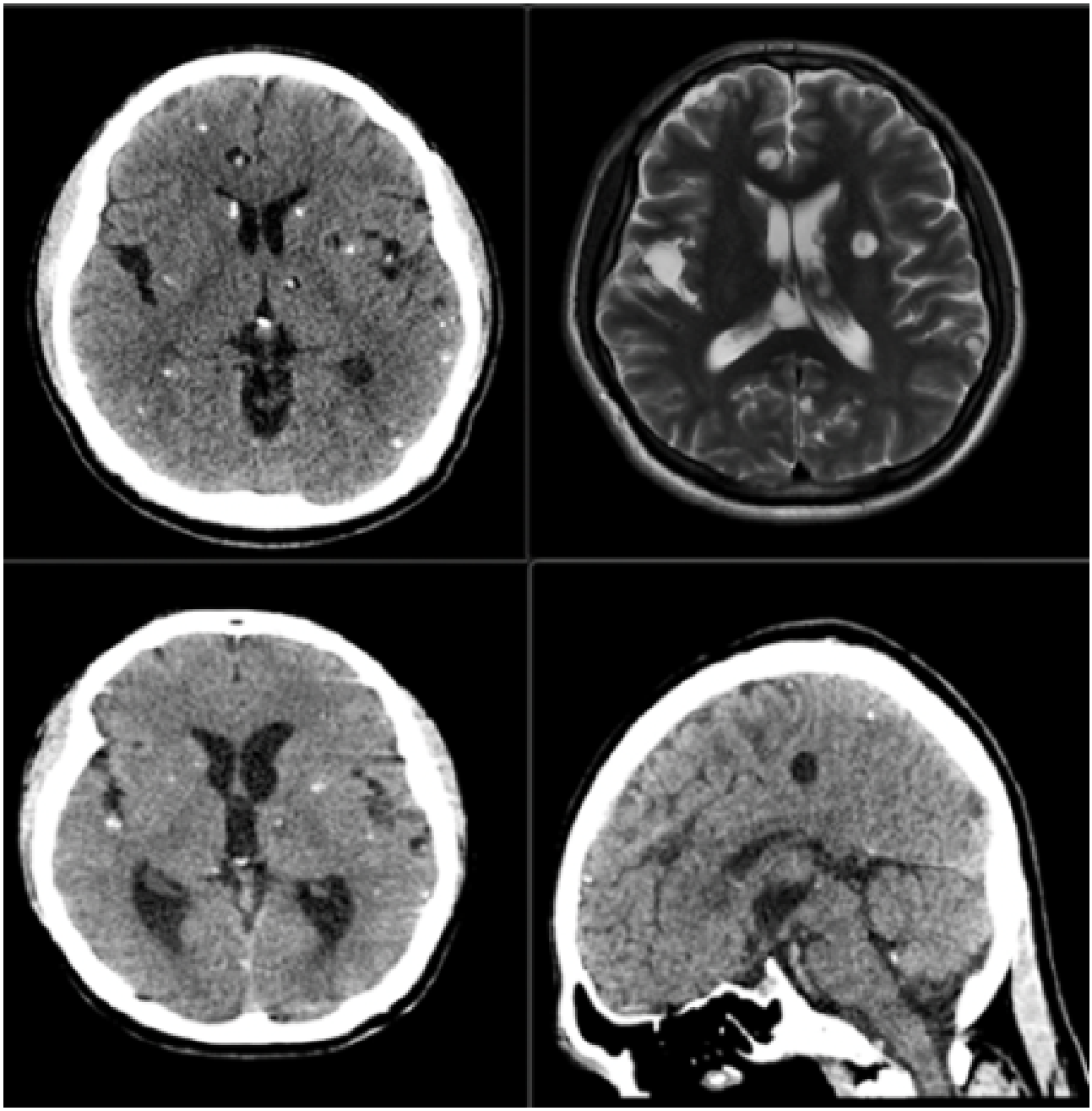


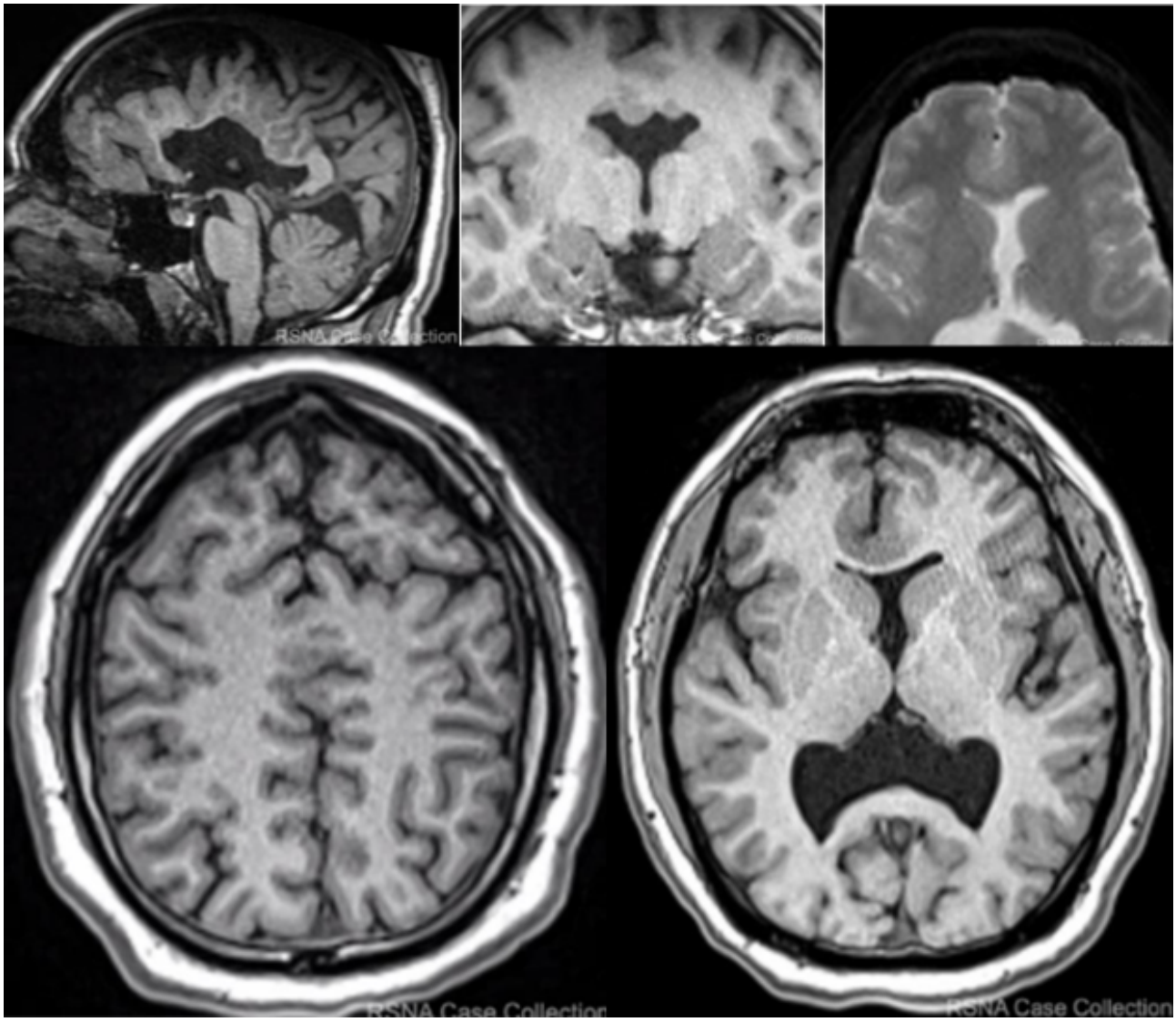
Figure 3 is a complex neuroradiology case (Table S6.1 in Multimedia Appendix 3, Case #19: Syntelencephaly) from this July 2024 cross-sectional study (Figure S5.1 in Multimedia Appendix 2) evaluating GPT-4V's diagnostic performance on 29 adult brain and central nervous system pathology cases from the RSNA Case Collection. This case involves a 31-year-old male patient with developmental brain abnormalities and seizures (Multimedia Appendix 5: complete case presentation with clinical vignette, developmental history). The case structure provided to GPT-4V included (1) clinical vignette describing patient demographics, symptoms (seizures), developmental history, and relevant neurological findings (textual data) and (2) complete series of diagnostic-quality neuroimaging studies obtained via CT or MRI in standardized PNG format (image format not documented); visual data (Part A in Multimedia Appendix 4 and Multimedia Appendix 5). This case exemplifies challenging diagnostic scenarios where developmental malformations present subtle imaging findings requiring expert-level integration of both clinical context and radiological interpretation (Table S6.1 in Multimedia Appendix 3 confirms correct diagnosis; Table S2.1 in Multimedia Appendix 2 shows Developmental category: 100% accuracy).

The mean self-reported modality contributions across diagnostic outcomes for 29 neuroradiology cases from the RSNA Case Collection are presented. The data are shown for all cases (N=29), correct diagnoses (n=22), and incorrect diagnoses (n=7). Image contributions are normalized to 1.0 (blue bars) to enable comparison across categories; text contributions appear as ratios (purple bars). Error bars represent 95% CIs for text:image ratios. Incorrect diagnoses showed significantly higher self-reported image reliance (76.7%) compared to correct diagnoses (62.8%),

XSL•FO
RenderX

with a mean difference of 13.9 percentage points (P<.001; Cohen d=4.08; Tables S4.2-S4.3 in Multimedia Appendix 9).

Figure 3. Representative developmental brain malformation case highlighting complex multimodal integration requirements for accurate artificial intelligence (AI) diagnosis.



Results

Overview of Primary and Exploratory Findings

Our primary finding is that GPT-4V achieved 76% diagnostic accuracy (22 out of 29 correct diagnoses) on expert-validated neuroradiology cases, significantly exceeding chance performance. Secondary exploratory findings regarding self-reported modality utilization should be interpreted cautiously, as they represent the model's characterization of its process rather than validated measurements of actual information use.

Diagnostic Performance

GPT-4V correctly diagnosed 22 out of 29 neuroradiology cases, yielding 76% accuracy (95% CI 57.9% - 87.8% by Wilson method), significantly above the 25% expected by chance (z=6.33; P<.001; Table S1 in Multimedia Appendix 1, Table S4.1 in Multimedia Appendix 9, and Table S6.1 in Multimedia Appendix 3). This exceeds the 43% accuracy (31/72 cases) reported by Mukherjee et al [34] for GPT-4V on RSNA "Case of the Day" challenges, suggesting that within focused domains under structured prompting, GPT-4V's performance can be enhanced. The multiple-choice format (Part A in Multimedia Appendix 4 and Multimedia Appendix 5) may have aided performance by providing plausible options rather than requiring open-ended diagnosis generation. However, we acknowledge that data leakage cannot be definitively ruled out with closed-source models, and any training data contamination could have contributed to this performance.

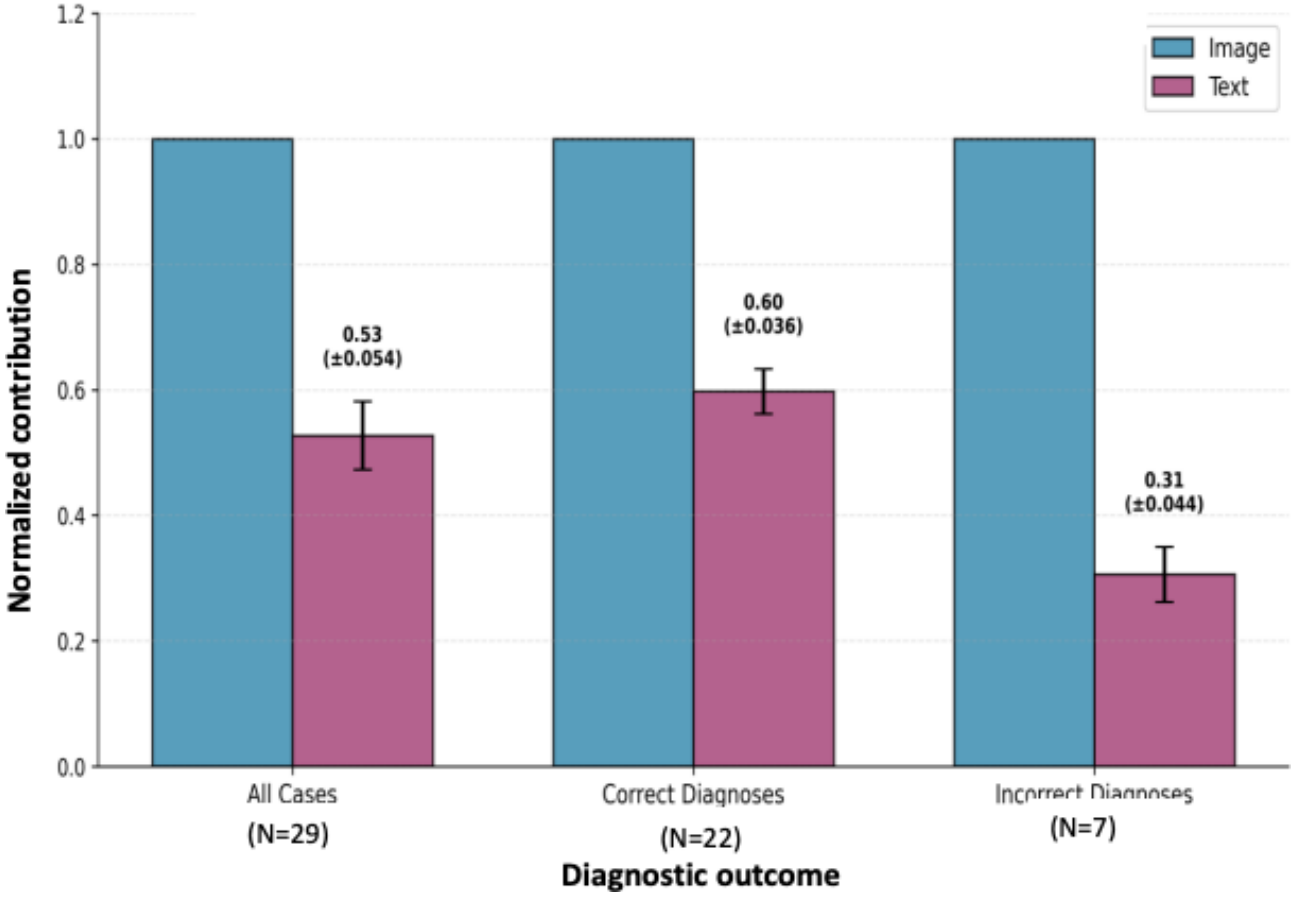
Multimodal Data Integration Patterns

As an exploratory analysis, we examined GPT-4V's self-reported percentage contributions of visual versus textual information to diagnostic decisions. The model reported that image data contributed 66% (95% CI 63.5% - 68.8%) and textual data 34% (95% CI 31.2% - 36.5%) on average (Table S1 in Multimedia Appendix 1 and Table S5.5 in Multimedia Appendix 7). The overall text:image ratio was 0.53; however, this ratio varied by diagnostic outcome: correct diagnoses

XSL•FO
RenderX

showed 0.60 versus incorrect diagnoses at 0.31 (Figure 4; Table S4.2 in Multimedia Appendix 9), suggesting that errors may be associated with self-characterized overreliance on imaging (Table S4.3 in Multimedia Appendix 9 shows that 100% of incorrect diagnoses had image-dominant attribution ≥70%, Fisher exact $P<.001$; Table S6.1 in Multimedia Appendix 3 provides case-level data). Whether this reflects actual information processing or post hoc rationalization requires rigorous experimental validation through controlled text-only and image-only conditions (Section 8.1 in Multimedia Appendix 6 documents interpretation limitations and validation requirements).

Figure 4. Self-reported image and text contributions by diagnostic outcome in GPT-4V neuroradiology evaluation.



Comparative Analysis

Our 76% accuracy (95% CI 57.9%-87.8%; Table S4.1 in Multimedia Appendix 9 and Table S1 in Multimedia Appendix 1) substantially exceeds the 43% from Mukherjee et al [34], who found that GPT-4V performed worse on imaging-dependent cases (39%) than on text-only inputs (50%), suggesting heavy reliance on textual information. Our focused neuroradiology approach (Table S6.1 in Multimedia Appendix 3 and Table S2.1 in Multimedia Appendix 2) with multiple-choice format (Part A in Multimedia Appendix 4 and Multimedia Appendix 5) was associated with more robust performance, though differences in case difficulty between the RSNA Case Collection and Annual Meeting case sets may also contribute. Domain-specific factors and question format appear to influence success.

Our findings contrast with those of Hirosawa et al [22], who reported that GPT-4V attributed only 30% of decisions to image data on general case reports, with text-only GPT-4 outperforming GPT-4V (55.9% vs 44.4%). Our substantially higher image utilization (66%; 95% CI 63.5%-68.8%; Table S1 in Multimedia Appendix 1 and Table S5.5 in Multimedia Appendix 7) may reflect our explicit prompting to consider and quantify image information (Part A in Multimedia Appendix 4 documents modality quantification instructions) or the imaging-centric nature of neuroradiology cases compared to general medical case reports. These results suggest that GPT-4V makes extensive use of visual data, but improved outcomes depend on accurate image interpretation within the clinical context.

Discussion

Summary of Main Findings

This study addressed 2 primary objectives: evaluating GPT-4V's diagnostic accuracy on expert-validated neuroradiology board-style questions and exploring self-reported reliance on imaging versus clinical text during diagnostic decision-making. Both objectives were successfully addressed. GPT-4V achieved 76% diagnostic accuracy on expert-validated neuroradiology cases, substantially exceeding prior performance on RSNA materials (76% vs 43% from Mukherjee et al [34] on case of the day challenges), though differences in case selection and difficulty limit direct comparison. Exploratory analysis revealed an inverse relationship: incorrect diagnoses were associated with higher self-reported visual reliance compared to correct

XSL•FO
RenderX

diagnoses. While these represent the model's self-characterization rather than validated measurements, this pattern generates testable hypotheses about multimodal integration failure modes. This is among the first studies systematically documenting how multimodal AI self-reports reliance on different information sources during clinical diagnosis.

Interpretation and Comparison to Literature

Our findings contribute to emerging evidence regarding multimodal AI diagnostic accuracy and information integration patterns. Superior performance compared to prior studies [34] demonstrates that diagnostic accuracy depends critically on domain specificity, question format, and prompt engineering, all of which suggest that performance cannot be characterized by single global metrics but varies substantially based on the implementation approach.

The exploratory finding regarding modality attribution, where diagnostic errors were associated with higher self-reported image reliance, aligns with multiple studies. Schramm et al [20] identified textual descriptions as the strongest contributor to performance, while Hayden et al [16] found that GPT-4V performed significantly worse on image-based questions (47.8%) compared to text-only questions (81.5%). Albaqshi et al [17] demonstrated that image-only accuracy plummeted compared to text-with-image performance. Our pattern may reflect a failure mode in which the model attempts to extract diagnostic information primarily from visual data despite limited visual interpretation capabilities, thereby neglecting critical clinical context that might correct misinterpretations.

This image-dominant failure pattern warrants deeper mechanistic consideration. Incorrect diagnoses showed substantially higher self-reported image reliance compared to correct diagnoses, with a very large effect size, suggesting this is not merely statistical noise but a consistent pattern in self-reported attribution that determines whether this reflects actual information processing or post hoc rationalization. The fact that all incorrect diagnoses demonstrated image-dominant attribution patterns is particularly striking and suggests a potential "tipping point" beyond which diagnostic accuracy deteriorates markedly.

Several mechanisms could explain this pattern, though experimental validation is required to test these hypotheses. First, if visual processing capabilities lack domain-specific fine-grained discrimination for subtle radiological findings, excessive reliance on visual inputs might lead to confident but incorrect diagnoses. Second, the insufficient integration of clinical context could contribute to diagnostic errors. Third, architectural factors in multimodal integration remain unexplored and warrant investigation through controlled experiments with systematic input manipulation. These proposed mechanisms are speculative and require rigorous testing through image-only, text-only, and combined conditions to validate whether self-reported attribution patterns reflect actual information processing.

Comparison with human diagnostic patterns is instructive. Experienced radiologists typically use iterative hypothesis refinement, beginning with clinical context to generate differential diagnoses, and then using imaging to confirm or refute specific possibilities [30]. This approach naturally balances modalities by forcing explicit integration. In contrast, GPT-4V may process visual and textual streams more independently, with final outputs reflecting whichever stream activates more strongly rather than true synthesis. Potential architectural or integration limitation could explain why adding visual capabilities sometimes degrades rather than enhances performance [16,22,23], though our observational data cannot establish this mechanism.

Our finding of higher overall image utilization compared to some studies [22] may reflect explicit prompting to quantify image contribution or the imaging-centric nature of neuroradiology cases. However, the consistency of the image-dominant failure pattern across diverse pathology categories within neuroradiology (Table S6.2 in Multimedia Appendix 3) suggests this is not merely an artifact of case selection within this domain. Whether this pattern generalizes to other radiological subspecialties or medical domains requires investigation in diverse clinical contexts. Our observed overall image utilization, while seemingly reasonable, may actually be excessive given that textual clinical information often carries disproportionate diagnostic weight relative to its volume.

That Busch et al [21] demonstrated GPT-4V's superiority over text-only approaches in some tasks suggests the relationship between modality contribution and diagnostic success is task-dependent and complex. This task dependency likely reflects varying degrees of diagnostic specificity achievable through visual inspection alone versus requiring clinical correlation. For conditions with pathognomonic imaging features (eg, calcified subependymal nodules in tuberous sclerosis), visual dominance may succeed. For conditions requiring clinical-radiological synthesis (eg, distinguishing demyelination patterns based on temporal profile), balanced integration becomes essential. Our results suggest that GPT-4V may not appropriately adjust modality weighting for different diagnostic scenarios, though whether this reflects limitations in actual processing versus self-assessment requires validation through controlled experiments.

The broader implications extend to fundamental questions about multimodal AI architecture. Current vision-language models typically use late fusion, where separate encoders process each modality before combining representations [35,36]. This approach, while computationally efficient, may fail to capture complex cross-modal dependencies essential for medical reasoning [37,38]. Early fusion architectures that enable deeper integration from initial processing stages, or attention mechanisms explicitly trained to modulate cross-modal influence based on task demands, may better support the dynamic modality balancing that expert diagnosis requires. Our finding that incorrect diagnoses systematically show imbalanced modality utilization provides empirical motivation for such architectural innovations.

Clinical Safety and Deployment Implications

Examination of incorrect responses revealed 2 critical failure patterns that have distinct clinical implications. First, the model frequently generated hallucinated rationales citing nonexistent

XSL•FO
RenderX

findings [24], consistent with documented hallucination rates of 35.5% to 63% in GPT-4V radiology applications [10,12,24]. Second, some errors reflected overemphasis on prominent visual findings while neglecting subtle clinical context, demonstrating that visual misinterpretations can lead the model astray when clinical information is insufficiently weighted.

These failure patterns carry distinct clinical risks requiring targeted mitigation strategies. Hallucinated findings are particularly dangerous because they appear authoritative and specific, potentially misleading clinicians who may not independently verify each claimed observation. In our study, hallucinations included references to imaging features not present in the provided images, incorrect anatomical localizations, and fabricated quantitative measurements. Such errors could lead to unnecessary interventions, incorrect diagnoses being entered into medical records, or delayed recognition of actual pathology.

The image-dominant failure mode presents a different risk profile. By over-weighting visual information that it cannot accurately interpret, GPT-4V may generate diagnoses that superficially align with prominent imaging features while missing the correct diagnosis that clinical context would suggest. This pattern is especially concerning in cases where imaging findings are nonspecific, but clinical history is highly discriminating. For example, ring-enhancing lesions have broad differential diagnoses, but patient age, immune status, and geographic location dramatically narrow possibilities [39]. A system that overrelies on imaging might suggest common etiologies based on visual appearance while missing the correct diagnosis apparent from clinical context.

These limitations mandate restricted deployment. GPT-4V should be implemented only as an educational tool or decision-support aid that highlights findings for human review but never as an autonomous diagnostic system. Any radiological application must include mandatory radiologist oversight, with AI output supplementing rather than replacing expert, as emphasized in multisociety professional guidelines [40-42]. Institutional protocols should explicitly prohibit applications bypassing human review. These restrictions remain necessary until multimodal integration capabilities achieve consistent, balanced utilization of both clinical and imaging information.

Specific implementation guidelines should include (1) interface design that presents AI outputs as preliminary suggestions explicitly requiring verification rather than definitive conclusions [43-45]; (2) transparent uncertainty quantification, ideally displaying the model's self-reported modality contributions alongside confidence estimates to flag high-risk image-dominant attributions; (3) training programs educating users about characteristic failure modes, particularly the tendency toward hallucinated findings and image-dominant errors; (4) future quality assurance protocols could explore whether AI attribution patterns predict diagnostic errors, though the 70% threshold observed in our small sample requires validation across larger, diverse datasets before clinical implementation; and (5) mandatory documentation of AI involvement in clinical reports to ensure appropriate medicolegal clarity and enable post hoc analysis of AI-associated diagnostic errors.

Workflow integration must preserve rather than undermine human expertise. Systems should be designed as "AI-assisted" rather than "AI-augmented" workflows, maintaining radiologist agency and encouraging critical evaluation. Evidence from other domains suggests that over-reliance on AI recommendations (automation bias) can degrade human performance, particularly when users lack mechanisms to assess AI reliability [43,44]. Interfaces should therefore facilitate the easy verification of AI claims, such as by highlighting specific image regions purportedly showing claimed findings, enabling radiologists to quickly confirm or refute visual interpretations.

Regulatory frameworks must evolve to address multimodal AI's unique challenges. Traditional medical device regulations focus on performance metrics including sensitivity, specificity, and accuracy but may inadequately address systematic failure modes, such as modality-specific overreliance or hallucination propensity [46,47]. Regulatory approval should require (1) comprehensive characterization of failure modes across diverse clinical scenarios, (2) validation that modality integration patterns align with domain expertise, (3) demonstration of appropriate uncertainty quantification, and (4) postmarket surveillance systems tracking AI-associated diagnostic errors. Our finding that image-dominant attribution predicts errors suggests that regulatory frameworks should incorporate modality balance metrics, potentially flagging deployments where typical attribution patterns diverge substantially from expert norms.

Educational implications are equally important. Radiology trainees must develop critical AI literacy, understanding both capabilities and characteristic failure modes of multimodal systems [48-50]. Training should include (1) recognition of hallucinated findings and strategies for systematic verification; (2) awareness that confident AI outputs may reflect overreliance on misinterpreted visual features; (3) skills in integrating AI suggestions with clinical reasoning rather than accepting them uncritically; and (4) understanding of when AI assistance is likely beneficial versus potentially misleading. Paradoxically, effective AI integration may require heightened rather than reduced emphasis on foundational clinical-radiological correlation skills [51-53], as overreliance on AI tools can diminish core competencies including diagnostic reasoning and clinical pattern recognition.

Comparison with human diagnostic errors provides important context. Radiologists also commit errors, with estimated miss rates varying by modality and pathology but often ranging from 3% to 5% for routine interpretations to 30% for subtle or complex findings [54,55]. However, human errors typically differ qualitatively from AI failures. Radiologists rarely hallucinate findings that do not exist; rather, they may overlook subtle abnormalities or misclassify ambiguous features [54,55]. Human errors often reflect attention limitations, cognitive biases, or knowledge gaps [56,57]. These are failure modes with well-established mitigation strategies, such as double-reading, checklists, and continuing education [58,59]. In contrast, AI hallucinations and systematic modality imbalances represent novel failure modes requiring new quality assurance approaches.

XSL•FO
RenderX

Our observed accuracy, while exceeding prior GPT-4V studies, remains below expert radiologist performance and insufficient for autonomous deployment. However, the more fundamental concern is not the accuracy level per se but the nature of failures. A system with 76% accuracy that fails randomly might be safely deployable with appropriate oversight, as human review would catch diverse errors. But a system showing systematic failure patterns (like our finding that image-dominant attribution reliably predicts errors) requires more cautious implementation, as certain case types may be systematically mishandled. Future deployment decisions must consider not only overall performance but failure pattern predictability and their alignment with human error patterns.

Our finding of higher overall image utilization compared to some studies [22] may reflect explicit prompting to quantify image contribution or the imaging-centric nature of neuroradiology cases. That Busch et al [21] demonstrated GPT-4V's superiority over text-only approaches in some tasks suggests the relationship between modality contribution and diagnostic success is task-dependent and complex.

Limitations

Several important limitations affect the interpretation of our findings. The primary concern is reliance on self-reported attribution of image versus text utilization, which may represent post hoc rationalizations rather than actual information processing. Rigorous validation requires controlled experiments comparing text-only, image-only, and multimodal conditions with information-theoretic metrics, such as mutual information between modalities and diagnostic accuracy.

The sample size of 29 neuroradiology cases limits statistical power for subgroup analyses and restricts generalizability to other radiological subspecialties. Performance in other subspecialties may differ substantially [21], and findings should not be extrapolated beyond adult neuroradiology. Multiple-choice format may overestimate performance relative to free-response clinical scenarios. The absence of ablation controls (text-only or image-only conditions) prevents the quantitative decomposition of relative modality contributions. Despite restricted RSNA access, data leakage cannot be definitively excluded with closed-source models. Finally, narrative justifications may not accurately reflect actual reasoning processes [24], limiting confidence in interpreting self-reported modality attribution.

Implications

For AI in radiology, these results highlight the importance of moving beyond simple accuracy metrics toward mechanistic understanding of how multimodal systems process heterogeneous data. Future research should prioritize controlled experimental validation through systematic input manipulation, development of information-theoretic frameworks for quantifying true (rather than self-reported) modality contributions, and standardized test sets with confirmed provenance postdating model training to definitively address data leakage concerns.

Technical improvements must focus on enhancing multimodal integration through architectural innovations or specialized training that forces explicit cross-referencing of visual and textual features. Interface design should evolve beyond simply adding AI outputs to workflows; instead, it enables systems to express uncertainty transparently, highlight specific image regions, and respond to targeted clinician queries. Domain specialization through fine-tuning on curated radiology datasets remains essential, as general-purpose models exhibit variable performance across subspecialties.

Most critically, broader implications extend to establishing evidence-based frameworks for human-AI collaboration in clinical medicine. Current multimodal AI systems show promise as educational tools and decision-support aids but remain inappropriate for autonomous diagnostic applications. The field must resist premature deployment driven by technological enthusiasm, instead insisting on rigorous validation of both diagnostic accuracy and decision-making transparency. With continued technological advancement focused on balanced, context-aware data integration and systematic evaluation methodologies, future generations of multimodal AI may achieve robust, reliable performance necessary for meaningful contribution to radiologic practice and patient care.

Conclusions: Broader Implications

This study contributes benchmark performance data and generates testable hypotheses about information integration patterns in diagnostic reasoning. The findings underscore that achieving high diagnostic accuracy requires more than adding visual capabilities to language models but demands sophisticated, balanced integration of clinical context and imaging findings. The exploratory observation that diagnostic failures may associate with imbalanced modality utilization suggests specific failure modes worthy of rigorous experimental investigation.

GPT-4V achieved 76% diagnostic accuracy on expert-validated neuroradiology cases, substantially exceeding prior GPT-4V performance on RSNA materials (43% by Mukherjee et al [34]). This improvement suggests that focused domain application with structured prompting may enhance performance, though experimental studies with controlled manipulation of these factors would be needed to establish causal relationships. However, the novel finding that all incorrect diagnoses associated with image-dominant attribution patterns, with substantially higher visual reliance than correct diagnoses and a very large effect size, identifies a potentially systematic failure mode requiring targeted mitigation. Until multimodal AI systems demonstrate consistent, balanced integration of clinical and imaging information with transparent uncertainty quantification, deployment should remain restricted to supervised educational and decision-support applications with mandatory radiologist oversight.

With continued technological advancement focused on balanced, context-aware data integration and systematic evaluation methodologies, future generations of multimodal AI may achieve the robust, reliable performance necessary for meaningful contribution to radiologic practice and patient care.

XSL•FO
RenderX

Acknowledgments

We gratefully acknowledge Adli Gates and Isaac Atkinson for their critical reading and meticulous editing of the manuscript. Their thoughtful feedback and detailed review significantly improved the clarity, structure, and overall quality of this work. We also thank the anonymous reviewers for their constructive comments that helped strengthen the methodological rigor and scholarly contribution of this study. This work would not have been possible without LECOM Research and Scholarship support.

We used ChatGPT and Claude for text proofreading and reference formatting, both of which were reviewed by the authors' team. We used Claude for drawing Table S1 and statistical analyses.

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative artificial intelligence (GAI) tools under full human supervision: idea generation, formulating research questions and hypotheses, feasibility assessment and risk evaluation, literature search and systematization, writing the literature review, evaluation of the novelty of the research and identification of gaps, development of experimental or research protocols, data collection, data curation and organization, data analysis, visualization, text generation, proofreading and editing, summarizing text, formulation of conclusions, reformatting, quality assessment, trend identification, and identification of limitations.

The GAI tool used was Claude Sonnet 4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

We used Claude to assist with statistical analysis.

Declaration submitted by: TW

Funding

The authors declared no financial support was received for this work.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Diagnostic performance and self-reported modality attribution of GPT-4 with Vision in neuroradiology cases.
[DOCX File, 47 KB - neuro_v5i1e69708_app1.docx]

Multimedia Appendix 2
Study flow diagram and case distribution by pathology category for cross-sectional evaluation of GPT-4 with Vision diagnostic performance in neuroradiology.
[DOCX File, 49 KB - neuro_v5i1e69708_app2.docx]

Multimedia Appendix 3
Complete case catalog with metadata, modality attribution, and performance summary by pathology category.
[DOCX File, 53 KB - neuro_v5i1e69708_app3.docx]

Multimedia Appendix 4
Complete prompt template used to elicit diagnostic responses and modality attribution from GPT-4 with Vision, with representative example response.
[DOCX File, 329 KB - neuro_v5i1e69708_app4.docx]

Multimedia Appendix 5
Example of neuroradiology case questions for GPT-4 with Vision evaluation.
[DOCX File, 440 KB - neuro_v5i1e69708_app5.docx]

Multimedia Appendix 6

XSL•FO
RenderX

Operational definitions, measurement specifications, and validation rules for primary and exploratory variables in GPT-4 with Vision neuroradiology diagnostic study.
[DOCX File, 49 KB - neuro_v5i1e69708_app6.docx]

Multimedia Appendix 7
Data quality verification, statistical assumptions testing, and comprehensive descriptive statistics, including completeness analysis, outlier detection, normality assessment, and variance homogeneity for GPT-4 with Vision neuroradiology study.
[DOCX File, 49 KB - neuro_v5i1e69708_app7.docx]

Multimedia Appendix 8
Post hoc power analysis, sensitivity analysis, and sample size justification for cross-sectional study of GPT-4 with Vision diagnostic performance in neuroradiology.
[DOCX File, 47 KB - neuro_v5i1e69708_app8.docx]

Multimedia Appendix 9
Complete statistical analysis results including primary diagnostic accuracy, exploratory modality attribution comparisons, and distribution patterns for GPT-4 with Vision neuroradiology evaluation.
[DOCX File, 49 KB - neuro_v5i1e69708_app9.docx]

References

1. Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science 2023 Sep 15;381(6663):adk6139. [doi: 10.1126/science.adk6139] [Medline: 37708283]
2. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
3. Lee JO, Zhou HY, Berzin TM, Sodickson DK, Rajpurkar P. Multimodal generative AI for interpreting 3D medical images and videos. NPJ Digit Med 2025 May 13;8(1):273. [doi: 10.1038/s41746-025-01649-4] [Medline: 40360694]
4. Simon BD, Ozyoruk KB, Gelikman DG, Harmon SA, Türkbey B. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. Diagn Interv Radiol 2025 Jul 8;31(4):303-312. [doi: 10.4274/dir.2024.242631] [Medline: 39354728]
5. Huang SC, Jensen M, Yeung-Levy S, Lungren MP, Poon H, Chaudhari AS. A systematic review and implementation guidelines of multimodal foundation models in medical imaging. Res Sq. Preprint posted online on Apr 28, 2025. [doi: 10.21203/rs.3.rs-5537908/v1]
6. Yildirim N, Richardson H, Wetscherek MT, et al. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. Presented at: CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing System; May 11-16, 2024; Honolulu HI, USA p. 1-22. [doi: 10.1145/3613904.3642013]
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
8. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. Acad Radiol 2024 Sep;31(9):3872-3878. [doi: 10.1016/j.acra.2024.06.046] [Medline: 39013736]
9. Sodhi KS, Tao TY, Seymore N. ChatGPT: chasing the storm in radiology training and education. Indian J Radiol Imaging 2023 Oct;33(4):431-435. [doi: 10.1055/s-0043-1774743] [Medline: 37811181]
10. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. Eur Radiol 2025 Apr;35(4):1959-1965. [doi: 10.1007/s00330-024-11035-5] [Medline: 39214893]
11. Sussan TT, Sussan RJ, Atkinson AG, et al. A comparative evaluation of GPT-4 Turbo and Gemini-Pro in medical licensing exams: enhancing artificial intelligence's role in medical education. Cureus 2026 Jan;18(1):e101101. [doi: 10.7759/cureus.101101] [Medline: 41658821]
12. Huppertz MS, Siepmann R, Topp D, et al. Revolution or risk?—Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol 2025 Mar;35(3):1111-1121. [doi: 10.1007/s00330-024-11115-6] [Medline: 39422726]
13. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-4V (GPT-4 with vision) on detection of radiologic findings on chest radiographs. Radiology 2024 May;311(2):e233270. [doi: 10.1148/radiol.233270] [Medline: 38713028]
14. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. Radiology 2024 Jul;312(1):e240273. [doi: 10.1148/radiol.240273] [Medline: 38980179]
15. Horiuchi D, Tatekawa H, Oura T, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol 2024 Dec;34(4):779-787. [doi: 10.1007/s00062-024-01426-y] [Medline: 38806794]

XSL•FO
RenderX

16. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text-and image-based ACR diagnostic radiology in-training examination questions. Radiology 2024 Sep;312(3):e240153. [doi: 10.1148/radiol.240153] [Medline: 39225605]
17. Albaqshi A, Ko JS, Suh CH, et al. Evaluating diagnostic accuracy of large language models in neuroradiology cases using image inputs from JAMA neurology and JAMA clinical challenges. Sci Rep 2025 Nov 27;15(1):43027. [doi: 10.1038/s41598-025-06458-z] [Medline: 41309648]
18. Nguyen D, Bronson I, Chen R, Kim YH. A systematic review and meta-analysis of GPT-based differential diagnostic accuracy in radiological cases: 2023-2025. Front Radiol 2025;5:1670517. [doi: 10.3389/fradi.2025.1670517] [Medline: 41229708]
19. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. NPJ Digit Med 2024 Aug 7;7(1):205. [doi: 10.1038/s41746-024-01208-3] [Medline: 39112822]
20. Schramm S, Preis S, Metz MC, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. Radiology 2025 Jan;314(1):e240689. [doi: 10.1148/radiol.240689] [Medline: 39835982]
21. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res 2024 May 1;26:e54948. [doi: 10.2196/54948] [Medline: 38691404]
22. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Evaluating ChatGPT-4's diagnostic accuracy: impact of visual data integration. JMIR Med Inform 2024 Apr 9;12:e55627. [doi: 10.2196/55627] [Medline: 38592758]
23. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol 2025 Jan;35(1):506-516. [doi: 10.1007/s00330-024-10902-5] [Medline: 38995378]
24. Jin Q, Chen F, Zhou Y, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med 2024 Jul 23;7(1):190. [doi: 10.1038/s41746-024-01185-7] [Medline: 39043988]
25. Deng J, Heybati K, Shammas-Toma M. When vision meets reality: exploring the clinical applicability of GPT-4 with vision. Clin Imaging 2024 Apr;108(3):110101. [doi: 10.1016/j.clinimag.2024.110101] [Medline: 38341880]
26. Parillo M, Vaccarino F, Beomonte Zobel B, Mallio CA. ChatGPT and radiology report: potential applications and limitations. Radiol Med 2024 Dec;129(12):1849-1863. [doi: 10.1007/s11547-024-01915-7] [Medline: 39508933]
27. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997. [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
28. Wada A, Akashi T, Shih G, et al. Optimizing GPT-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. Diagnostics (Basel) 2024 Jul 17;14(14):1541. [doi: 10.3390/diagnostics14141541] [Medline: 39061677]
29. Strotzer QD, Nieberle F, Kupke LS, et al. Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. Radiology 2024 Nov;313(2):e240955. [doi: 10.1148/radiol.240955] [Medline: 39589253]
30. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024 Jan;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
31. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
32. Dicle O, Özan S, Şahin H, Seçil M. How to perform an excellent radiology board examination: a web-based checklist. Insights Imaging 2021 Jan 7;12(1):4. [doi: 10.1186/s13244-020-00924-0] [Medline: 33411060]
33. Definitions for purposes of this policy (45 CFR §46.102). National Code of Federal Regulations. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-04-20]
34. Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 case of the day questions. Radiology 2024 Oct;313(1):e240609. [doi: 10.1148/radiol.240609] [Medline: 39352277]
35. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 2020;3(1):136. [doi: 10.1038/s41746-020-00341-z] [Medline: 33083571]
36. Alsentzer E, Murphy J, Boag W. Publicly available clinical BERT embeddings. 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 7, 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/W19-1909]
37. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? Ann Intern Med 2020 Jan 7;172(1):59-60. [doi: 10.7326/M19-2548] [Medline: 31842204]
38. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]

XSL•FO
RenderX

39. Schwartz KM, Erickson BJ, Lucchinetti C. Pattern of T2 hypointensity associated with ring-enhancing brain lesions can help to differentiate pathology. Neuroradiology 2006 Mar;48(3):143-149. [doi: 10.1007/s00234-005-0024-5] [Medline: 16447037]
40. Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR and RSNA. Radiol Artif Intell 2024 Jan;6(1):e230513. [doi: 10.1148/ryai.230513] [Medline: 38251899]
41. Dreyer KJ, Geis JR. When machines think: radiology's next frontier. Radiology 2017 Dec;285(3):713-718. [doi: 10.1148/radiol.2017171183] [Medline: 29155639]
42. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020 Mar;2(2):e200029. [doi: 10.1148/ryai.2020200029] [Medline: 33937821]
43. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127. [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
44. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA 2023 Dec 19;330(23):2275-2284. [doi: 10.1001/jama.2023.22295] [Medline: 38112814]
45. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3:41. [doi: 10.1038/s41746-020-0253-3] [Medline: 32219182]
46. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021 Apr;27(4):582-584. [doi: 10.1038/s41591-021-01312-x] [Medline: 33820998]
47. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
48. van Kooten MJ, Tan CO, Hofmeijer EIS, et al. A framework to integrate artificial intelligence training into radiology residency programs: preparing the future radiologist. Insights Imaging 2024 Jan 17;15(1):15. [doi: 10.1186/s13244-023-01595-3] [Medline: 38228800]
49. Tajmir SH, Alkasab TK. Toward augmented radiologists: changes in radiology education in the era of machine learning and artificial intelligence. Acad Radiol 2018 Jun;25(6):747-750. [doi: 10.1016/j.acra.2018.03.007] [Medline: 29599010]
50. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018 Oct 24;2(1):35. [doi: 10.1186/s41747-018-0061-6] [Medline: 30353365]
51. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
52. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019 Dec 3;5(2):e16048. [doi: 10.2196/16048] [Medline: 31793895]
53. Park CJ, Yi PH, Siegel EL. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. Curr Probl Diagn Radiol 2021;50(5):614-619. [doi: 10.1067/j.cpradiol.2020.06.011] [Medline: 32680632]
54. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015 Oct;35(6):1668-1676. [doi: 10.1148/rg.2015150023] [Medline: 26466178]
55. Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive error in radiology. AJR Am J Roentgenol 2017 Apr;208(4):739-749. [doi: 10.2214/AJR.16.16963] [Medline: 28026210]
56. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. J Am Coll Radiol 2010 Sep;7(9):698-704. [doi: 10.1016/j.jacr.2010.03.004] [Medline: 20816631]
57. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. Radiographics 2018;38(1):236-247. [doi: 10.1148/rg.2018170107] [Medline: 29194009]
58. Eng J, Mysko WK, Weller GE, et al. Interpretation of emergency department radiographs: a comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. AJR Am J Roentgenol 2000 Nov;175(5):1233-1238. [doi: 10.2214/ajr.175.5.1751233] [Medline: 11044013]
59. Larson DB, Nance JJ. Rethinking peer review: what aviation can teach radiology about performance improvement. Radiology 2011 Jun;259(3):626-632. [doi: 10.1148/radiol.11102222] [Medline: 21602501]

Abbreviations

AI: artificial intelligence
CT: computed tomography
GPT-4V: GPT-4 with Vision
LLM: large language model
MRI: magnetic resonance imaging
RSNA: Radiological Society of North America

XSL•FO
RenderX

Edited by S Brini; submitted 05.Dec.2024; peer-reviewed by F Liu, S Fitzek, S Zawada, S Chen, VSK Kancharla; revised version received 20.Jan.2026; accepted 21.Jan.2026; published 30.Apr.2026.

Please cite as:
Sussan TT, Brawley RR, Eckroth J, Mossell JE, Weitao T
Diagnostic Accuracy of GPT-4 With Vision in Neuroradiology Board-Style Exam Questions: Cross-Sectional Case-Based Study
JMIR Neurotech 2026;5:e69708
URL: https://neuro.jmir.org/2026/1/e69708
doi:10.2196/69708

© Tom T Sussan, Rebekah R Brawley, Joshua Eckroth, James E Mossell, Tao Weitao. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 30.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

xHD-Vox, an Automated Speech Model for Estimating Motor and Cognitive Scores in Huntington Disease: Development and Longitudinal Validation

Tiphaine Le Ludec[1,2,3], MSc; Andres Gil-Salcedo[1,2,3], PhD; Hadrien Titeux[1,2,3], MSc; Robin Louiset[1,2,3], PhD; Clément Le Moine Veillon[1,2,3], PhD; Renaud Massart[1,2,3], MBA, PhD; Anne-Catherine Bachoud-Lévi[1,2,3,4,5], PhD, Prof Dr Med

[1]Team 01 - Neuropsychologie Interventionnelle, INSERM U955, Mondor Institute of Biomedical Research, Créteil, France

[2]NeurATRIS, Créteil, France

[3]Department of Cognitive Science, Ecole Normale Supérieure (ENS), 29 rue d'Ulm, Paris, France

[4]Reference Center for Huntington's Disease, Department of Neurology, Henri Mondor-Albert Chenevier University Hospitals, Créteil, France

[5]Faculty of Medicine, Université Paris-Est Créteil, Créteil, France

Corresponding Author:
Anne-Catherine Bachoud-Lévi, PhD, Prof Dr Med
Team 01 - Neuropsychologie Interventionnelle, INSERM U955, Mondor Institute of Biomedical Research, Créteil, France

Abstract

Background: Huntington disease (HD) is a rare genetic neurodegenerative disease that causes progressive motor, cognitive, and psychiatric symptoms over decades after onset. Clinical care is typically provided in specialized centers with only annual clinical assessments, highlighting the need for more frequent and cost-effective monitoring.

Objective: This study aimed to develop and validate xHD-Vox, a fully automated, interpretable, speech-based model for predicting the composite Unified Huntington Disease Rating Scale (cUHDRS) and its cognitive, motor, and functional components.

Methods: We included 181 HD gene carriers (341 annual visits) from three French prospective cohorts: BIO-HD (NCT01412125), REPAIR-HD (NCT03119246), and MIG-HD (NCT00190450). Participants had ≥40 cytosine-adenine-guanine (CAG) repeats, available cUHDRS scores, and audio recordings of forward and backward counting (1-20). For model development and feature selection, we used a speech pathologist-annotated subset (145 visits and 90 participants). Selected speech features were then automated using Whisper, an open-source speech recognition tool. The final linear regression model, xHD-Vox, was calibrated on the training set of 269 visits (157 participants, and annotated subset included). Performance was evaluated on an independent longitudinal test set (24 participants, with 3 annual visits each) using mean absolute error, explained variance (R^2), and intraclass correlation coefficient. Longitudinal decline was assessed with 2-way repeated-measures ANOVAs. Predicted 1-year and 2-year changes were compared with clinician-assessed 95% CIs.

Results: Feature selection identified four key predictors: standardized CAG-age-product score, CAG repeat length, rate of numbers pronounced per second, and the SD of that rate. On the test set, xHD-Vox achieved a mean absolute error of 2.1 for cUHDRS and explained 57% of its variance, compared with 38% when using only demographic features. Longitudinal analyses using repeated-measures ANOVAs with post hoc Tukey tests confirmed a significant decline over the 2-year follow-up for both clinician-assessed measures and xHD-Vox predictions. At the group level, the mean 1-year and 2-year changes predicted by xHD-Vox were consistent with clinically measured changes, falling within the corresponding 95% CIs.

Conclusions: We developed xHD-Vox, an interpretable and automated model that predicts clinical scores in HD using a short speech task. Predicted scores were consistent with clinician-assessed scores, supporting its potential use in mobile apps for remote monitoring. This approach could facilitate scalable, real-time tracking of disease progression, especially in underserved regions, and enable personalized and responsive clinical care.

Trial Registration: ClinicalTrials.gov NCT01412125; https://clinicaltrials.gov/study/NCT01412125 and NCT03119246; https://clinicaltrials.gov/study/NCT03119246 and NCT00190450; https://clinicaltrials.gov/study/NCT00190450

(JMIR Neurotech 2026;5:e83838) doi:10.2196/83838

KEYWORDS

huntington; speech; voice; automated; cUHDRS; UHDRS; machine learning; longitudinal; telemonitoring; motor and cognitive function; composite Unified Huntington Disease Rating Scale; Unified Huntington Disease Rating Scale

XSL•FO
RenderX

Introduction

Huntington disease (HD) is a rare autosomal neurodegenerative disorder caused by an expanded cytosine-adenine-guanine (CAG) repeat (>35) in the *huntingtin* gene [1]. The disease manifests through a variety of motor, cognitive, and psychiatric symptoms, requiring lifelong multidisciplinary care. While no cure is currently available despite extensive research on disease modifiers, symptomatic treatments are effective in managing disability for many years. Therefore, continuous and personalized monitoring of disease progression is essential to optimize patient care. However, access to specialized expertise remains limited, highlighting the need for scalable, automated approaches to support clinical decision-making.

The composite Unified Huntington disease Rating Scale (cUHDRS) is a robust metric for tracking disease progression, integrating motor, cognitive, and functional scores from the UHDRS [2]—the worldwide reference scale for HD clinical evaluation. Its assessment is typically conducted once a year during face-to-face examinations by trained neurologists and neuropsychologists with restricted availability [3]. This calls for cost-effective and remote-friendly solutions to enable continuous disease monitoring.

Speech analysis has emerged as a promising, noninvasive, and practical tool for tracking HD progression [4,5]. While other biomarkers, such as striatal atrophy on magnetic resonance imaging, cerebrospinal fluid neurofilament levels, and cognitive batteries, have been validated to monitor disease progression, speech-based assessments are well-suited for remote, home-based implementation, reducing the logistical burden and costs associated with in-clinic visits. In our previous study [5], we demonstrated that speech features could predict cUHDRS scores and their cognitive and motor components more accurately than the combination of demographic and genetic data [5]. These features were extracted from a simple counting task, in which participants counted forward and backward from 1 to 20 in less than 45 seconds. However, the need for manual annotations by speech pathologists still limited the model's scalability and implementation as a telemonitoring solution [6,7].

To address these limitations, we developed xHD-Vox (x for eXplainable, HD for Huntington disease, and Vox for voice), an interpretable and automated model that predicts cUHDRS and its motor and cognitive components from brief speech recordings in HD gene carriers. This study comprises the following two main parts: (1) xHD-Vox development, where we ensured interpretability by using a standard linear regression model and selecting a subset of relevant speech features identified in our previous work [5]. Manual annotations were replaced with features extracted using Whisper, an open-source automatic speech recognition (ASR) system developed by OpenAI, known for its high transcription accuracy and multilingual support [8]. (2) xHD-Vox evaluation on an independent longitudinal dataset to evaluate its ability to track disease progression over time.

Methods

Participants

Participants were enrolled from three prospective longitudinal studies: BIO-HD (NCT01412125), REPAIR-HD (NCT03119246), and MIG-HD (NCT00190450) before transplants.

We selected data from HD gene carriers with a CAG repeat length of greater than or equal to 40 to ensure the complete penetrance of HD [9]. We excluded visits with missing speech recordings or clinical scores required to compute cUHDRS. This resulted in the selection of 181 participants, with a total of 341 visits (Figure 1).

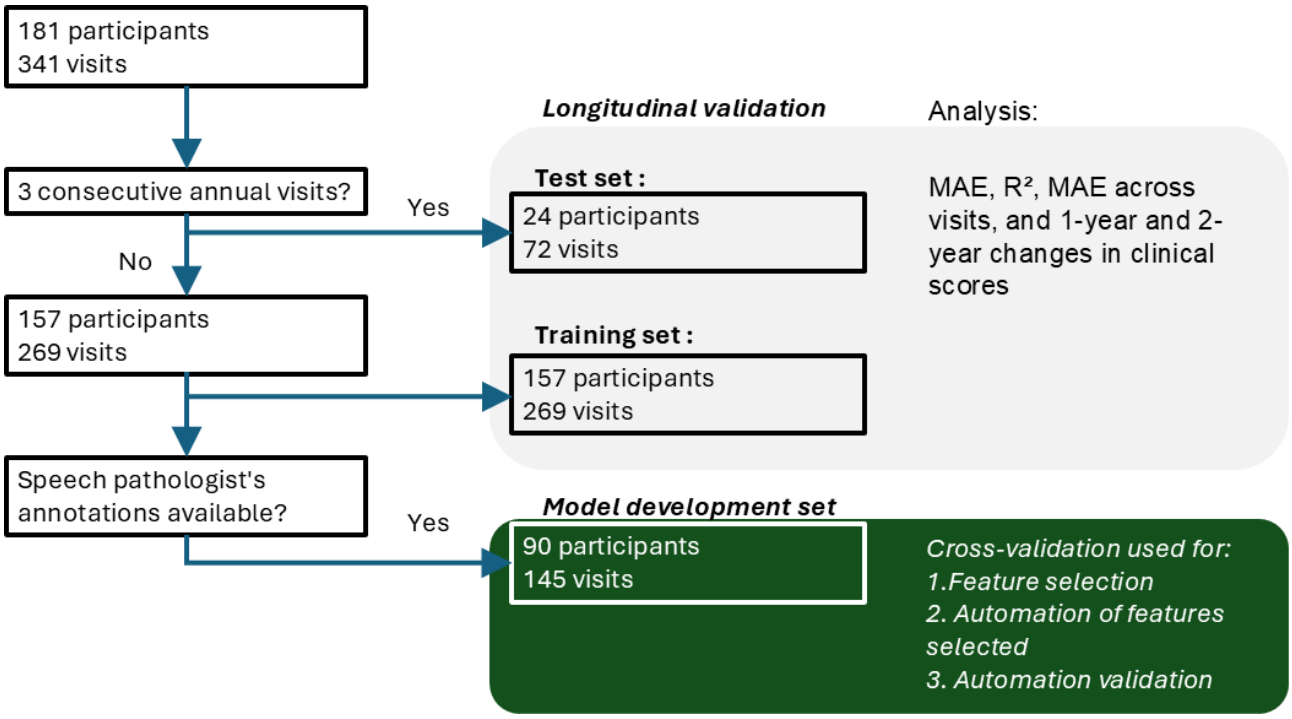
For the evaluation of our model on an independent longitudinal test set, we selected participants who had 3 consecutive annual visits, which resulted in a cohort of 24 participants (72 visits). Assessing performance over a 2-year follow-up period ensured a sufficiently long observation window to capture disease progression.

The remaining 157 participants, comprising 269 visits (mean 1.7, SD 1.1), were used for model training.

Among these, 145 recordings from 90 participants were annotated by speech pathologists and used to develop the automated model.

XSL•FO
RenderX

Figure 1. Participant selection flowchart for the development and validation study. The test set for validation is first isolated from the database. Model development is performed on a subset of the training set with speech pathologists' annotations available. MAE: mean absolute error; R^2: explained variance.



Clinical Evaluation

Participants were assessed by annually certified examiners with the Unified Huntington Disease Rating Scale (UHDRS). For this study, we used the cUHDRS and its four components to evaluate the disease progression: Total Motor Score (TMS), the Symbol Digit Modalities Test (SDMT), the Stroop Word (SW), and the Total Functional Capacity (TFC). The cUHDRS was calculated using the following formula [2]:

$$cUHDRS = TFC - 10419 \cdot TMS - 29749 + SDMT - 284113 + SW - 661201 + 10$$

The CAG-age-product (CAP) score was calculated using the standardized formula calibrated to reach 100 at the expected age of onset [10]:

$$CAP = age \times CAG - 306.49$$

Standardized Lightweight Speech Task

Speech samples were recorded in a clinical research environment at Henri Mondor Hospital through 2 brief controlled tasks. Each participant was asked by a neuropsychologist to (1) count aloud numbers from 1 to 20 (forward counting), then (2) to count backward from 20 to 1 while holding their hands up and closing their eyes (backward counting). The backward task increases cognitive load by inhibiting the forward counting process [5]. Recordings were made using an external microphone (Zoom H4n Pro; 44 kHz recording, resampled at 16 kHz) for the BIO-HD and REPAIR-HD cohorts and were extracted from the audio track of video recordings for MIG-HD. The microphone was placed on the neuropsychologist's desk, facing the patient, who was seated approximately 1 - 2 meters away.

For the annotated samples, speech recordings were transcribed by speech therapists at the word level using the Praat (developed by Paul Boersma, David Weenink, and Anastasia Shchupak) and Seshat (developed by Hadrien Titeux, Rachid Riad, Xuan-Nga Cao, Nicolas Hamilakis, Kris Madden, Alejandrina Cristia, Anne-Catherine Bachoud-Lévi, and Emmanuel Dupoux) platforms [11]. Mispronunciations (using the Speech Assessment Methods Phonetic Alphabet), paraphasias, phoneme perseverations, abnormal breathing, vocal noises, and filled pauses ("euh" and "um") were identified.

Analysis Implementation

This study followed the TRIPOD+AI (Transparent Report of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) reporting guideline. We conducted the analysis in Python using the scikit-learn library [12]. Features were normalized with RobustScaler to enhance feature importance analysis and improve the model's convergence. For cross-validation, we generated 50 folds, with each fold containing 80% of participants in the training set and the remaining 20% in the test set. Each participant was assigned exclusively to either the training or test set within each fold.

Automated Model Development

Overview

In this first analysis, we automated the model developed by Riad et al [5] for predicting cUHDRS and its components using a parsimonious approach to reduce model complexity while preserving predictive accuracy. This approach involved the following three steps: (1) ranking the most predictive features of Riad et al [5] model, (2) identifying the smallest subset of features that achieved at least the same performance as the full feature set by progressively adding features in order of importance, and (3) automating the annotation process for the selected features using Whisper.

XSL•FO
RenderX

Ranking the Most Predictive Features of our Previous Model

Features Investigated

We applied the automatic pipeline for feature extraction based on speech pathologists' annotations previously developed by Riad et al [5]. This process resulted in the extraction of 60 speech features (30 features for each of the 2 tasks), including articulatory and phonatory deficiencies, rhythm and temporal statistics, filled pauses, vocalization additions, and sequence-related metrics (eg, errors in number order and perseveration errors) (Table S1 in Multimedia Appendix 1 provides the detailed list of features extracted). We excluded the feature phones per second due to their high correlation (r=0.99) with pronounced numbers per second. Demographic features, including CAP score, age, and CAG repeat length, were added, leading to a total of 63 features being investigated.

Features Importance Ranking

Riad et al [5] model is a 63-feature ElasticNet regression, a standard linear regression with L1 and L2 regularization to prevent overfitting and mitigate multicollinearity. The study demonstrated that using only features from the backward task yields performance comparable to that achieved using features from both forward and backward tasks (ablation study). Performance was assessed using mean absolute error (MAE) across 50 cross-validation folds, which quantifies the average absolute difference between observed and predicted clinical scores.

Given similar results in our replication (Tables S2 and S3 in Multimedia Appendix 1), we retained only features from the backward counting task for further selection.

To identify the most relevant features from this task, we calculated the average absolute value of the ElasticNet regression coefficients across 50 cross-validation training folds. Features were then ranked by their relative importance for each clinical score prediction. Finally, we computed an average rank for each feature across cUHDRS, TMS, and cognitive scores. TFC was excluded from this average ranking because the model showed low explained variance (R^2) for this score, indicating poor predictive performance.

Identifying the Minimal Predictive Features Set

To identify the minimal set of features that achieved performance at least equivalent to that of the initial model [5], features were progressively added in order of importance, as determined in step (1), in a standard linear regression model. We stopped adding features once the model's performance matched or exceeded that of the initial model. Performance was evaluated using the mean MAE across 50 cross-validation folds. This method identified a minimal subset; however, it may not be globally optimal, as other combinations of features could potentially provide better predictive performance.

Automating the Selected Features' Annotation Process Using Whisper

The two selected features in step (2) were "numbers pronounced per second" and "SD of the duration of individual numbers." Numbers pronounced per second were computed as the total number of numbers spoken (including perseverations and sequence errors) divided by the task duration. Task duration was defined as the time interval from the start of the first number pronounced to the end of the last number pronounced. The SD of duration was calculated across the duration of individual spoken numbers.

To automate the extraction of these 2 features, it was necessary to identify the sequence of numbers produced by each participant within recordings that could include hesitations, fillers, or unrelated speech (eg, questions addressed to the neuropsychologist). This required accurate transcription of the recordings as well as precise text-audio alignment to compute timing-related features.

We used Whisper, a pretrained ASR model developed by OpenAI [8], which provides both transcription and timestamp alignment, thereby facilitating automated extraction of timing features. Whisper was trained on 680,000 hours of labeled audio, including multilingual datasets, and has demonstrated a low word error rate in French (word error rate=5.3 on FLEURS and 10.8 on Common Voice [version 15]). Its robustness to noisy recordings and coverage of 96 languages made it well suited for this task.

Numbers pronounced per second and the SD of the duration of individual numbers were compared using Pearson and Spearman correlations between the values computed from Whisper's annotations and those derived from speech-language pathologist annotations.

Finally, we validated our automated model by comparing the MAE of predictions based on features extracted from speech pathologists' annotations with those obtained using Whisper's alignment.

Model Evaluation Using Longitudinal Data

Overview

We evaluated our automated model on longitudinal data using a leave-one-out setup to simulate its real-world performance. The model was calibrated on the training set before being tested on unknown participants (test set).

All longitudinal assessment results are reported on the test set to ensure evaluation on an unseen cohort and avoid inflated performance estimates. Given the test set selection approach, we verified a posteriori that (1) the training set contained enough participants for proper model calibration, (2) the training/testing split followed an 80/20 ratio in terms of visits, and (3) the test and training sets were balanced in terms of disease severity (refer to the Participants subheading in the Results section for details).

Model Performance on the Test Set

The model was trained using linear regression, and its performance was evaluated on the test set using the MAE, R^2, and intraclass correlation coefficient (ICC). The R^2 metric quantifies the proportion of variance explained by the model, whereas the ICC measures the agreement between predicted and observed clinical scores. To highlight the potential contribution of speech features, we compared the performance

XSL•FO
RenderX

of our model with that of a baseline demographic model (based on CAP score and CAG repeat length).

Functional Classification

Given the low performance of the regression model for TFC, we explored functional decline using a logistic regression model trained to classify stage 3 vs stages 0 - 2 of the Huntington Disease Integrated Staging System (HD-ISS). The model included two predictors: CAP score and speech rate (measured as numbers pronounced per second). Selection of these predictors was supported by cross-validation results (Figure S3 and methods in Multimedia Appendix 1). Model performance was evaluated using Receiver Operating Characteristic–Area Under the Curve (ROC AUC), accuracy, and recall.

Longitudinal Change

To test the hypothesis of a longitudinal decline in clinical scores, both as measured by clinicians and as predicted by the model on the test set, we performed 2-way repeated-measures ANOVAs separately for each score. The 2 within-subject factors were visit time (baseline, year 1, and year 2) and type of measure (clinician-assessed vs model-predicted). Statistical significance was set at α=.05, and Bonferroni corrected for the 5 tested scores.

For each ANOVA, we examined the main effects of time and type of measure, as well as their interaction. When a significant effect of time was observed, Tukey post hoc comparisons were conducted to characterize differences between visits.

To quantify the magnitude of longitudinal change, we computed 1-year and 2-year changes in each score for both clinician-assessed and model-predicted measures. One-year change was defined as the difference between the score at year 1 and baseline, whereas 2-year change was defined as the difference between the score at year 2 and baseline. Group-level agreement between clinical and predicted changes was assessed by comparing mean changes and their 95% CIs.

To evaluate model performance at the individual level, we examined the relationship between predicted and clinically measured 1-year changes using scatterplots, Pearson correlation coefficients, and Spearman correlation coefficients. This analysis was restricted to 1-year changes to maximize the number of observations (48 paired changes), thereby increasing statistical power. Sensitivity to outliers was assessed and reported in Multimedia Appendix 1.

Ethical Considerations

All participants signed an informed consent. Ethical approval was obtained from the institutional review board of Henri Mondor Hospital for BIO-HD and MIG-HD and of Saint-Louis Hospital for the French part of REPAIR-HD. The study complied with the Declaration of Helsinki, current Good Clinical Practice guidelines, and local laws and regulations.

Results

Participants

The full database (N=181 participants, 341 visits) included individuals at early stages of the disease, with low to moderate functional impairment (mean TFC 11.2, SD 2.3). Detailed demographic and clinical characteristics are provided in Table 1.

This database was split into training and testing sets (refer to Figure 1 for the flow of subset construction). The test set comprised 21% of the total visits and was balanced in terms of disease severity compared with the training set: 50% of participants were at stage 3 in both sets, and 19% were at stage 0 - 1 (vs 32% in the training set). Clinical scores (cUHDRS, TFC, SDMT, SW, and TMS) were also close between the 2 subsets.

Our sample size analysis (Figure S2 in Multimedia Appendix 1) indicated that 100 participants were sufficient to calibrate the model with 95% accuracy; this condition was met with 157 participants in the training set.

The speech pathologist-annotated database used for model development showed a disease severity profile similar to that of the full database, with a slightly lower mean TFC (mean 10.8, SD 2.3).

A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of data selection is provided in Figure S1 in Multimedia Appendix 1. Only 1.2% of clinical data were missing.

XSL•FO
RenderX

Table . Demographic and clinical characteristics across all database subsets used in the different analyses.

Characteristic	Full database	Test set	Training set	Speech pathologist–annotated database
Number of participants	181	24	157	90
Sex, n				
Female	100	7	93	55
Male	81	17	64	35
Number of visits	341	72	269	145
Number of visits, mean (SD)	1.9 (1.1)	3 (0)	1.7 (1.1)	1.6 (0.8)
Age at first visit, mean (SD)	49.4 (11.3)	53.5 (10)	48.7 (11.4)	51.7 (10.5)
CAG[a] repetition, mean (SD)	43.6 (2.9)	42.5 (1.7)	43.7 (3)	43.6 (3.4)
HD-ISS[b] 0 - 1, n visits (%)	100 (0.29)	14 (0.19)	86 (0.32)	37 (0.25)
HD-ISS 2, n visits (%)	73 (0.22)	22 (0.3)	51 (0.19)	27 (0.19)
HD-ISS 3, n visits (%)	168 (0.49)	36 (0.5)	132 (0.49)	81 (0.56)
cUHDRS[c], mean (SD)	12.1 (4.6)	12.0 (4.1)	12.1 (4.8)	11.2 (4.6)
TMS[d], mean (SD)	21.0 (18.8)	21.3 (16.5)	20.9 (19.4)	25.2 (19)
TFC[e], mean (SD)	11.2 (2.3)	11.4 (1.9)	11.1 (2.4)	10.8 (2.3)
SDMT[f], mean (SD)	35.7 (16.3)	34.4 (14)	36 (16.9)	32.7 (16.3)
SW[g], mean (SD)	74.9 (23.7)	73.4 (21.8)	75.2 (24.1)	71.1 (23.6)

[a]CAG: cytosine-adenine-guanine.

[b]HD-ISS: Huntington Disease Integrated Staging System.

[c]cUHDRS: composite Unified Huntington Disease Rating Scale.

[d]TMS: Total Motor Score.

[e]TFC: Total Functional Capacity.

[f]SDMT: Symbol Digit Modalities Test.

[g]SW: Stroop Word.

Automated Model Development

Ranking the Most Predictive Features

Table S3 in Multimedia Appendix 1 shows that the R^2 for the TFC score was low (0.13) compared with that of the other scores, which were closer to 0.5. Although we continued to report results for TFC, it was excluded from the feature selection process because of its poor predictive performance.

The four most predictive features, based on their average rank, were the CAP score, CAG repeat length, and 2 rhythm-related speech features: numbers pronounced per second and the SD of the duration of pronounced numbers. Table 2 shows the top 15 most predictive features from the backward speech task for predicting clinical scores.

XSL•FO
RenderX

Table . Coefficient ranks of demographic and speech features from the backward task for each clinical score. The average rank is calculated as the mean rank across all clinical scores except TFC[a] (low predictive performance based on explained variance [R^2]). Features are ordered based on their average rank.

Feature	cUHDRS[b]	TMS[c]	Stroop Word	SDMT[d]	Average rank	TFC[a]
CAP[e] score	1	1	1	1	1.00	1
Pronounced numbers per second	2	5	3	2	3.00	8
CAG[f] repetition	3	2	5	7	4.25	2
SD of the duration of pronounced numbers	4	11	2	3	5.00	32
Total number of pronunciation errors	7	4	6	8	6.25	17
Temporal rate of the pronounced numbers	6	3	14	4	6.75	10
Range of the fundamental frequency	5	6	10	13	8.50	4
Temporal rate of silences	9	7	17	5	9.50	6
Pronunciation errors per second	11	8	9	11	9.75	28
Task duration	8	9	7	18	10.50	3
SD of normalized intensity of vocalizations	13	13	8	12	11.50	21
Total number of silences	12	10	20	6	12.00	14
Mean duration of pronounced numbers	10	21	4	16	12.75	13
Ratio of pronunciation errors	14	12	16	9	12.75	12
Temporal rate of the filled pauses	15	16	11	14	14.00	20
Normalized range of intensity of vocalizations	17	17	12	15	15.25	31
Participant age	16	20	15	10	15.25	23

[a]TFC: Total Functional Capacity.

[b]cUHDRS: composite Unified Huntington Disease Rating Scale.

[c]TMS: Total Motor Score.

[d]SDMT: Symbol Digit Modalities Test.

[e]CAP: CAG-age-product.

[f]CAG: cytosine-adenine-guanine.

Identifying the Minimal Predictive Features Set

Table 3 shows that using the top 4 ranked features identified in step 1.1 (Table 2) resulted in MAE values that were equal to or lower than those of the initial 63-feature model (column 6 vs column 1). Using only the 2 top-ranked features (CAP score and numbers pronounced per second) was already sufficient for cUHDRS and SDMT to achieve lower MAE compared with the 63-feature model (column 1 vs column 4). Adding CAG repeat length further reduced the MAE for TMS (column 5 vs column 4), whereas incorporating the SD of the duration of pronounced numbers leads to an additional decrease in MAE for SW prediction (column 6 vs column 5).

XSL•FO
RenderX

Table . Model performance in terms of average MAE[a] (SD) across 50 cross-validation folds for different sets of features. The goal was to identify the minimal set of predictive features that achieved performance equal to or better than the initial model from our previous study [5], which included 60 speech features, CAP[b] score, CAG[c] repeat length, and age (column 1). In columns 3-6, features were added iteratively based on their importance ranking using a standard linear regression model. The top 4 ranked features were, in order, CAP score, backward numbers per second, CAG repeat length, and the SD of backward numbers per second. Column 2 reports performance using only CAP score and CAG repeat length, that is, the 2 demographic features among the top 4 ranked predictors, for comparison. Equivalent performance with the initial 63-feature model was achieved using these 4 features. The last column reports performance using features automatically extracted with Whisper.

Clinical score	Features extracted from annotations used in the regression model						Features extracted using Whisper
	63-feature initial model [5]	Demographic baseline (CAP[b] score + CAG[c] repeat length)	Top 1 highest-ranked features	Top 2 highest-ranked features	Top 3 highest-ranked features	Top 4 highest-ranked features (minimal set)	Top 4 highest-ranked features (model automated)
cUHDRS[d], mean (SD)	2.6 (0.4)	2.7 (0.5)	2.7 (0.5)	2.2 (0.4)	2.2 (0.4)	2.2 (0.3)	2.1 (0.3)
UHDRS-TMS[e], mean (SD)	10.6 (1.1)	12.0 (1.6)	12.3 (1.8)	10.7 (1.5)	10.4 (1.3)	10.6 (1.3)	10.1 (1.1)
UHDRS-TFC[f], mean (SD)	1.7 (0.2)	1.5 (0.2)	1.5 (0.2)	1.5 (0.2)	1.4 (0.2)	1.4 (0.2)	1.3 (0.2)
SW[g], mean (SD)	11.9 (1.9)	14.4 (2.3)	14.5 (2.3)	12.1 (2.1)	12.2 (2.0)	11.6 (1.8)	11.6 (1.7)
SDMT[h], mean (SD)	8.8 (1.5)	9.7 (1.8)	9.7 (1.7)	8.3 (1.3)	8.3 (1.3)	8.4 (1.3)	8.4 (1.3)

[a]MAE: Mean Absolute Error

[b]CAP: CAG-age-product.

[c]CAG: cytosine-adenine-guanine.

[d]cUHDRS: composite Unified Huntington Disease Rating Scale.

[e]UHDRS-TMS: Unified Huntington Disease Rating Scale Total Motor Score.

[f]UHDRS-TFC: Unified Huntington Disease Rating Scale Total Functional Capacity.

[g]SW: Stroop Word.

[h]SDMT: Symbol Digit Modalities Test.

The final 4-feature model outperformed a demographic baseline model based on CAP score and CAG repeat length (column 6 vs column 2; paired *t* test with Bonferroni correction, $P<.001$ for all scores). R^2 values, reported in Table S4 in Multimedia Appendix 1, complement the MAE analysis: for example, the final model achieved a mean R^2 of 0.59 for cUHDRS (vs 0.43 for the baseline) and 0.57 for SW (vs 0.37).

The demographic baseline includes only the demographic features present in the final model to evaluate the incremental value of speech features. Age was not included as a sensitivity analysis showed that adding age to the demographic features did not improve MAE or R^2 compared to using only CAP score and CAG repeat length (Tables S5 and S6 in Multimedia Appendix 1) and introduced severe multicollinearity (variance inflation factor [VIF] >12 for all features). By contrast, VIF values were acceptable when using only CAP score and CAG repeat length for both the demographic baseline model (VIF=1.2) and the final 4-feature model (maximum VIF of 1.56). This demographic baseline model was used for subsequent comparisons with xHD-Vox to quantify the specific contribution of speech features.

Automating the Selected Features' Annotation Process Using Whisper

Whisper-extracted features and annotations-based features showed Pearson and Spearman correlations of 0.84/0.81 for numbers pronounced per second and 0.42/0.61 for the SD of the duration of pronounced numbers.

The automated pipeline achieved equal or lower MAE values than the speech pathologist-annotated pipeline (column 7 vs column 6 in Table 3), with statistically significant differences observed for TMS and TFC (paired *t* test; Bonferroni corrected; $P<.001$).

The final performance of the automated model yielded relative errors ranging from 12% for SDMT to 40% for TMS, with 13% for TFC, 19% for cUHDRS, and 35% for SW. Relative error was calculated as the MAE (column 7 in Table 3) divided by the mean cohort score (last column in Table 1).

Model Evaluation Using Longitudinal Data

Overview

First, we compared the cross-validation performance of our model obtained in (Table 3, last column) with results previously reported in the literature using comparable evaluation

XSL•FO
RenderX

frameworks (Table 4), which is further discussed in the Discussion section.

XSL•FO
RenderX

Table . Literature review of prediction models and group analyses based on speech markers in Huntington disease. For studies of speech markers of disease, *P* values are indicated when P<.05 for controls vs preHD (Saft) and preHD vs HD (Vogel).

Objective of study	Prediction model				Speech markers of disease		
Authors	This study	Riad et al [5]	Riad et al [13]	Nunes et al [4]	Vogel et al [14]	Skodda et al [15]	Saft et al [16]
Language	French	French	French	English	English	German	German
Type of study	Longitudinal	Cross-sectional	Cross-sectional	Cross-sectional analysis	Cross-sectional	Cross-sectional	Longitudinal (Y0+21 months)
No of participants	181: 41 HD-ISS[a] 0-1, 33 HD-ISS 2, and 107 HD-ISS 3	103: 87 HDs[b] and 16 PreHDs[c]	85: 45 HDs, 16 preHDs, and 24 controls	36: 18 HDs, 7 PreHDs, and 11 controls	45: 17 HDs, 13 PreHDs, and 15 controls	56: 28 preHDs and 28 controls	13 preHDs
Number of visits	341	126	85	NR:[d] follow-up every 3 - 6 months for 3 years	45	56	26
Automated?	Yes	No (speech pathologist annotations)	Yes (acoustic features)	Automated using BioDigit Speech (proprietary ASR[e])	Yes (purely acoustic)	Yes (purely acoustic)	Yes (purely acoustic)
Tasks	Counting backward	Counting backward + forward	Maximum phonation	Passage reading, counting forward (1-20), counting backward (50-30 by increments of 3), and used for classification only.	Passage reading, monologue with positive content, and automated sample (days of the week)	Reading, sustained phonation, maximum syllable repetition capacity, and steadiness of syllable repetition.	Reading, sustained phonation, maximum syllable repetition capacity, and steadiness of syllable repetition.
Task duration	≤40 seconds	≤80 seconds	3 seconds	≤2 minutes	NR (presumably >2 minutes)	NR (presumably >2 minutes)	NR (presumably >2 minutes)
Number of features	4	63	16	14	6	6	6
Features	• CAP[f] score • Numbers pronounced per second • CAG[g] repeat length • SD of numbers pronounced per second	• Sixty speech features (listed in Table S1 in Multimedia Appendix 1). • CAP score • CAG repeat length • Age	• Phonatory features and modulation power spectrum features	• Features for reading and counting task: rhythm features, sequence-error perseverations, and articulatory and phonatory deficiencies. Additional features for counting task: correct and incorrect counts and their ratio.	• Total speech time (*P*<.01) • Total silence time (*P*<.01) • % silence • Speech rate (*P*<.0001) • F0 coefficient of variation • Alpha ratio	• Net speech rate (*P*<.0001) • Pause ratio • Vowel keeping time (VKT) and MaxSylRep • COV1[h] (*P*<.0001) • COV2[i] • IntDur: interval duration (ms) • PatiRatio[r]	• Net speech rate • Pause ratio • VKT and MaxSylRep • COV1[h] • COV2[i] • IntDur: interval duration (ms; *P*<.001) • PatiRatio[r] (*P*<.001)
Model	Linear regression	Elastic net	Elastic net	Random forest	—[j]	—	—
MAE[k], mean (SD)							
cUHDRS[l]	2.1 (0.3)	2.4 (0.4)	2.78 (0.5)	NA[m]	—[j]	—	—

XSL•FO
RenderX

Objective of study	Prediction model				Speech markers of disease		
Authors	This study	Riad et al [5]	Riad et al [13]	Nunes et al [4]	Vogel et al [14]	Skodda et al [15]	Saft et al [16]
TMS[n]	10.1 (1.1)	12 (1.8)	13.14 (1.7)	9.64	—	—	—
TFC[o]	1.3 (0.2)	1.3 (0.2)	1.64 (0.3)	2.43	—	—	—
SW[p]	11.6 (1.7)	13.3 (2.2)	NA	NA	—	—	—
SDMT[q]	8.4 (1.3)	8.9 (1.8)	NA	NA	—	—	—

[a]HD-ISS: Huntington Disease Integrated Staging System.
[b]HD: Huntington Disease.
[c]PreHD: premanifest stage.
[d]NR: not reported.
[e]ASR: automatic speech recognition.
[f]CAP: CAG-age-product.
[g]CAG: cytosine-adenine-guanine.
[h]COV1: relative coefficient of variation for single syllable repetition.
[i]COV2: relative coefficient of variation of the pairs of syllables.
[j]Not applicable.
[k]MAE: mean absolute error.
[l]cUHDRS: composite Unified Huntington Disease Rating Scale.
[m]Not available.
[n]TMS: Total Motor Score.
[o]TFC: Total Functional Capacity.
[p]SW: Stroop Word.
[q]SDMT: Symbol Digit Modalities Test.
[r]PatiRatio: average interval duration of syllables (pa) divided by average interval duration of syllable (ti)

Model Equations

The regression coefficients of the xHD-Vox model, calibrated on the training set, are provided in Table 5.

XSL•FO
RenderX

Table . Regression coefficients of the xHD-Vox model: a standard linear regression model predicting clinical scores. All speech features were extracted from the backward counting task. Variables were not normalized before regression for reproducibility purposes. Exact *P* values are reported when *P*≥.001. As an example, the regression equation for predicting cUHDRS[a] is: cUHDRS=33.05−0.12 × CAP[b] score + 2.5 × numbers pronounced per second − 0.27 × CAG[c] repetition − 2.32 × SD of numbers pronounced per second. Definition of speech features: task duration = time between the first and last spoken numbers (from 20 to 1). Numbers pronounced per second = total number of numbers spoken divided by task duration (including perseverations). SD of numbers pronounced per second = variability in duration across individual spoken numbers.

Feature	cUHDRS[a]	TMS[d]	TFC[e]	Stroop Word	SDMT[f]
Intercept	33.05[g] (*P*≤.001)	−78.10[g] (*P*≤.001)	19.52[g] (*P*≤.001)	164.87[g] (*P*≤.001)	97.33[g] (*P*≤.001)
CAP[b] score	−0.12[g] (*P*≤.001)	0.45[g] (*P*≤.001)	−0.04[g] (*P*≤.001)	−0.56[g] (*P*≤.001)	−0.42[g] (*P*≤.001)
Numbers pronounced per second	2.50[g] (*P*≤.001)	−8.40[g] (*P*≤.001)	0.94[g] (*P*≤.001)	14.3[g] (*P*≤.001)	8.25[g] (*P*≤.001)
CAG[c] repetition	−0.27[g] (*P*≤.001)	1.44[g] (*P*≤.001)	−0.11[h] (*P*=.009)	−1.15[g] (*P*≤.001)	−0.66[i] (*P*=.02)
SD of numbers pronounced per second	−2.32[g] (*P*≤.001)	10.2[g] (*P*≤.001)	−1.27[h] (*P*=.002)	−10.54[h] (*P*=.001)	−5.06[i] (*P*=.046)

[a]cUHDRS: composite Unified Huntington Disease Rating Scale.

[b]CAP: CAG-age-product.

[c]CAG: cytosine-adenine-guanine.

[d]TMS: Total Motor Score.

[e]TFC: Total Functional Capacity.

[f]SDMT: Symbol Digit Modalities Test.

[g]*P*≤.001.

[h]*P*≤.01.

[i]*P*<.05.

Model Performance on the Test Set

The R^2 values of xHD-Vox on the test set were 0.57 for cUHDRS and TMS, 0.49 for SW, and 0.48 for SDMT (Figure 2). The lowest R^2 was observed for TFC (0.26).

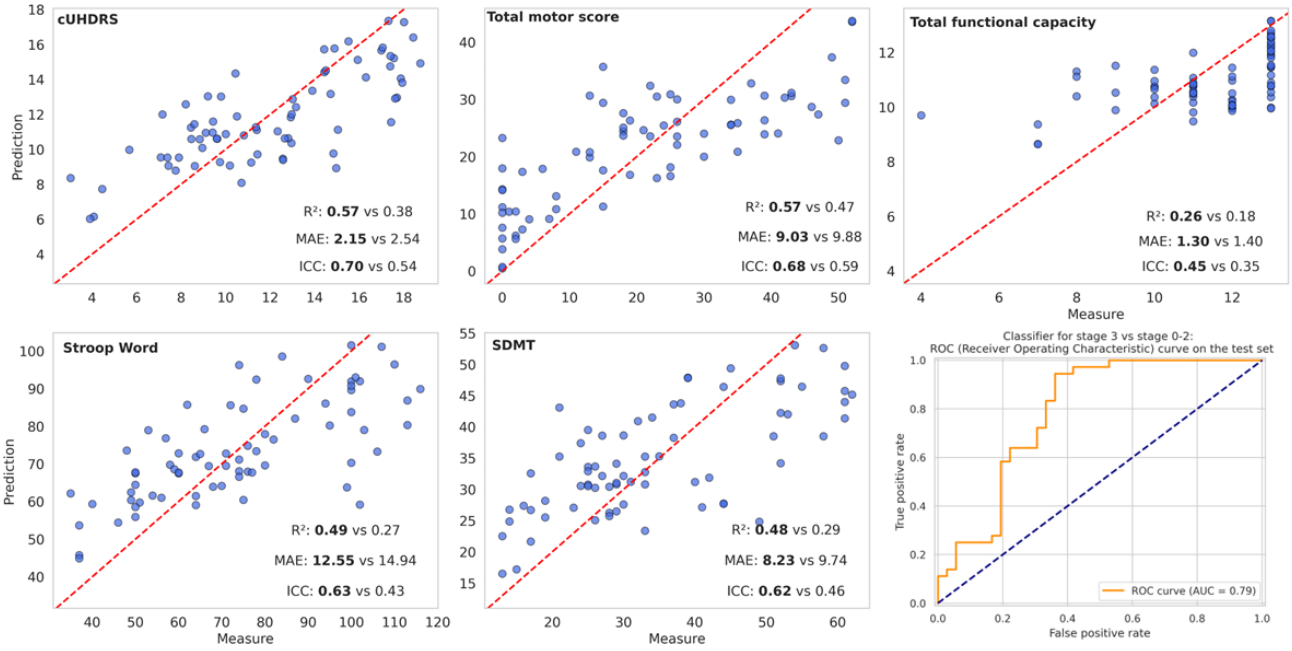
Compared with a demographic baseline model including CAP score and CAG repeat length, xHD-Vox improved cUHDRS prediction, increasing R^2 by 50% and reducing MAE by 15%. The largest gains were observed for cognitive outcomes, with relative R^2 increases of 81% for SW and 66% for SDMT, along with corresponding MAE reductions of 16% and 13%, respectively.

Based on the model's cross-validation results in Table 3 (last column), the test set MAE for cUHDRS matched expectations (2.1), equal to the average MAE across the 50 folds. For Symbol Digit Modality and TMS, the test-set MAEs were lower than expected (8.2 vs 8.4 and 9.0 vs 10.1, respectively). In contrast, SW showed poorer test-set performance, with a higher MAE than during cross-validation (12.5 vs 11.6).

Mean and median absolute errors per visit remained stable over the 2-year follow-up (Figure S4 in Multimedia Appendix 1). MAE varied by HD-ISS stage (Figure S5 in Multimedia Appendix 1). For example, for cUHDRS, the MAE decreased from 2.6 at stage 0 - 1 to 1.7 at stage 2 before increasing to 2.2 at stage 3.

XSL•FO
RenderX

Figure 2. Model performance on the testing set (72 visits, 24 participants). Predicted clinical scores vs measured values (blue dots). The red line represents perfect predictions. The mean absolute error (MAE) corresponds to the vertical distance between a point and this line. The explained variance (R^2), MAE, and intraclass correlation coefficient (ICC) of our model (bold) vs those obtained with a demographic baseline (cytosine-adenine-guanine (CAG)-age-product [CAP] score + CAG repetition). The bottom-right panel displays the Receiver Operating Characteristic curve for a logistic classifier distinguishing Huntington Disease Integrated Staging System stage 3 from stages 0 - 2, based on CAP score and speech rate. Refer to Figure S3 in Multimedia Appendix 1 for comparison of the classifier's performance compared to CAP score alone. cUHDRS: composite Unified Huntington Disease Rating Scale; HD-ISS: Huntington Disease Integrated Staging System; SDMT: Symbol Digit Modalities Test.



Functional Classification

Model performance for TFC prediction was modest (R^2=0.26), likely reflecting the limited variance in TFC scores (mean 11.1, SD 2.4). However, speech rate was significantly lower in participants at HD-ISS stage 3 compared to earlier stages (Figure S6 in Multimedia Appendix 1; 1-way ANOVA followed by post hoc Tukey test; P<.001).

We therefore investigated whether speech rate could serve as a marker of functional impairment and trained a logistic regression classifier using CAP score and speech rate as the 2 predictors (methods and Figure S3 in Multimedia Appendix 1). The model distinguished stage 3 from stages 0 - 2 with good performance on the held-out test set: recall=0.83, accuracy=0.73, and ROC AUC=0.79 (bottom right panel in Figure 2). For comparison, a model based only on CAP score achieved recall=0.68, accuracy=0.68, and ROC AUC=0.74.

Longitudinal Change

We conducted repeated-measures 2-way ANOVAs with time (Y0, Y1, and Y2) and type of measure (clinical vs xHD-Vox) as within-subject factors for all scores (Table 6). A significant main effect of time was observed for all scores (P<.001 for all scores), with no significant time × type interaction, indicating similar temporal trajectories for clinical measures and xHD-Vox predictions. Post hoc Tukey tests revealed significant differences between Y0 and Y2 for all scores, and between all time points (Y0, Y1, and Y2) for cUHDRS and SW. Mean scores decreased consistently across time points for both clinical and xHD-Vox measures, demonstrating a statistically significant longitudinal decline captured by both measures. No significant main effect of type of measure was observed. Detailed ANOVA results for the main effect of type and the time × type interaction are provided in Table S7 in the Multimedia Appendix 1. These results demonstrate that xHD-Vox reliably captures longitudinal clinical decline over a 2-year follow-up.

XSL•FO
RenderX

Table . Mean clinical and xHD-Vox scores over time and results of repeated-measures 2-way ANOVA for the main effect of time. Repeated-measures 2-way ANOVA was conducted for each clinical score, with factors time (Y0, Y1, and Y2) and type of measure (clinical vs xHD-Vox). All ANOVAs were corrected for multiple comparisons using the Bonferroni method. Results for the main effect of time are reported here. Post hoc Tukey tests were performed for pairwise comparisons when the time effect was significant.

Clinical score	Clinical measures			xHD-Vox measure			Time effect (repeated-measured ANOVA)			
	Y0, mean (SD)	Y1, mean (SD)	Y2, mean (SD)	Y0, mean (SD)	Y1, mean (SD)	Y2, mean (SD)	*F*(1, 2)	*P* value	np²[a]	Post hoc (Tukey, Bonferroni corrected)
cUHDRS[b]	12.75 (3.65)	11.99 (4.02)	11.29 (4.41)	12.17 (2.49)	11.73 (2.83)	11.51 (2.57)	17.83 (2, 46)	<.001	0.02	Y1 vs Y2, Y1 vs Y0, and 2 vs Y0
UHDRS-TMS[c]	19.25 (15.11)	21.29 (16.72)	23.33 (18.10)	20.06 (9.23)	21.74 (10.50)	22.61 (9.58)	8.76 (2, 46)	.001	0.01	Y1 vs Y2 and Y2 vs Y0
UHDRS-TFC[d]	11.88 (1.48)	11.38 (1.88)	11.04 (2.30)	11.15 (0.97)	10.97 (1.12)	10.88 (1.02)	8.85 (2, 46)	.001	0.02	Y1 vs Y0 and Y2 vs Y0
SW[e]	77.83 (22.00)	73.75 (21.30)	69.04 (22.10)	75.43 (12.56)	73.26 (14.40)	72.21 (13.05)	14.23 (2, 46)	<.001	0.02	Y1 vs Y2, Y1 vs Y0, and Y2 vs Y0
SDMT[f]	36.21 (13.35)	34.42 (14.04)	32.71 (14.80)	35.94 (8.30)	34.54 (9.30)	33.76 (8.50)	9.94 (2, 46)	<.001	0.01	Y2 vs Y0

[a]ηp²: partial eta squared.

[b]cUHDRS: composite Unified Huntington Disease Rating Scale.

[c]UHDRS-TMS: Unified HuntingtonRating Scale Disease Rating Scale Disease Rating Scale Total Motor Score.

[d]UHDRS-TFC: Unified HuntingtonRating Scale Disease Rating Scale Total Functional Capacity.

[e]SW: Stroop Word.

[f]SDMT: Symbol Digit Modalities Test.

The magnitude of the 1-year and 2-year decline at the group level is reported in Table 7, which presents mean changes and their 95% CIs. Corresponding plots are shown in Figure S7 in Multimedia Appendix 1 for illustrative purposes. For all clinical scores, the mean 2-year decline measured by clinicians was approximately twice the 1-year decline, indicating a consistent and approximately linear progression over time. Predicted changes derived from xHD-Vox fell within the 95% CIs of the corresponding clinician-assessed changes for all scores except the 2-year change in SW. These results suggest that xHD-Vox provides a group-level estimate of the magnitude of longitudinal clinical decline.

XSL•FO
RenderX

Table . One-year and 2-year mean changes in clinical scores for model predictions and clinician assessments on the test set (N=24). Predicted mean changes are compared with the clinician-derived 95% CIs to evaluate alignment between model outputs and expert ratings. For example, the predicted 1-year change in cUHDRS[a] (−0.44) falls within the 95% CI of the clinician assessments (−1.3 to −0.2).

Clinical score	1-year change, mean (95% CI)		2-year change, mean (95% CI)	
	Clinician	Predicted	Clinician	Predicted
cUHDRS[a]	−0.8 (−1.3 to −0.2)	−0.44 (−0.9 to −0.0)	−1.5 (−2.3 to −0.6)	−0.7 (−1.1 to −0.3)
TMS[b]	2.0 (−1.2 to 5.3)	1.7 (0.2 to 3.2)	4 (0.1 to 8.0)	2.6 (1.1 to 4.0)
TFC[c]	−0.5 (−0.9 to −0.1)	−0.2 (−0.4 to 0)	−0.8 (−1.5 to −0.1)	−0.3 (−0.4 to −0.1)
SW[d]	−4.1 (−7.4 to −0.8)	−2.2 (−4.4 to 0.1)	−8.8 (−13.6 to −3.9)	−3.2 (−5.4 to −1.1)
SDMT[e]	−1.8 (−3.8 to 0.2)	−1.4 (−2.6 to −0.2)	−3.5 (−6.2 to −0.8)	−2.2 (−3.4 to −1)

[a]cUHDRS: composite Unified Huntington Disease Rating Scale.

[b]TMS: Total Motor Score.

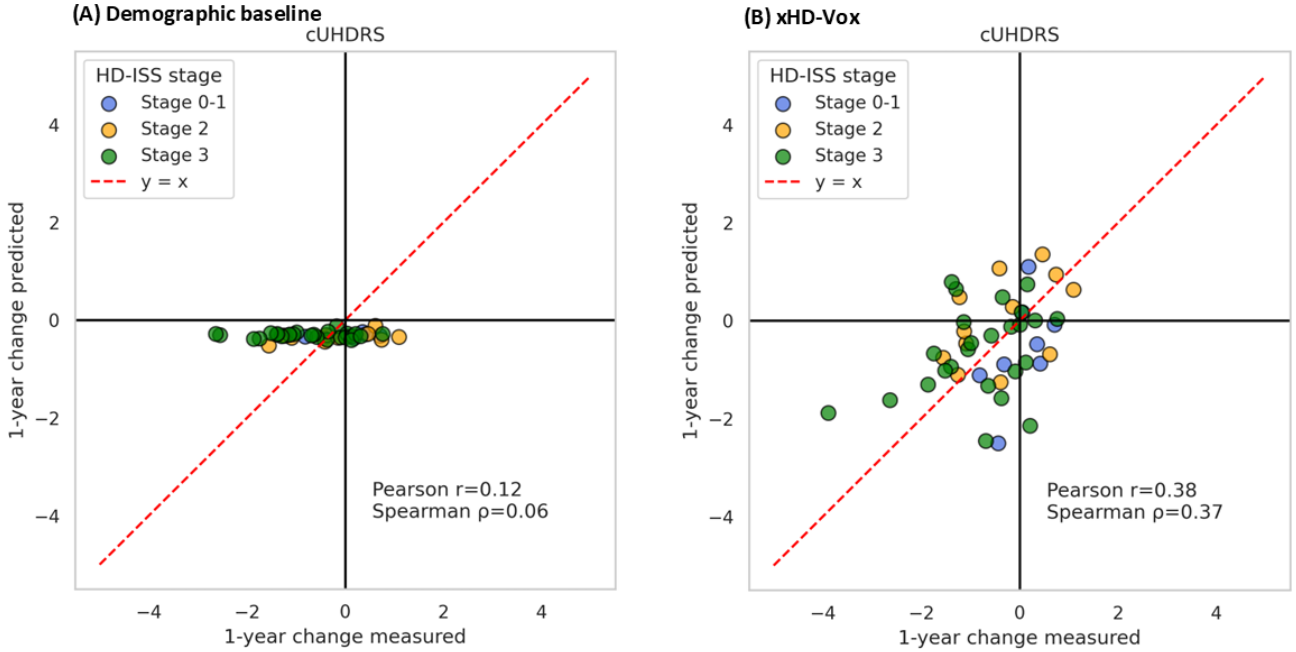
[c]TFC: Total Functional Capacity.

[d]SW: Stroop Word.

[e]SDMT: Symbol Digit Modalities Test.

Figure 3 compares predicted 1-year changes in cUHDRS with clinically measured changes at the individual level for both xHD-Vox and the demographic baseline model (45 observations; 3 influential outliers removed). A sensitivity analysis including all 48 observations is reported in Figure S8 in Multimedia Appendix 1 and discussed in the following paragraphs. The demographic baseline captured little intraindividual variability, with Pearson and Spearman correlations of 0.12 and 0.06, respectively, and produced identical slopes across participants.

Figure 3. One-year change in composite Unified Huntington Disease Rating Scale (cUHDRS) predicted vs clinically measured for xHD-Vox (panel B) and the demographic baseline model (panel A). The red line represents the line of perfect agreement (y = x). Pearson and Spearman correlation coefficients are reported. The main analysis includes 45 observations, with 3 influential outliers removed; a sensitivity analysis including all 48 observations is shown in Figure S8 in Multimedia Appendix 1. HD-ISS: Huntington Disease Integrated Staging System.



In contrast, xHD-Vox exhibited greater interindividual variability, with Pearson and Spearman correlations of 0.38 and 0.37, respectively. The model also captured positive changes in some participants at earlier disease stages (HD-ISS 0 - 2), as illustrated in Figure 3 and in the individual trajectories shown in Figure S9 in Multimedia Appendix 1.

A sensitivity analysis including all 48 observations (24 participants with 2 annual change measures each) showed that correlation estimates were strongly influenced by 3 influential outliers. When all data points were included, the Pearson correlation was close to 0, whereas the Spearman correlation remained higher (ρ=0.17). The discrepancy between Pearson and Spearman correlations suggested a monotonic relationship that was disproportionately affected by a small number of extreme values, as identified in Figure S8 in Multimedia Appendix 1.

XSL•FO
RenderX

Discussion

This study presented an interpretable, automated speech-based model, xHD-Vox, for monitoring clinical progression in HD. Building on Riad et al [5] prior work using counting tasks, we reduced the feature set from 63 to 4 key predictors while maintaining comparable MAE: CAP score, CAG repeat length, and rhythm-related measures—speech rate (numbers pronounced per second) and its SD. These 2 rhythm features were automatically derived from Whisper's transcriptions and timestamps, an open-source ASR model [8].

xHD-Vox was trained on data from 157 participants and evaluated on a longitudinal cohort of 24 participants monitored annually during 2 years. On this test set, xHD-Vox explained 57% of the variance in cUHDRS, representing a 50% improvement over the demographic baseline. The MAE was 2.2, with an ICC of 0.70—lower than the 0.92 reported for clinicians [17], but promising for a fully automated tool. Similar results were obtained for TMS (MAE=9.0; ICC=0.68 vs ICC=0.84 for clinicians [18]).

Cognitive outcomes also showed significant improvement relative to the demographic baseline, with an R^2 increase of 66% for SDMT, reaching 0.48. To assess clinical relevance, we compared our MAE values with normative benchmarks from Mills et al [19]. The model's MAE for SW (12.5) approximated the difference between the 50th and 5th percentiles (11) for highly educated 50-year-olds, a demographic representative of our cohort. Similarly, the MAE for SDMT (8.2) closely matched the normative spread (8).

Although the linear regression model yielded a relatively low R^2 for TFC (0.26), a logistic regression using speech rate and CAP score effectively classified HD-ISS stage 0 - 2 vs stage 3 with an ROC AUC of 0.79, supporting its potential utility in detecting functional decline.

Longitudinal analyses using repeated-measures ANOVAs with post hoc Tukey tests confirmed a significant decline over the 2-year follow-up in the test set for both clinician-assessed measures and xHD-Vox predictions. At the group level, the mean 1-year and 2-year changes predicted by xHD-Vox were consistent with clinically measured changes, falling within the corresponding 95% CIs. Exploratory individual-level analyses showed that xHD-Vox reasonably captured longitudinal 1-year change compared with a demographic baseline model, although results were sensitive to a small number of influential outliers. Collectively, these findings support the potential of xHD-Vox to model longitudinal clinical trajectories, while underscoring the need for larger and more diverse test samples to further assess robustness at the individual level.

Compared to prior speech studies in HD (summarized in Table 4), our model stands out for its larger cohort (n=181 gene carriers), improved performance, and longitudinal scope. In terms of performance, the MAE here was reduced for SW and TMS by 13% and 16%, respectively, relative to Riad et al [5]. For TFC, a 46% improvement was achieved compared with Nunes et al [4], although this should be interpreted cautiously, considering the difference in sample size in cross-validation (n=25 in their study). The only published longitudinal study did not model clinical outcomes but focused on the decline of speech biomarkers over time in premanifest participants (n=13) [16], which precludes direct comparison [14,16].

Unlike most previous studies, we relied on a backward counting task, a task widely conceptualized as a working memory task. Individual differences in working memory capacity are known to covary with processing speed and broader cognitive efficiency [20,21], which are core determinants of SDMT and SW performance. In addition, backward counting and the SW task both rely on inhibitory control processes: inhibition of the overlearned forward number sequence in backward counting and inhibition of the automatic reading response in the Stroop task. Thus, rhythm-related speech features likely capture these shared cognitive efficiencies, explaining their predictive value for broader cognitive performance beyond working memory alone. Speech rate was also demonstrated to be a marker of disease progression in HD [14,15]. Consistent with Vogel et al [14], we observed a reduction in speech rate in stage 3 participants compared with controls and earlier-stage patients (Figure S6 in Multimedia Appendix 1). We did not detect significant differences between stages 0 - 1 and controls in our measure of speech rate, in agreement with Vogel et al [14] but in contrast to findings reported by Skodda et al [15] (refer to Table 4 for sample sizes in each group and study). Although speech rate in stage 2 participants was lower than in stages 0 - 1, this difference was not statistically significant in our cohort (P=.07), whereas it was significant in Vogel et al [14]. Differences across studies may reflect variability in sample size, statistical approaches, and, importantly, the experimental task used. Indeed, Vogel et al [14] highlighted the task-dependent nature of speech rate as a biomarker, showing weak correlations between speech rate and disease burden score in a days-of-the-week recitation task (R=−0.22), but stronger correlations for reading (R=−0.68). These considerations underscore the importance of task selection when using speech rate as a cognitive and disease-related marker.

Our model can be implemented in mobile apps for remote monitoring. We provided all model equations and code on GitHub, ensured clinical interpretability through identified features, and automated feature extraction using the open-source algorithm Whisper. Whisper produced accurate transcriptions and sequence-level timestamps, resulting in high agreement with speech pathologists' annotations for speech rate. Agreement was lower for the SD of numbers pronounced, which relies on word-level timestamps, but this limitation does not substantially compromise overall model performance. Here, the use of Whisper addressed reproducibility concerns raised by a previous study [4] that relied on proprietary ASR systems or performance issues of acoustic-based automated speech analysis [13]. Despite being trained on general speech, Whisper has demonstrated strong performance for dysarthric speech, including in alpha-synucleopathies (Pearson correlations ranging from 0.6 to 0.94 depending on linguistic features) [22]. Fine-tuning Whisper for dysarthric speech (eg, CrisperWhisper [23]) or leveraging its audio encoder opens opportunities for other predictive approaches [24].

XSL•FO
RenderX

XHD-Vox opens the door for several applications. First, it could be used in underserved areas. About 40 % of individuals with Parkinson disease currently do not receive care from a specialist [25]. Given that HD is a rare disease, we anticipate that the situation may be similar or even worse. Like Parkinson disease, HD specialists are concentrated in urban medical centers, while patients are geographically dispersed, making travel to these centers more difficult due to disability and age. By offering a simple and ecological approach for tracking disease progression, the need for clinic visits may decrease, whereas increasing the number of people benefiting from appropriate monitoring. Second, by allowing for frequent measurements of disease severity, xHD-Vox could assist clinicians in following patients, especially during crises such as COVID-19, when patients could not attend their annual visits. Third, our approach could be used for screening patients' disease severity for clinical trials, especially between stages 0 - 2 and stage 3, using our classifier.

However, xHD-Vox has not yet been tested in other languages. Fahed et al [26] conducted the only multilingual study (English, Spanish, and Polish) on acoustic biomarkers in HD, focusing on reading tasks, syllable repetition, and sustained vowels. They showed that all temporal features are language-dependent in reading tasks. In Parkinson disease, the articulation rate also differed across languages [27]. With Whisper supporting 96 languages, adapting xHD-Vox to a new language would require fewer than 100 participants to achieve 95% accuracy (Figure S2 in Multimedia Appendix 1). While we have tested our model on an unseen sample (test set) with good results, replication in cohorts from other centers would strengthen generalizability. Additionally, testing real-world data collection by recording patients on their mobile phones, as done for Parkinson disease [28], would validate the approach for remote monitoring. Building on this framework, we have developed an Android prototype of xHD-Vox, currently under internal testing, to explore its potential for mobile deployment and remote monitoring.

In conclusion, we provide novel insights into the robustness of an automated speech-based clinical estimation model for longitudinal monitoring of HD gene carriers, xHD-Vox. Its strengths lie in its automation and interpretability, facilitating deployment on mobile platforms and acceptance by clinicians. This work represents a step forward in telemedicine for HD monitoring, with the potential to detect disease progression, particularly in regions with limited medical access or during periods of health care disruption, such as lockdowns.

Acknowledgments

The authors acknowledge the patients and families whose cooperation was vital to generate this study. They also acknowledge Tiffany Monnier and Audrey Fabre for ensuring the quality of clinical data reported and annotations of speech recordings, respectively. The author used ChatGPT-4 (OpenAI) to refine the grammar and scientific language of the paper.

Funding

MIG-HD was funded from the "Direction de la Recherche Clinique" (Assistance Publique- Hôpitaux de Paris) by AOM00139 and AOM04021 grants, REPAIR-HD from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°602245 , BIO-HD by the Henri-Mondor Hospital National Reference Centre for Huntington's disease (Ministry of Health). The team is supported by NeurATRIS (ANR-11-INBS-0011) and ANR-17-EURE-0017, both funded by the Agence Nationale de la Recherche (French National Research Agency).

Data Availability

The anonymized patient-level datasets generated during this study are available from the corresponding author upon reasonable request. The code used for model validation is openly available on GitHub (repository Tiphainell/xHD-Vox) and archived on Zenodo [29].

Authors' Contributions

TLL contributed to drafting and revision of the paper for content, study concept and design, analysis, and interpretation of data. AG-S contributed to drafting and revision of the paper for content. HT contributed to the development of a feature extraction pipeline and had a major role in the acquisition of data. RL contributed to the revision of the paper for content. CLMV contributed to the revision of the paper for content. RM contributed to drafting and revision of the paper for content and study concept or design. ACB-L contributed to drafting and revision of the paper for content, study concept or design, major role in the acquisition of data, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional figures, tables, and methods.
[DOCX File, 1245 KB - neuro_v5i1e83838_app1.docx]

XSL•FO
RenderX

Checklist 1
TRIPOD+AI checklist.
[PDF File, 712 KB - neuro_v5i1e83838_app2.pdf]

References

1. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 1993 Mar 26;72(6):971-983. [doi: 10.1016/0092-8674(93)90585-e] [Medline: 8458085]
2. Schobel SA, Palermo G, Auinger P, et al. Motor, cognitive, and functional declines contribute to a single progressive factor in early HD. Neurology (ECronicon) 2017 Dec 12;89(24):2495-2502. [doi: 10.1212/WNL.0000000000004743] [Medline: 29142089]
3. Pham Nguyen TP, Bravo L, Gonzalez-Alegre P, Willis AW. Geographic barriers drive disparities in specialty center access for older adults with Huntington's disease. J Huntingtons Dis 2022;11(1):81-89. [doi: 10.3233/JHD-210489] [Medline: 35253771]
4. Nunes AS, Pawlik M, Mishra RK, et al. Digital assessment of speech in Huntington disease. Front Neurol 2024;15:1310548. [doi: 10.3389/fneur.2024.1310548] [Medline: 38322583]
5. Riad R, Lunven M, Titeux H, et al. Predicting clinical scores in Huntington's disease: a lightweight speech test. J Neurol 2022 Sep;269(9):5008-5021. [doi: 10.1007/s00415-022-11148-1] [Medline: 35567614]
6. Bartolomeu Pires S, Kunkel D, Goodwin N, et al. Are people living with Huntington's disease experiencing person-centered integrated care? J Huntingtons Dis 2024 Nov;13(4):535-546. [doi: 10.1177/18796397241288449] [Medline: 39973375]
7. Willock R, Rickards H, Rosser AE, et al. An overview of specialist services for Huntington's disease in the United Kingdom. J Huntingtons Dis 2023;12(4):363-370. [doi: 10.3233/JHD-220560] [Medline: 38108355]
8. Radford A, Kim JW, Xu T, Brockman G, McLeavey C. Robust speech recognition via large-scale weak supervision. Presented at: 40th International Conference on Machine Learning; Jul 23-29, 2023 URL: https://proceedings.mlr.press/v202/radford23a.html [accessed 2026-06-12]
9. Tabrizi SJ, Schobel S, Gantman EC, et al. A biological classification of Huntington's disease: the Integrated Staging System. Lancet Neurol 2022 Jul;21(7):632-644. [doi: 10.1016/S1474-4422(22)00120-X] [Medline: 35716693]
10. Warner JH, Long JD, Mills JA, et al. Standardizing the CAP score in Huntington's disease by predicting age-at-onset. J Huntingtons Dis 2022;11(2):153-171. [doi: 10.3233/JHD-210475] [Medline: 35466943]
11. Titeux H, Rachid R, Xuan-Nga C, et al. Seshat: a tool for managing and verifying annotation campaigns of audio data. Presented at: Twelfth Language Resources and Evaluation Conference; May 11-16, 2020 URL: https://aclanthology.org/2020.lrec-1.861/ [accessed 2026-06-12]
12. Pedregosa F, Varoquaux G, Gramfort A, Michel V. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
13. Riad R, Titeux H, Lemoine L, et al. Vocal markers from sustained phonation in Huntington's disease. Presented at: Interspeech 2020; Oct 25-29, 2020 p. 1893-1897 URL: https://www.isca-archive.org/interspeech_2020 [doi: 10.21437/Interspeech.2020-1057]
14. Vogel AP, Shirbin C, Churchyard AJ, Stout JC. Speech acoustic markers of early stage and prodromal Huntington's disease: a marker of disease onset? Neuropsychologia 2012 Dec;50(14):3273-3278. [doi: 10.1016/j.neuropsychologia.2012.09.011] [Medline: 22982606]
15. Skodda S, Grönheit W, Lukas C, et al. Two different phenomena in basic motor speech performance in premanifest Huntington disease. Neurology (ECronicon) 2016 Apr 5;86(14):1329-1335. [doi: 10.1212/WNL.0000000000002550] [Medline: 26962067]
16. Saft C, Jessen J, Hoffmann R, Lukas C, Skodda S. Speech biomarkers in Huntington's Disease: a longitudinal follow-up study in premanifest mutation carriers. J Huntingtons Dis 2024;13(3):369-373. [doi: 10.3233/JHD-240021] [Medline: 38995795]
17. Trundell D, Palermo G, Schobel S, et al. F23 validity, reliability, ability to detect change and meaningful within-patient change of the CUHDRS. J Neurol Neurosurg Psychiatry 2018;89. [doi: 10.1136/jnnp-2018-EHDN.127]
18. Reilmann R, Schubert R. Motor outcome measures in Huntington disease clinical trials. Handb Clin Neurol 2017;144:209-225. [doi: 10.1016/B978-0-12-801893-4.00018-3] [Medline: 28947119]
19. Mills JA, Long JD, Mohan A, Ware JJ, Sampaio C. Cognitive and motor norms for Huntington's disease. Arch Clin Neuropsychol 2020 Aug 28;35(6):671-682. [doi: 10.1093/arclin/acaa026] [Medline: 32407458]
20. Fry AF, Hale S. Relationships among processing speed, working memory, and fluid intelligence in children. Biol Psychol 2000 Oct;54(1-3):1-34. [doi: 10.1016/s0301-0511(00)00051-x] [Medline: 11035218]
21. Brébion G, David AS, Jones HM, Pilowsky LS. Working memory span and motor and cognitive speed in schizophrenia. Cogn Behav Neurol 2009 Jun;22(2):101-108. [doi: 10.1097/WNN.0b013e3181a722a0] [Medline: 19506426]
22. Šubert M, Tykalová T, Novotný M, Dušek P, Klempíř J, Rusz J. Automated analysis of spoken language differentiates multiple system atrophy from Parkinson's disease. J Neurol 2025 Jan 15;272(2):113. [doi: 10.1007/s00415-024-12828-w] [Medline: 39812820]

XSL•FO
RenderX

23. Zusag M, Wagner L, Thallinger B. CrisperWhisper: accurate timestamps on verbatim speech transcriptions. Presented at: Interspeech 2024; Sep 1-5, 2024. [doi: 10.21437/Interspeech.2024-731]
24. Riad R, Denais M, de Gennes M, et al. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: algorithm development and validation study. Psychiatry and Clinical Psychology. Preprint posted online on 2024. [doi: 10.1101/2024.03.20.24304577]
25. Willis AW, Schootman M, Evanoff BA, Perlmutter JS, Racette BA. Neurologist care in Parkinson disease: a utilization, outcomes, and survival study. Neurology (ECronicon) 2011 Aug 30;77(9):851-857. [doi: 10.1212/WNL.0b013e31822c9123] [Medline: 21832214]
26. Fahed VS, Doheny EP, Collazo C, et al. Language-independent acoustic biomarkers for quantifying speech impairment in Huntington's disease. Am J Speech Lang Pathol 2024 May;33(3):1390-1405. [doi: 10.1044/2024_AJSLP-23-00175] [Medline: 38530396]
27. Rusz J, Hlavnička J, Novotný M, et al. Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. Ann Neurol 2021 Jul;90(1):62-75. [doi: 10.1002/ana.26085] [Medline: 33856074]
28. Tsanas A, Little MA, Ramig LO. Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network. IEEE Access 2021;9:11024-11036. [doi: 10.1109/ACCESS.2021.3050524] [Medline: 33495722]
29. Le Ludec T. Code for: development and validation of xHD-Vox: an automated speech-based model predicting motor and cognitive scores in Huntington's disease. Zenodo. 2025. URL: https://doi.org/10.5281/zenodo.17084076 [accessed 2026-06-12]

Abbreviations

ASR : automatic speech recognition
CAG: cytosine-adenine-guanine
CAP: CAG-age-product
HD: Huntington disease
HD-ISS : Huntington Disease Integrated Staging System
ICC: intraclass correlation coefficient
MAE: mean absolute error
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROC AUC: Receiver Operating Characteristic–Area Under the Curve
SDMT : Symbol Digit Modalities Test
SW: Stroop Word
TFC: Total Functional Capacity
TMS: Total Motor Score
TRIPOD+AI: Transparent Report of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UHDRS: Unified Huntington Disease Rating Scale
VIF: variance inflation factor

Edited by A Zampogna; submitted 09.Sep.2025; peer-reviewed by J Rusz, M Lowery; revised version received 29.Jan.2026; accepted 23.Feb.2026; published 08.Jul.2026.

Please cite as:
Le Ludec T, Gil-Salcedo A, Titeux H, Louiset R, Le Moine Veillon C, Massart R, Bachoud-Lévi AC
xHD-Vox, an Automated Speech Model for Estimating Motor and Cognitive Scores in Huntington Disease: Development and Longitudinal Validation
JMIR Neurotech 2026;5:e83838
URL: https://neuro.jmir.org/2026/1/e83838
doi:10.2196/83838

©Tiphaine Le Ludec, Andres Gil-Salcedo, Hadrien Titeux, Robin Louiset, Clément Le Moine Veillon, Renaud Massart, Anne-Catherine Bachoud-Lévi. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 8.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Self-Guided Beta-Responsive Visual Neurofeedback in a Public Exhibition: Feasibility and Usability Randomized Controlled Trial

Melanie Baumgartner, MSc; Raphael Weibel, PhD; Tobias Hoesli, MA; Florian von Wangenheim, Prof Dr; Joseph Ollier, PhD

Mobiliar Lab for Analytics, Department of Management, Technology, and Economics, ETH Zurich, Weinbergstrasse 58, Zurich, Switzerland

Corresponding Author:
Melanie Baumgartner, MSc
Mobiliar Lab for Analytics, Department of Management, Technology, and Economics, ETH Zurich, Weinbergstrasse 58, Zurich, Switzerland

Abstract

Background: Neurofeedback is studied for stress and anxiety regulation, but most systems require supervision or research-grade equipment. Consumer electroencephalography (EEG) could broaden access, yet autonomous first-contact feasibility, usability, and baseline neural-state effects remain unclear.

Objective: We assessed autonomous completion and usability, tested whether a fixed breathing pacer improved outcomes over visual neurofeedback alone, and explored whether pretask resting beta was associated with usability.

Methods: We conducted a single-session, 2-arm randomized feasibility and usability trial at a public science exhibition in Bern, Switzerland, between March and June 2025. Eligible adults were aged 18 years or older, could read German or French, and reported no epilepsy or implanted electronic device. A tablet app guided participants without experimenter assistance through consent, setup, a 30-second eyes-closed baseline, 1:1 randomization, a 120-second beta-responsive task, and questionnaires. The final analytic sample included participants assigned to visual neurofeedback only (VN; n=27) or visual neurofeedback plus a 6 breaths/minute breathing pacer (VN+BP; n=25). Outcomes were usability, relaxation, and state mindfulness. Data were collected prospectively; no secondary data were used. Analyses used Mann-Whitney *U*, chi-square, Spearman, and permutation tests (2-sided, α=.05); EEG outcome and item analyses were exploratory.

Results: Of 110 visitors initiating a session, 79 completed the task with a valid EEG recording (71.8%, 95% CI 62.8-79.4) and 52 met analytic criteria (47.3%, 95% CI 38.2%-56.5%). Participants (35 women) had a mean age of 39.0 (SD 11.3) years. Usability was good (System Usability Scale [SUS] mean 72.3, SD 14.7; 95% CI 68.3-76.2), and 31/52 participants scored at least 68 (59.6%, 95% CI 46.1%-71.8%). Usability did not differ between VN and VN+BP (median 75.0, IQR 62.5-81.3 vs median 72.5, IQR 65.0-90.0; Mann-Whitney *U*=362.5, *P*=.65; rank-biserial *r*=–0.07, 95% CI –0.39 to 0.25). Relaxation (median 63.0, IQR 50.0-80.0 vs median 61.0, IQR 50.0-67.0; Mann-Whitney *U*=276.5, *P*=.27; *r*=0.18, 95% CI –0.14 to 0.49) and mindfulness (median 2.4, IQR 2.1-2.8 vs median 2.5, IQR 2.0-3.2; Mann-Whitney *U*=346.5, *P*=.68; *r*=–0.07, 95% CI –0.39 to 0.26) did not differ. Higher resting beta was associated with lower SUS before correction (Spearman ρ=–0.31, 95% CI –0.54 to –0.04; raw *P*=.03), but not after Benjamini-Hochberg correction (adjusted *P*=.30). SUS item analyses showed a global pattern (permutation *P*=.02), strongest for functional integration (ρ=–0.43, 95% CI –0.63 to –0.18; Holm-adjusted *P*=.01).

Conclusions: This trial evaluated an autonomous consumer-EEG neurofeedback system integrating self-fitting, signal-quality checking, baseline calibration, adaptive feedback, randomization, and self-report without operator support. Many first-time users completed the autonomous workflow with acceptable-to-good usability; a fixed, open-loop breathing pacer did not improve usability, relaxation, or state mindfulness; and higher resting beta showed a preliminary association with lower perceived usability. These findings support autonomous consumer-EEG neurofeedback and motivate work on adaptive onboarding, validated EEG pipelines, and objective self-regulation outcomes.

(JMIR Neurotech 2026;5:e89256) doi:10.2196/89256

KEYWORDS

neurofeedback; brain-computer interfaces; electroencephalography; beta band activity; mobile health; usability; human-computer interaction

XSL•FO
RenderX

Introduction

Neurofeedback (NFB) trains individuals to change their own brain activity by transforming ongoing neural signals into immediate visual, auditory, or haptic feedback [1]. In classical self-regulation models, such closed-loop systems increase perceived control and self-efficacy because users receive immediate feedback on their internal states and experience that they are capable of voluntarily changing their states and activities [2]. From an embodied and interoceptive perspective, regulation emerges when neural processes, bodily signals, and environmental surroundings are aligned such that internal states become sufficiently explicit for the user to act on them [3]. In other words, effective self-regulation depends on individuals being able to monitor their physiological state effectively [4,5]. For autonomously delivered consumer-EEG (electroencephalography), where expert assistance is unavailable, usability is a prerequisite for successful intervention delivery: users must be able to fit the headset, establish acceptable signal quality, understand the feedback, and complete the task without expert support.

Extant research has shown that NFB can enhance attention, emotion regulation, and stress management in both clinical and nonclinical populations [6,7]. In particular, training to reduce fast neural activity has shown benefits for individuals affected by anxiety symptoms, tension, and depressive symptoms, conditions which are highly prevalent among the general public [8-11]. EEG-based protocols have been used adjunctively in posttraumatic stress disorder (PTSD) and stress-related conditions to stabilize arousal, and performance-oriented NFB has improved affect regulation and cognitive control in nonclinical users [12-14]. A typical target in NFB is oscillatory band modulation. Alpha rhythms (8 - 13 Hz) are typically associated with relaxed, inward-focused attention, whereas beta rhythms (13 - 30 Hz) are linked to cortical activation, arousal, vigilance, and cognitive effort [8,15-17]. Repeated training can shift these rhythms in desired directions and has been associated with gains in attention, affect regulation, and cognitive performance [14]. Of particular relevance for stress and anxiety disorders is excessive beta activity. Elevated beta power has been associated with hyperarousal and anxious states, and beta downtraining has been reported to reduce tension and improve emotional stability [18-20]. This makes beta-responsive NFB training a theoretically grounded target for supporting self-regulation under stress. Most NFB is still delivered in controlled laboratory or clinical settings with research-grade EEG and expert support [1,21]. Commercial and app-based systems offer a potentially more scalable mode of delivery. Platforms such as Myndlift (Myndlift, Inc) provide remote home NFB with therapist supervision [12], while consumer-grade headsets such as Muse make mobile deployment technically feasible [22-24]. However, these approaches commonly rely on repeated sessions, human guidance, or remote supervision, and evidence on the autonomous first-contact use of consumer-EEG NFB remains limited. This is important because technical availability alone does not ensure successful independent use: without expert support, users must fit the headset, establish acceptable signal quality, understand the feedback, and remain engaged throughout the task [5,25-27]. In addition, little is known about whether users' pretask neural state is associated with how they experience and evaluate an autonomous NFB system.

We therefore evaluated a fully autonomous, tablet-based, beta-responsive visual NFB system using a consumer EEG headband in a public first-contact setting. The app guided participants through headset fitting, signal-quality checks, baseline recording, an NFB task, and postsession questionnaires without experimenter assistance. Participants received either visual NFB alone or visual NFB with an additional fixed paced-breathing guide. Paced breathing at approximately 6 breaths/minute has been associated with increased heart rate variability, greater parasympathetic influence, and improved baroreflex function [28-30]. The randomized comparison therefore allowed us to test whether adding this open-loop breathing cue improved usability or subjective experience beyond visual NFB alone.

Perceived usability was assessed using the System Usability Scale (SUS) [31,32]. Drawing on physiological computing and neuroadaptive human-computer interaction, we additionally examined pretask neural activity as a potential contextual factor associated with how users experienced the system [5,26,27,33]. Resting beta activity was selected as an exploratory feature because tonic beta has been associated with activation, vigilance, and maintenance of the current cognitive state [15,17]. Examining its relationship with first-contact usability may generate hypotheses for future state-sensitive onboarding approaches, such as adapting pacing, instruction density, or feedback support to the user's initial state [25,33]

Accordingly, the study addressed three research questions:

1. To what extent can first-contact users successfully complete a fully autonomous NFB workflow in a public setting, and how usable do they perceive the system to be?
2. Does adding a fixed paced-breathing guide to visual NFB affect perceived usability, subjective relaxation, state mindfulness, or descriptive EEG band-power estimates?
3. Is pretask resting beta activity associated with perceived usability in this autonomous first-contact setting?

Methods

Trial Design

We conducted a single-session, 2-arm parallel-group randomized feasibility and usability trial with a 1:1 allocation ratio at a public science exhibition in Bern, Switzerland, between March and June 2025. The trial evaluated a fully autonomous, tablet-based, beta-responsive visual NFB system using a consumer EEG headband. The study took place in a semi-enclosed booth that provided a seated, low-distraction environment separated from the main visitor area. The setting was chosen to approximate first-contact, low-support use outside a controlled laboratory or clinical environment. All interactions with the system were designed to be autonomous: participants received instructions, provided consent, completed EEG setup, performed the NFB task, and completed questionnaires without experimenter assistance. No changes to the trial design,

XSL•FO
RenderX

eligibility criteria, interventions, or outcome measures were made after trial commencement.

Reporting Guidelines

This trial is reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) 2025 statement [34], the CONSORT extension for randomized pilot and feasibility trials [35], and the CONSORT-EHEALTH checklist for reports of randomized trials of web-based and mobile health interventions [36]. The intervention is described in accordance with the TIDieR (Template for Intervention Description and Replication) checklist and guide [37]. The completed checklists are provided in Checklist 1-4.

Patient and Public Involvement

Patients and members of the public were not involved in the design, conduct, analysis, or reporting of this study beyond their participation as research participants.

Participants

Participants were identified and recruited by convenience sampling from visitors attending a public science exhibition in Bern, Switzerland, between March and June 2025. Recruitment occurred through exhibition signage and on-site information describing an opportunity to try a NFB exercise. Participation was voluntary and unpaid. Data were collected in a semi-enclosed booth that provided a seated, low-distraction environment separated from the main exhibition area. Eligibility criteria required participants to be at least 18 years old, able to read and respond in German or French, able to provide informed consent, and to report no history of epilepsy or seizures and no implanted electronic medical device, such as a pacemaker or defibrillator. No prior experience with EEG, NFB, meditation, or related technologies was required.

Before entering the study, participants viewed an on-screen information sheet describing the study purpose, procedures, eligibility requirements, potential minor risks, voluntary participation, the right to withdraw, and data handling. They provided digital informed consent through the tablet app by confirming that they understood the study information, met the eligibility criteria, and agreed to participate. For the final analytic sample, records were excluded if participants reported alcohol consumption or stimulant medication use within 8 hours before the session. These exclusions were applied to limit acute physiological confounding and were treated as analytic exclusion criteria rather than safety-related eligibility criteria. Additional data-quality exclusions are described in the “Statistical Methods” section.

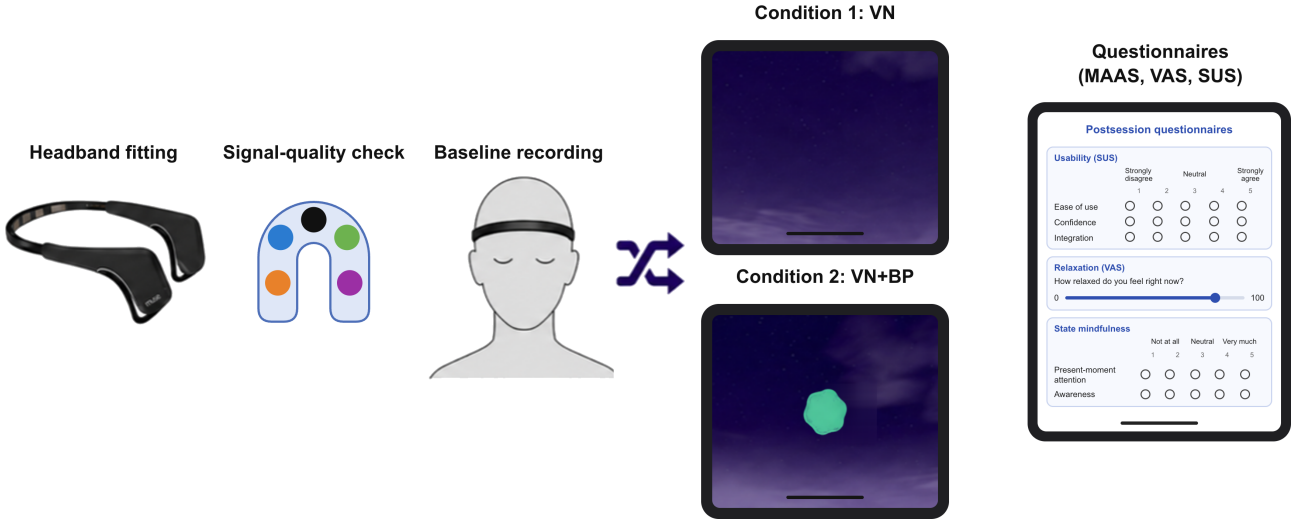
Interventions

The intervention was delivered entirely through a tablet app without direct experimenter assistance. After initiating the session on the tablet, participants watched a 2-minute instructional video explaining the study purpose, safety requirements, eligibility criteria, and procedural steps. They were then guided step by step by on-screen prompts. Participants fitted the Muse 2 consumer EEG headband and adjusted it until the app indicated acceptable contact quality. If acceptable contact quality was not reached, participants could proceed after an on-screen warning; these sessions were flagged by the app, were not recorded, and were excluded from all analyses. Participants then completed a 30-second eyes-closed baseline recording. Immediately after baseline, the app automatically assigned participants in a 1:1 ratio, using simple randomization without blocking or stratification, to visual neurofeedback only (VN) or visual neurofeedback plus breathing pacer (VN+BP). Participants subsequently completed a 120-second beta-responsive visual NFB task in their assigned condition. Before the task, the tablet displayed a short instruction screen with example images of the sky-based feedback. Participants were told that the display would respond to their current brain activity relative to their own baseline. They were encouraged to relax, observe the display, and explore whether they could influence it. The wording avoided explicit claims that the task would necessarily reduce beta activity below the eyes-closed baseline.

After the task, participants removed the headset. The Unity app (Unity Technologies) then directed them to a separate postsession questionnaire in Qualtrics on the same tablet. The questionnaire assessed demographic characteristics, previous NFB-related experience, alcohol consumption, and stimulant medication use within the preceding 8 hours, perceived usability, subjective relaxation, and state mindfulness. Total session time was approximately 10 - 12 minutes from consent to completion of the questionnaire. The complete participant workflow was autonomous and conducted on the same tablet without experimenter assistance. Figure 1 illustrates the autonomous experimental workflow.

XSL•FO
RenderX

Figure 1. Autonomous experimental workflow of the tablet-guided session, from the instructional video and consent to the postsession questionnaires. MAAS: Mindful Attention Awareness Scale; SUS: System Usability Scale; VAS: Visual Analog Scale; VN: visual neurofeedback only; VN+BP: visual neurofeedback plus breathing pacer.



NFB System

Overview

The NFB system was implemented as a custom Unity app developed using Unity 2022.3.10f1 and integrated with the Muse Unity software development kit (SDK), version 7.2.2. The Unity app handled digital consent and eligibility confirmation, headset setup, signal-quality checking, baseline acquisition, randomized condition assignment, and real-time visual feedback. After completion of the NFB task, the app directed participants to the separate postsession questionnaire in Qualtrics on the same tablet. The visual feedback consisted of a sky-themed environment. The display changed continuously as a function of each participant's current beta activity relative to their own baseline. Higher relative beta values were mapped to a darker and cloudier sky, whereas values closer to or below baseline were mapped to a clearer sky with greater star visibility. The mapping was continuous and smoothed to avoid abrupt visual changes. If the headset disconnected or the SDK headband-fit status indicated unacceptable signal quality, the display reverted to a neutral intermediate state. The Unity app ran on an Apple iPad with a 10.2-inch display running iPadOS. The tablet was positioned approximately 50 cm in front of the participant at eye level. NFB computation, randomized condition implementation, and display rendering were handled by the Unity app on the tablet. The postsession questionnaire was administered separately through Qualtrics on the same device. The seating position, tablet position, and booth lighting were kept stable across sessions to reduce variation in posture, viewing distance, and visual distraction. For sessions that met the contact-quality threshold, the Unity app stored session metadata, condition assignment, timestamps, SDK-derived band-power values, signal-quality information, and technical events under an anonymous session identifier. The same identifier was entered in Qualtrics, enabling the Unity app and questionnaire records to be linked without using directly identifying information. A short demonstration of the NFB training with the breathing pacer is provided in Multimedia Appendix 1.

EEG Acquisition and Feedback Signal

EEG data were acquired using a Muse 2 consumer EEG headband with 4 dry electrodes positioned at AF7, AF8, TP9, and TP10. The Muse Unity SDK provided manufacturer-processed band-power estimates and channel-quality indicators. These SDK-derived values were used for real-time feedback and descriptive EEG analyses. The real-time feedback signal was SDK-derived relative beta power, averaged across AF7, AF8, TP9, and TP10. Relative beta was used for feedback because it expresses beta activity relative to each participant's concurrent broadband EEG activity, reducing sensitivity to inter-individual differences in absolute signal amplitude, headset fit, and electrode contact. For each participant, the baseline reference was computed as the temporal mean of relative beta during the 30-second eyes-closed baseline.

During the NFB task, the app continuously computed the deviation of current relative beta from the participant-specific baseline:

$$\Delta\beta t=(\beta rel,t-\beta baseline)\times 10$$

This value was smoothed in Unity using frame-based linear interpolation:

$$\beta\sim t=Lerp(\beta\sim t-1,\Delta\beta t,0.5\times\Delta t)$$

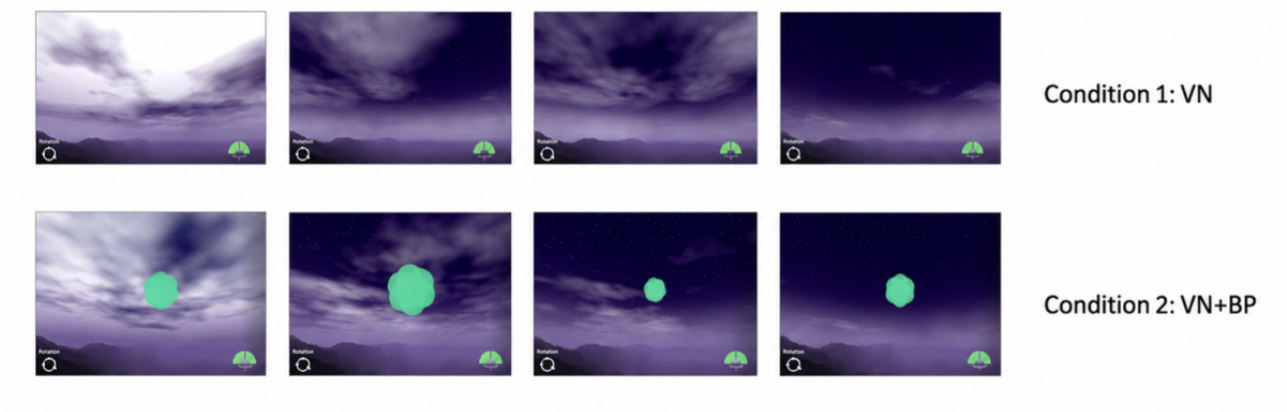
The resulting smoothed value controlled the visual feedback display. This baseline-referenced relative beta signal was used only for the real-time feedback. For descriptive EEG analyses, we used separately logged SDK-derived absolute band-power estimates, expressed on the manufacturer's log-scaled arbitrary-unit scale. These absolute band-power estimates did not drive the visual feedback. The SDK-derived estimates were treated as manufacturer-provided features and were not independently validated against an offline EEG preprocessing pipeline. Raw EEG time series were not retained by the acquisition app. The available EEG data therefore consist of SDK-derived band-power streams and associated signal-quality metadata.

XSL•FO
RenderX

Experimental Conditions

Participants were randomized to 1 of 2 conditions. In the VN, the screen displayed only the beta-responsive sky environment. Participants were instructed to observe the display, relax, and explore whether they could influence the visual feedback. In the VN+BP, participants saw the same beta-responsive sky environment with an additional centrally overlaid circular breathing guide. The circle expanded during inhalation and contracted during exhalation at a fixed pace of approximately 6 breaths/minute. The breathing pacer was not physiologically closed-loop and did not adapt to EEG or respiratory data. Figure 2 illustrates the beta-responsive visual feedback interface in the VN and VN+BP conditions. The two conditions were otherwise identical in duration, headset setup, baseline procedure, visual environment, questionnaire delivery, and autonomous workflow.

Figure 2. Beta-responsive visual feedback interface in the visual neurofeedback only (VN) and visual neurofeedback plus breathing pacer (VN+BP) conditions.



Intervention Fidelity

For recorded sessions, the app automatically recorded the anonymous session identifier, condition assignment, phase timestamps, SDK-derived band-power values, signal-quality indicators, and technical events. Completion of the 120-second task was used as the principal measure of intervention completion. The Unity app and Qualtrics questionnaire stored data separately, and records were linked using the anonymous session identifier. Questionnaire responses were not used to control the NFB display or breathing pacer. The intervention was delivered once. No individualized adaptation, additional training, intervention modification, or follow-up session was provided.

Outcomes

Overview

Perceived usability was the primary outcome and was assessed using the 10-item SUS, a widely used instrument for evaluating interactive systems [5,31]. Items are rated on a 5-point scale from 1=strongly disagree to 5=strongly agree. Standard SUS scoring yields a total score from 0 to 100, with higher scores indicating better usability. SUS was analyzed as a continuous variable. For descriptive purposes, scores below 68 were categorized as low usability and scores of 68 or higher as acceptable-or-better usability, following established interpretive guidelines [32]. Subjective relaxation was assessed using a single-item Visual Analog Scale (VAS): “How relaxed did you feel during the neurofeedback session?” Participants responded on a 0 - 100 slider, with 0 indicating not at all relaxed and 100 indicating extremely relaxed. State mindfulness was assessed using the state version of the Mindful Attention Awareness Scale–State (MAAS-State) [38]. Items assessed present-moment attention and awareness during the session. After reverse scoring as specified by the scale instructions, higher mean scores indicated greater state mindfulness.

Exploratory EEG features consisted of SDK-derived absolute band-power estimates at baseline and during the task. The main exploratory EEG feature of interest was pretask resting beta power, averaged across AF7, AF8, TP9, and TP10 during the 30-second eyes-closed baseline. Additional descriptive EEG features included theta, alpha, beta, and gamma band-power estimates. Delta-band activity was not included in the main EEG summaries because low-frequency activity in brief, unsupervised consumer EEG recordings with dry electrodes is particularly vulnerable to movement, ocular artifacts, and slow drift. System-performance and feasibility measures included session initiation, task completion, signal-quality failure, setup duration, Bluetooth disconnections, app errors, and data completeness. No formal progression criteria for proceeding to a future definitive randomized trial were prespecified.

Sample Size

This trial was designed as an exploratory feasibility and usability study rather than as a confirmatory efficacy trial. The sample size was determined pragmatically by the exhibition period (approximately 3 months), expected visitor flow, and the aim of estimating autonomous completion and perceived usability with reasonable precision. No formal power calculation was performed. Feasibility rates and key usability estimates are therefore reported with 95% CIs, and exploratory EEG-user experience analyses are interpreted cautiously.

Randomization

Sequence Generation

The tablet app generated the condition assignment at the participant level using an internal randomization routine configured for equal allocation to VN and VN+BP. Assignment

XSL•FO
RenderX

occurred immediately after completion of the 30-second baseline recording. No participant characteristics were used to determine the assignment. The assigned condition was stored automatically together with the EEG recording.

Allocation Concealment Mechanism

Condition assignment was generated and implemented automatically only after completion of the baseline recording. Neither participants nor on-site study personnel had access to the upcoming assignment before allocation. The assigned condition became apparent only after randomization, when the task interface was displayed. No researcher could manually select, alter, or override an assignment.

Implementation

Participants initiated and completed the procedure independently. The tablet app enrolled the participant session, generated the allocation, presented the assigned intervention, and recorded the condition. No member of the research team manually generated or communicated assignments.

Blinding

Participants could not be blinded to the breathing pacer because it was visually apparent during the task. Both conditions otherwise used the same headset, baseline procedure, beta-responsive sky environment, task duration, general instructions, and postsession questionnaire. No experimenter delivered the intervention or communicated the assigned condition. Questionnaire responses were entered directly by participants through Qualtrics. Analytic exclusion rules were applied without reference to outcome values. Statistical analyses used the recorded condition labels, and no formal analyst-blinding procedure was used.

Statistical Methods

Analysis Populations and Exclusion Criteria

Feasibility analyses used genuine visitor session initiators as the denominator. Sessions involving device testing or kiosk inactivity were not counted as participant sessions. Task-completion analyses included sessions that reached the end of the 120-second NFB task with a valid recording. Sessions that ended before allocation or that proceeded despite insufficient EEG contact quality were included in the feasibility assessment but not in the primary outcome analyses. The analytic sample consisted of participants who completed the NFB task, had a linkable postsession questionnaire record, and met the analytic inclusion criteria. Participants were analyzed according to their randomized condition.

Participants who did not complete the postsession questionnaire were treated as lost to follow-up for the primary outcome. Records were excluded from the primary analytic sample when participants reported alcohol consumption or stimulant medication use during the preceding 8 hours and completed the postsession questionnaire in under 30 seconds.

Participants who reported recent alcohol or stimulant medication use were classified under this exclusion criterion regardless of whether they completed the remaining questionnaire items. Alcohol and stimulant medication questions were administered after the intervention. These responses therefore did not affect eligibility, randomization, or intervention completion. Because allocation could not be retrieved for 17 randomized sessions with insufficient contact quality, and because records were excluded after randomization, the primary outcome analyses were not full intention-to-treat analyses.

Missing Data

Participant-level and item-level missingness were assessed separately. Failure to complete the postsession questionnaire was treated as participant-level (unit) nonresponse, that is, as loss to follow-up for the primary outcome. Missing individual responses within otherwise completed questionnaires were treated as item-level missingness. Item-level missingness was assessed across the 10 SUS items, the relaxation VAS, and the 5 MAAS-State items. Derived total and mean scores were not entered alongside their constituent items in the missingness test. A Little missing completely at random (MCAR) test was applied to the 16 raw questionnaire items in the final analytic dataset. Because item-level missingness was minimal, was limited to the secondary MAAS-State outcome, and did not affect the primary SUS outcome, multiple imputation was not performed. Available-case analyses were used, and outcome-specific denominators were reported.

Statistical Analyses

Statistical tests were 2-sided with an alpha level of .05. Continuous variables are reported as mean and SD, median and IQR or range, as appropriate. Categorical variables are reported as counts and percentages. Feasibility proportions and other principal proportions were reported with Wilson 95% CIs. Between-condition differences in SUS, subjective relaxation, MAAS-State, and descriptive EEG band-power estimates were examined using Mann-Whitney *U* tests. Rank-biserial correlations and 95% CIs were reported as effect sizes for key continuous comparisons. Categorical variables, including the proportion of participants with SUS scores of at least 68, were compared using chi-square tests or Fisher exact tests when expected cell counts were small. Associations between baseline EEG estimates and subjective outcomes were examined using Spearman rank correlations. The principal exploratory association examined pretask resting beta power and SUS total score. Additional exploratory analyses examined associations between baseline band-power estimates and subjective relaxation and state mindfulness. Partial Spearman correlations adjusted for age and gender were conducted as sensitivity analyses where reported. Multiplicity was addressed separately for the EEG-outcome and SUS item-level analysis families. Benjamini-Hochberg correction was applied across the baseline band-power by subjective-outcome correlation family. For the 10 SUS items, all items were oriented so that higher values indicated better perceived usability. Holm correction and permutation-based family-level tests with 10,000 permutations were used to evaluate the overall item-level pattern and the strongest individual item association.

Additional sensitivity analyses of the beta-usability association included bootstrap and permutation tests of the total-score correlation, leave-one-out influence analyses, robust regression, and item-level modeling. These analyses are reported in

XSL•FO
RenderX

Multimedia Appendix 2. EEG-user experience analyses were considered exploratory and hypothesis-generating. Statistical analyses were performed in Python using *NumPy*, *pandas*, *SciPy*, and *statsmodels*. Figures were produced using *matplotlib*. Little MCAR test was performed in R (version 4.6.1; R Core Team) using the mcar_test function from the *naniar* package.

Ethical Considerations

The trial was reviewed and approved by the ETH Zurich Ethics Commission (reference 25 ETHICS-086). All procedures were conducted in accordance with the Declaration of Helsinki and the institutional guidelines for research with human participants [39]. Before participation, each participant viewed an on-screen information sheet describing the study objective, voluntary participation, the right to withdraw at any time without penalty, data handling, anonymity, minor potential risks such as headband pressure, and eligibility criteria. Participants provided digital informed consent by confirming that they met the eligibility criteria, understood the study information, and agreed to participate. The procedure was available in German or French. No directly identifiable information was collected. Each session was stored under an anonymous session code. Participants received no financial or material compensation. The figures in this paper and the supplementary material do not contain identifiable participant images; the system schematic uses a nonidentifiable illustration.

Registration

This randomized feasibility and usability trial was not registered prospectively or retrospectively in a primary trial registry. Study materials, deidentified data, and analysis documentation were deposited retrospectively in the Open Science Framework repository.

Protocol

No prospectively registered trial protocol was available. The Open Science Framework repository contains available study documentation, questionnaire scoring procedures, analytic exclusion rules, deidentified data, and analysis scripts.

Results

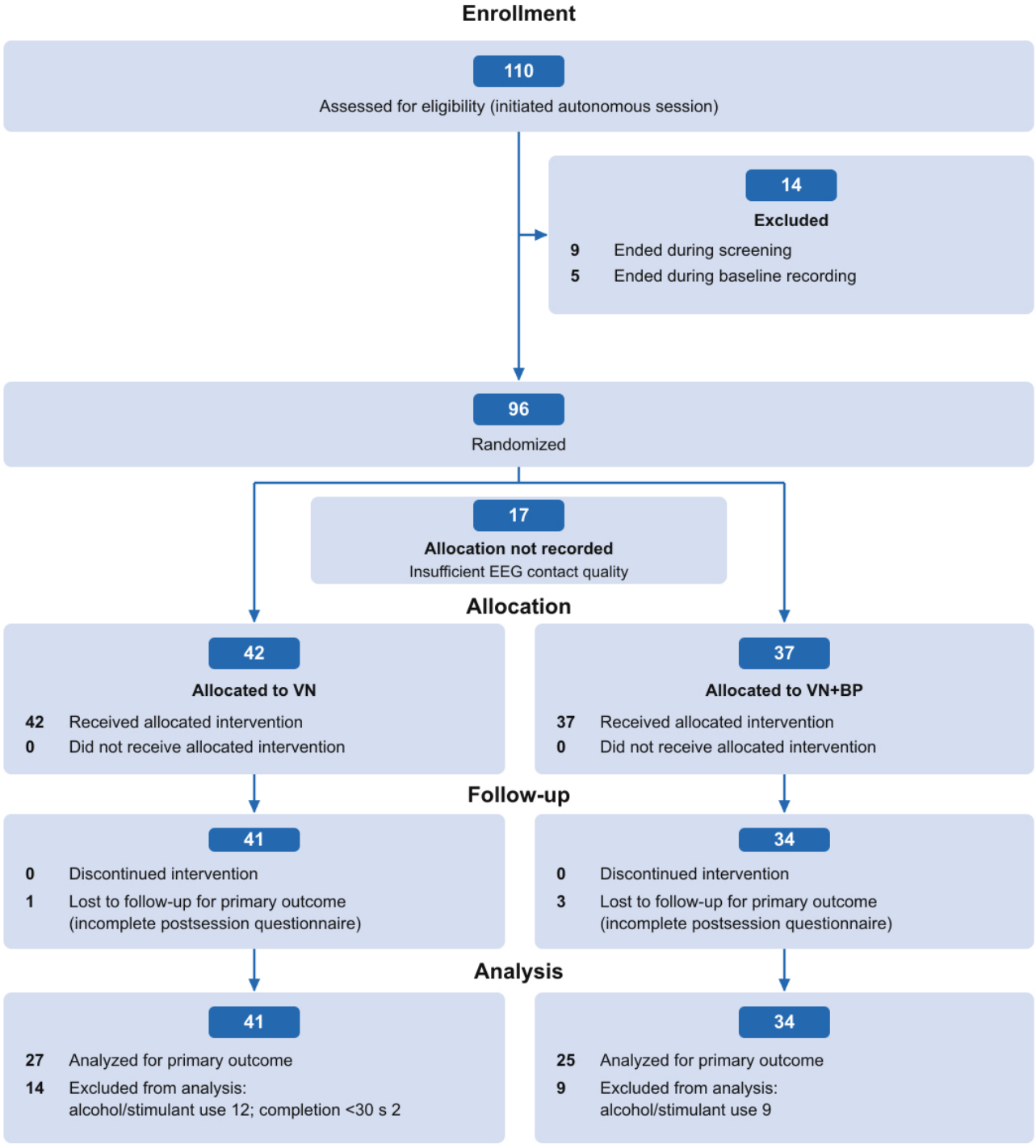
Participant Flow

Overview

A total of 110 visitors initiated an autonomous study session (Figure 3). A total of 14 (12.7%, 95% CI 7.7 - 20.2) sessions ended before allocation: 9 during screening and 5 during the baseline recording. The remaining 96 sessions were randomized. In 17 of these (17.7% of randomized sessions, 95% CI 11.4 - 26.5), EEG contact quality did not meet the app's threshold before allocation; participants continued despite an on-screen warning, but these sessions were not recorded and were excluded from all analyses, and their allocation could not be retrieved. Because contact quality was assessed before allocation, these losses are unlikely to be related to the assigned condition. Overall, 79 of 110 (71.8%, 95% CI 62.8 - 79.4) session initiators and 79 of 96 (82.3%, 95% CI 73.5 - 88.6) randomized participants completed the 120-second NFB task with a valid recording: 42 were allocated to VN and 37 to VN+BP. Among these 79 participants, 4 did not complete the postsession questionnaire and were therefore lost to follow-up for the primary outcome (VN, n=1; VN+BP, n=3). A further 23 participants were excluded from the primary analysis: 21 reported alcohol consumption or stimulant medication use within the preceding 8 hours (VN, n=12; VN+BP, n=9), and 2 completed the questionnaire in under 30 seconds (VN, n=2; VN+BP, n=0). The final analytic sample therefore comprised 52 participants: 27 in VN and 25 in VN+BP. The analytic sample represented 65.8% (52/79) of task completers, 54.2% (52/96) of randomized participants, and 47.3% (52/110; 95% CI 38.2 - 56.5) of all initiated sessions. Figure 3 presents the participant flow by randomized condition.

XSL•FO
RenderX

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) 2025 flow diagram of participants through the trial. EEG: electroencephalography; VN: visual neurofeedback only; VN+BP: visual neurofeedback plus breathing pacer.



Recruitment

Recruitment and data collection occurred during a public science exhibition in Bern, Switzerland, between March and June 2025. Because the trial involved a single session, there was no follow-up period. The trial ended as planned at the conclusion of the exhibition period.

Baseline Data

Participants in the final analytic sample were between 18 and 66 (mean 38.9, SD 11.6; median 37.0, IQR 30.0-47.3) years old; 35 of 52 (67.3%) participants were women. The final analytic sample comprised 27 participants in VN and 25 participants in VN+BP. The groups did not differ in age (Mann-Whitney U=329.0, P=.84) or gender distribution (χ^2_1 =0.002, P=.97).

Table 1 summarizes demographic characteristics, system-performance metrics, and data completeness for the final analytic sample.

XSL•FO
RenderX

Table . Analytic sample characteristics, system performance, and data completeness in a single-session, 2-arm randomized feasibility and usability study of autonomous consumer-EEG (electroencephalography) beta-responsive visual neurofeedback at a public science exhibition in Bern, Switzerland, between March and June 2025[a].

Characteristic	Total (N=52)	VN[b] (n=27)	VN+BP[c] (n=25)	Test statistic	*P* value
Demographics					
Age (years)					
Mean (SD)	38.9 (11.6)	39.1 (11.5)	38.7 (11.9)	329.0[d]	.84
Median (IQR)	37.0 (30.0-47.3)	37.0 (31.0-43.5)	43.0 (30-50)	—[e]	—
Sex, n (%)					
Female	35 (67.3)	18 (66.7)	17 (68.0)	0.002 (1)[f]	.97
Male	17 (32.7)	9 (33.3)	8 (32)	—	—
System performance					
Session duration (min), mean (SD)	11.2 (1.8)	11.0 (1.7)	11.4 (1.9)	318.0[d]	.70
Setup to acceptable signal quality (s), mean (SD)	42.3 (18.6)	40.7 (17.6)	44.1 (19.8)	312.0[d]	.63
Baseline recording duration (s)	30.0 (0.0)	30.0 (0.0)	30.0 (0.0)	—	—
Neurofeedback task duration (s)	120.0 (0.0)	120.0 (0.0)	120.0 (0.0)	—	—
Acceptable EEG signal quality, n (%)	52 (100)	27 (100)	25 (100)	—	—
Bluetooth disconnections, n	0	0	0	—	—
App errors, n	0	0	0	—	—
Data completeness					
Baseline EEG available, n (%)	52 (100)	27 (100)	25 (100)	—	—
Task EEG available, n (%)	52 (100)	27 (100)	25 (100)	—	—
SUS[g] complete, n (%)	52 (100)	27 (100)	25 (100)	—	—
Relaxation VAS[h] complete, n (%)	52 (100)	27 (100)	25 (100)	—	—
MAAS-State[i] score available, n (%)	51 (98.1)	27 (100)	24 (96)	—	—
Overall completeness, n/N (%)	826/832 (99.3)	432/432 (100)	394/400 (98.5)	—	—

[a]Mann-Whitney *U* tests were used for continuous variables, and chi-square tests were used for categorical variables.

[b]VN: visual neurofeedback only.

[c]VN+BP: visual neurofeedback plus breathing pacer.

[d]Mann-Whitney *U* test value.

[e]Not applicable.

[f]Chi-square (*df*) test value.

[g]SUS: System Usability Scale.

[h]VAS: Visual Analog Scale.

[i]MAAS-State: Mindful Attention Awareness Scale–State.

XSL•FO
RenderX

Numbers Analyzed

Participant-level and item-level missing data were assessed separately. At the participant level, 31 of the 110 (28.2%) session initiators did not yield a valid task recording: 14 ended before allocation, and 17 proceeded despite insufficient EEG contact quality. Among the 79 task completers with a valid recording, 4 did not complete the postsession questionnaire and were treated as participant-level nonresponse (5.1%). SUS scores, relaxation ratings, baseline EEG estimates, and task EEG estimates were available for all 52 participants in the final analytic sample: 27 in VN and 25 in VN+BP. A MAAS-State score was available for 51 participants: 27 in VN and 24 in VN+BP. One VN+BP participant had no MAAS-State items available and therefore had no score. Another VN+BP participant completed 4 of the 5 MAAS-State items and received a score based on the available items.

Among the 52 records retained in the analytic dataset, all SUS items and relaxation VAS responses were complete. Item-level missingness was confined to the MAAS-State and affected 2 of 52 (3.8%) participants. Six of 832 (0.7%) expected questionnaire-item responses were missing. Little MCAR test provided no evidence against the assumption that item-level missingness was completely at random, χ^2_{26}=15.2, *P*=.95. A sensitivity analysis among these 69 participants yielded the same conclusion, χ^2_{26}=16.0, *P*=.94. The remaining 6 participants excluded from the primary analysis (4 of the 21 excluded for recent alcohol or stimulant use and both participants excluded for a questionnaire duration under 30 seconds) had not provided any responses to the 16 questionnaire items and therefore could not contribute to this test. These results were interpreted cautiously because only 2 records contained missing items, all within the MAAS-State, so the test had limited power to detect departures from MCAR. Multiple imputation was not performed because item-level missingness was minimal and confined to the MAAS-State measure. Outcome-specific available-case analyses were used.

Outcomes and Estimation

Feasibility of Autonomous Use

All 52 retained participants reached acceptable signal quality and completed the baseline and NFB task. Mean setup time from headset placement to acceptable signal quality was 42.3 (SD 18.6) seconds. No Bluetooth disconnections or app errors were recorded in the analytic sample. EEG-derived data, SUS scores, and relaxation ratings were available for all 52 participants, and MAAS-State scores were available for 51 of 52 (98.1%) participants.

Psychological Outcomes by Randomized Condition

Table 2 summarizes perceived usability, subjective relaxation, and state mindfulness by randomized condition. Overall usability was good, with a mean SUS score of 72.3 (SD 14.7; 95% CI 68.3 - 76.2) and a median score of 75.0 (IQR 63.8-82.5). Using the conventional SUS reference value of 68, 31/52 (59.6%; 95% CI 46.1 - 71.8) participants rated the system in the acceptable-or-better usability range. Usability did not differ between VN and VN+BP (median 75.0, IQR 63.8-82.5 vs median 72.5, IQR 65.0-90.0; Mann-Whitney *U*=363.0, *P*=.65; rank-biserial *r*=−0.07, 95% CI −0.39 to 0.25). The proportion of participants with SUS scores of 68 or higher was also similar between conditions (15/25, 60%, vs 16/27, 59.3%; χ^2_1=0.002, *P*=.96). The overall median relaxation rating was 62.0 (IQR 50.0-72.3). Relaxation ratings did not differ between VN and VN+BP (median 63.0, IQR 50.0-80.0 vs median 61.0, IQR 50.0-67.0; Mann-Whitney *U*=277.0, *P*=.27; rank-biserial *r*=0.18, 95% CI −0.14 to 0.49). State mindfulness also did not differ between VN and VN+BP (median 2.4, IQR 2.1-2.8 vs median 2.5, IQR 2.0-3.2; Mann-Whitney *U*=346.0, *P*=.68; rank-biserial *r*=−0.07, 95% CI −0.39 to 0.26). Figure 4 visualizes the distributions of perceived usability, subjective relaxation, and state mindfulness by randomized condition.

XSL•FO
RenderX

Table . Psychological outcomes by randomized condition in a single-session, 2-arm randomized feasibility and usability study of autonomous consumer-EEG (electroencephalography) beta-responsive visual neurofeedback[a].

Outcome	Total (N=52)	VN[b] (n=27)	VN+BP[c] (n=25)	Effect size, *r* (95% CI)	Test statistic	*P* value
SUS[d] score, mean (SD)	72.3 (14.7)	71.2 (12.6)	73.4 (16.8)	—[e]	—	—
SUS score, median (IQR)	75.0 (63.8-82.5)	75.0 (62.5-81.3)	72.5 (65.0-90.0)	−0.07 (−0.39 to 0.25)[f]	363.0[g]	.65
SUS ≥68, n/N (%)	31/52 (59.6)	16/27 (59.3)	15/25 (60.0)	—	0.002 (1)[h]	.96
Relaxation VAS[i], mean (SD)	59.4 (18.7)	61.8 (19.7)	56.8 (17.6)	—	—	—
Relaxation VAS, median (IQR)	62.0 (50.0-72.3)	63.0 (50.0-80.0)	61.0 (50.0-67.0)	0.18 (−0.14 to 0.49)[f]	277.0[g]	.27
MAAS-State[j], mean (SD)	2.5 (0.7)	2.4 (0.5)	2.5 (0.8)	—	—	—
MAAS-State, median (IQR)	2.4 (2.0-2.8)	2.4 (2.1-2.8)	2.5 (2.0-3.2)	−0.07 (−0.39 to 0.26)[f]	346.0[g]	.68

[a]Mann-Whitney *U* tests were used for continuous or ordinal outcomes; chi-square tests were used for categorical outcomes. MAAS-State scores were available for 51 participants (VN, n=27; VN+BP, n=24). One VN+BP participant completed 4 of the 5 items and received a score based on the available items; another VN+BP participant had no MAAS-State items available and therefore had no score.

[b]VN: visual neurofeedback only.

[c]VN+BP: visual neurofeedback plus breathing pacer.

[d]SUS: System Usability Scale.

[e]Not applicable.

[f]Rank-biserial.

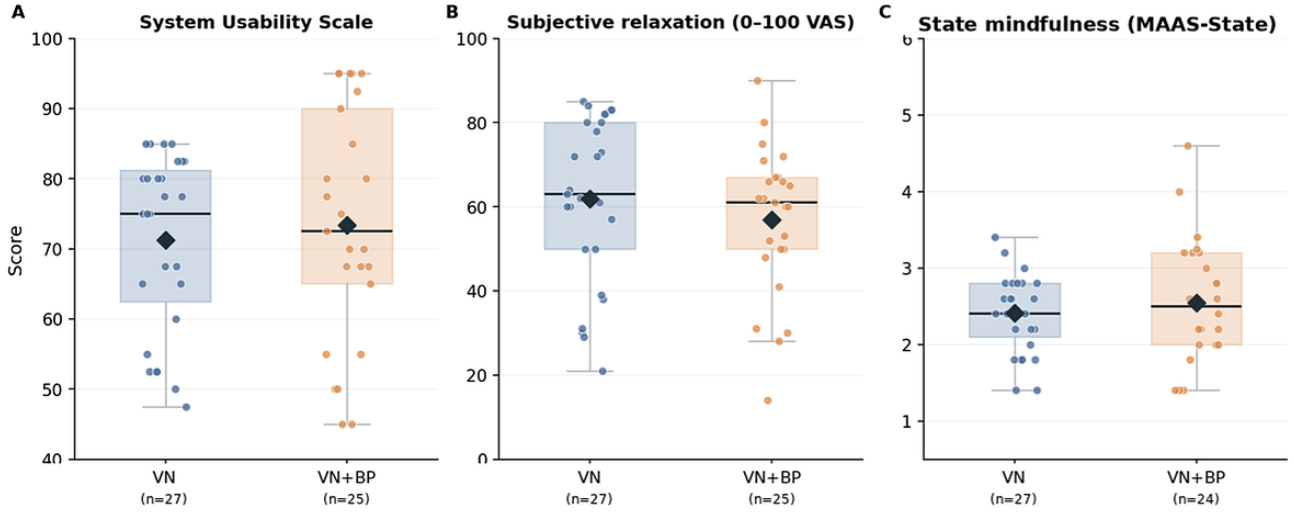
[g]Mann-Whitney *U* test value.

[h]Chi-square (*df*) test value.

[i]VAS: Visual Analog Scale.

[j]MAAS-State: Mindful Attention Awareness Scale–State.

Figure 4. Psychological outcomes by randomized condition. MAAS-State: Mindful Attention Awareness Scale–State; SUS: System Usability Scale; VAS: Visual Analog Scale; VN: visual neurofeedback only; VN+BP: visual neurofeedback plus breathing pacer.



Descriptive SDK-Derived EEG Band-Power Estimates

Table 3 reports SDK-derived absolute band-power estimates during the 30-second eyes-closed baseline and the 120-second NFB task. These values are reported descriptively to characterize the EEG feature space in the retained analytic sample. They should not be confused with the real-time feedback signal: the visual display was driven by baseline-referenced relative beta, whereas Table 3 reports separately logged SDK-derived absolute band-power estimates.

XSL•FO
RenderX

Table . Software development kit–derived absolute electroencephalography band-power estimates at baseline and during the 120-second neurofeedback task[a].

Band	Baseline total (N=52), mean (SD)	Baseline VN[b] (n=27), mean (SD)	Baseline VN+BP[c] (n=25), mean (SD)	Task total (N=52), mean (SD)	Task VN (n=27), mean (SD)	Task VN+BP (n=25), mean (SD)
Theta, 4 - 8 Hz	0.39 (0.31)	0.37 (0.29)	0.41 (0.33)	0.42 (0.34)	0.40 (0.32)	0.44 (0.36)
Alpha, 8 - 13 Hz	0.51 (0.33)	0.48 (0.31)	0.54 (0.36)	0.48 (0.31)	0.46 (0.29)	0.50 (0.33)
Beta, 13 - 30 Hz	0.36 (0.29)	0.34 (0.27)	0.38 (0.31)	0.38 (0.31)	0.36 (0.28)	0.40 (0.34)
Gamma, 30 - 50 Hz	0.10 (0.28)	0.09 (0.26)	0.11 (0.30)	0.11 (0.29)	0.10 (0.27)	0.12 (0.31)

[a]Values are software development kit–derived absolute band-power estimates averaged across AF7, AF8, TP9, and TP10 and expressed on the manufacturer's log-scaled arbitrary-unit scale. Baseline refers to the 30-second eyes-closed recording; task refers to the 120-second beta-responsive neurofeedback task. These descriptive absolute band-power estimates did not drive the visual feedback display; real-time feedback was driven by baseline-referenced relative beta.

[b]VN: visual neurofeedback only.

[c]VN+BP: visual neurofeedback plus breathing pacer.

Across participants, the descriptive spectral profile was similar at baseline and during the task. Mean beta power was 0.36 (SD 0.29) during baseline and 0.38 (SD 0.31) during the task. Between-condition comparisons showed no statistically significant differences in SDK-derived absolute band power for theta, alpha, beta, or gamma at baseline or during the task.

Between-condition comparisons showed no significant differences in absolute band power at baseline (theta: Mann-Whitney U=297.0, P=.46; alpha: Mann-Whitney U=312.0, P=.63; beta: Mann-Whitney U=289.0, P=.37; gamma: Mann-Whitney U=301.0, P=.50) or during the task (theta: Mann-Whitney U=305.0, P=.55; alpha: Mann-Whitney U=318.0, P=.70; beta: Mann-Whitney U=294.0, P=.42; gamma: Mann-Whitney U=308.0, P=.58).

Ancillary Analyses

Exploratory Resting Beta-Usability Association

Because between-condition differences were minimal, exploratory individual-difference analyses were collapsed across randomized conditions. Baseline beta power refers to SDK-derived absolute beta-band power during the 30-second eyes-closed baseline, averaged across AF7, AF8, TP9, and TP10 and expressed on the manufacturer's log-scaled arbitrary-unit scale.

Higher pretask resting beta power was associated with lower SUS total scores in the uncorrected Spearman rank correlation (ρ=−0.31, 95% CI −0.54 to −0.04, raw P=.03; Table 4). This total-score association did not survive Benjamini-Hochberg correction across the EEG-outcome family (adjusted P=.30). The corresponding Pearson correlation was not statistically significant (r=−0.24, P=.08). A sensitivity analysis adjusting for age and gender showed a similar negative association (partial Spearman ρ=−0.33, P=.02).

Baseline beta power was not significantly associated with subjective relaxation or state mindfulness (Table 4). Figure 5 visualizes the exploratory association between pretask resting beta power and SUS total score.

XSL•FO
RenderX

Table . Exploratory associations between pretask resting beta power and subjective outcomes[a].

Analysis	N	Spearman ρ estimate	95% CI	Raw *P* value	Adjusted or permutation *P* value
Baseline beta × SUS[b] total score	52	−0.31	−0.54 to −0.04	.03	.30[c]
Baseline beta × relaxation VAS[d]	52	−0.13	−0.39 to 0.15	.37	.61[c]
Baseline beta × MAAS-State[e]	51	0.11	−0.17 to 0.37	.44	.61[c]
SUS item 5: functional integration	52	−0.43	−0.63 to −0.18	.001	.01[f]; .02[g]
Global SUS item pattern	52	9/10 item correlations negative	—	—	.02[h]

[a]Baseline beta refers to software development kit–derived absolute beta-band power during the 30-second eyes-closed baseline, averaged across AF7, AF8, TP9, and TP10. Electroencephalography-outcome analyses were corrected across the baseline band-power by subjective-outcome family. System Usability Scale item analyses used permutation-based family-level tests with 10,000 permutations. All analyses in this table are exploratory.

[b]SUS: System Usability Scale.

[c]Benjamini-Hochberg–adjusted.

[d]VAS: Visual Analog Scale.

[e]MAAS-State: Mindful Attention Awareness Scale–State.

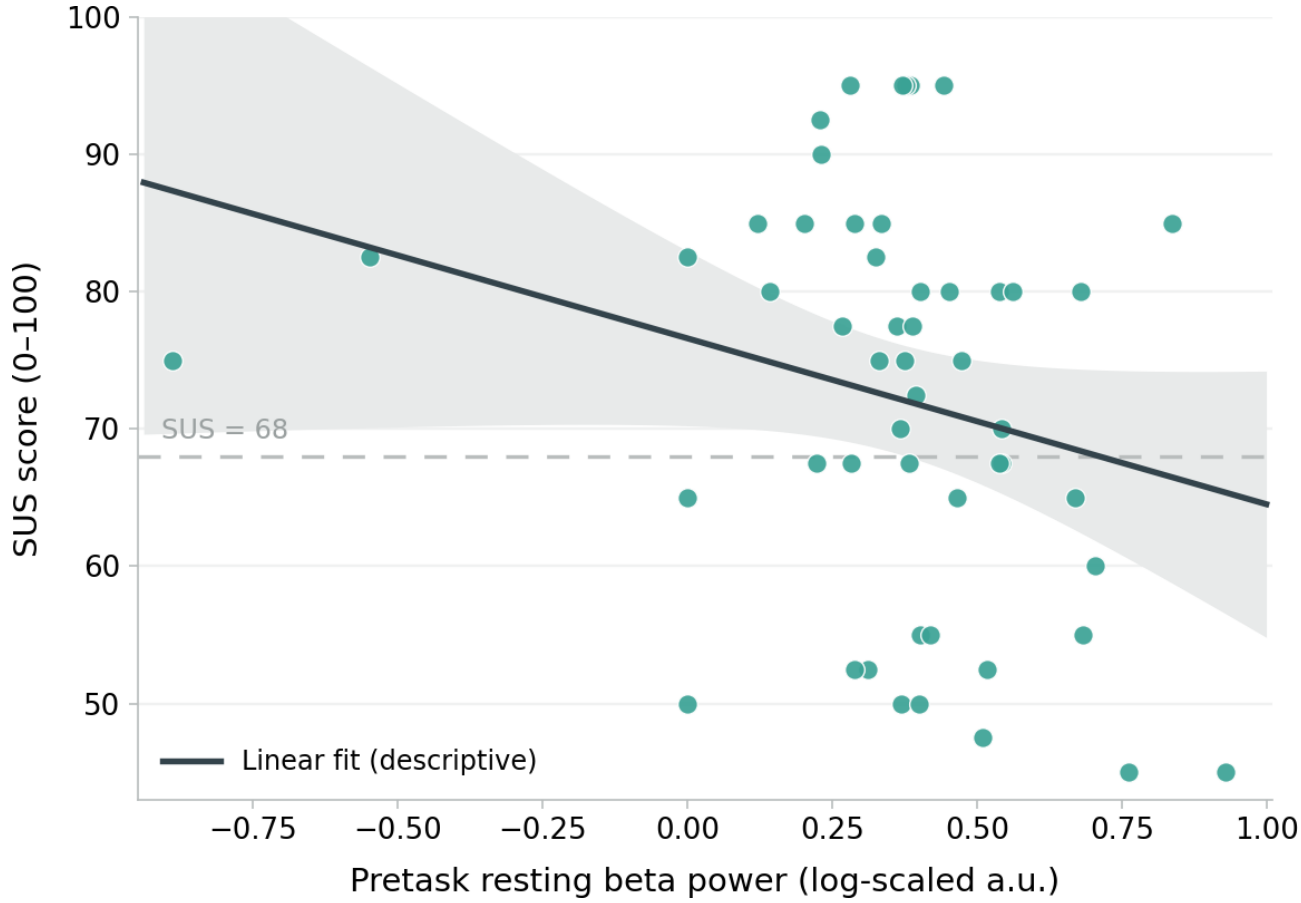
[f]Holm-adjusted.

[g]Max-item permutation.

[h]Permutation.

XSL•FO
RenderX

Figure 5. Pretask resting beta power and perceived usability. Scatterplot of software development kit–derived absolute beta-band power during the 30-second eyes-closed baseline and System Usability Scale (SUS) score in the final analytic sample (N=52). Baseline beta power was averaged across AF7, AF8, TP9, and TP10 and is expressed on the manufacturer's log-scaled arbitrary-unit scale. The dashed horizontal line indicates the conventional SUS reference value of 68. The fitted line is shown for descriptive visualization; inferential testing used Spearman rank correlation. Higher resting beta power was associated with lower SUS scores in the uncorrected Spearman correlation (ρ=−0.31, 95% CI −0.54 to −0.04, raw *P*=.03), but the total-score association did not survive Benjamini-Hochberg correction across the electroencephalography-outcome family (adjusted *P*=.30). Point colors, if used, indicate whether SUS scores were below or above 68 and are shown only for descriptive orientation. The figure should not be interpreted as evidence of prediction, causality, beta suppression, or learned beta regulation.



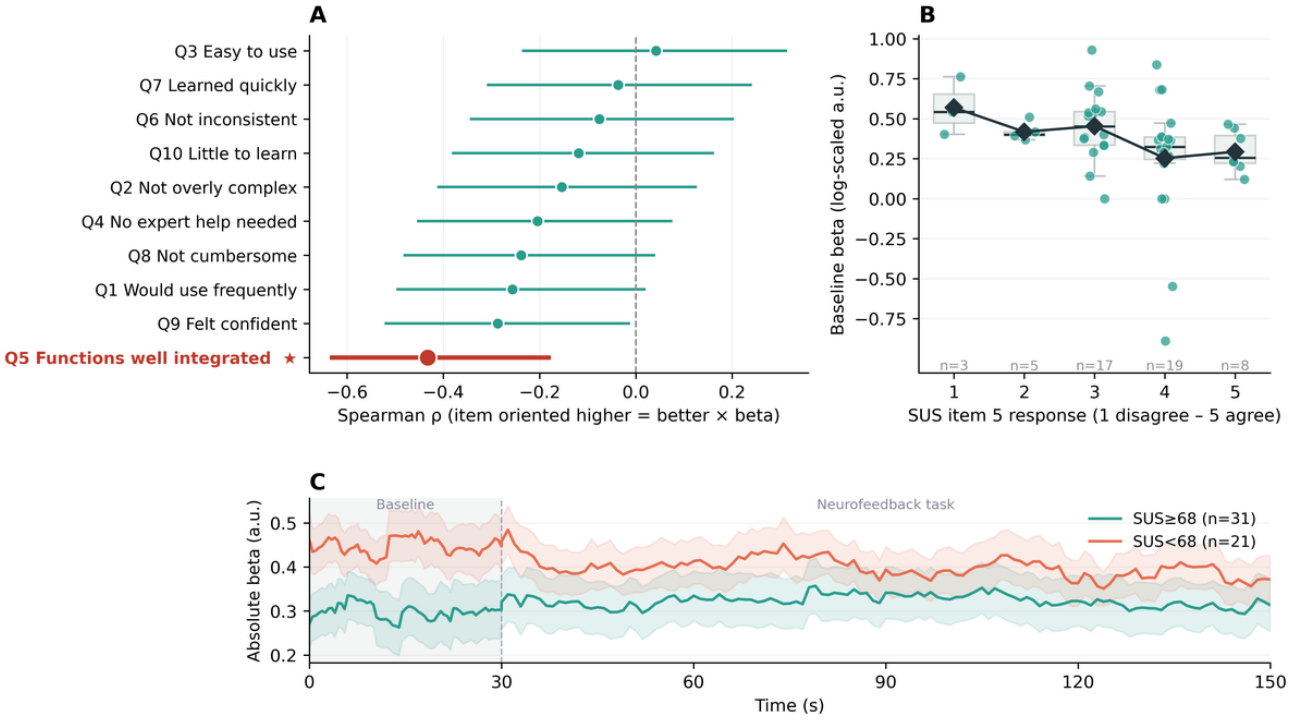
Exploratory SUS Item-Level Analyses

Exploratory SUS item-level analyses were conducted to examine the association pattern across the 10 SUS items. All items were oriented so that higher scores indicated better perceived usability. The family-level permutation test across SUS items was significant (global permutation *P*=.02). The strongest item-level association was observed for SUS item 5, perceived functional integration: higher baseline beta power was associated with lower perceived functional integration (Spearman ρ=−0.43, 95% CI −0.63 to −0.18, raw *P*=.001; Holm-adjusted *P*=.01; max-item permutation *P*=.02).

Figure 6 shows the item-level association pattern, the strongest item-level association for SUS item 5, and a descriptive time-course visualization of SDK-derived beta power across baseline and task by conventional SUS category.

XSL•FO
RenderX

Figure 6. Exploratory System Usability Scale (SUS) item-level associations and descriptive beta time course. (A) Spearman correlations between software development kit (SDK)–derived absolute beta-band power during the 30-second eyes-closed baseline and the 10 SUS items, with all SUS items oriented so that higher scores indicate better perceived usability. Points indicate Spearman ρ estimates and horizontal bars indicate 95% CIs. The dashed vertical line marks ρ=0. SUS item 5, perceived functional integration, is highlighted as the strongest item-level association. (B) SDK-derived absolute beta-band power by response level on SUS item 5, "I found the various functions in this system were well integrated." Points represent individual participants; boxes show median and IQR; diamonds indicate category means; response counts are shown below each category. The connecting line between category means is shown only for visual orientation. (C) Mean SDK-derived absolute beta-band power across the 30-second eyes-closed baseline and the 120-second neurofeedback task for participants with SUS scores below 68 and participants with SUS scores of 68 or higher. The vertical dashed line indicates the transition from baseline to the neurofeedback task. Shaded regions indicate ±1 SEM. In all panels, beta power was averaged across AF7, AF8, TP9, and TP10 and is expressed on the manufacturer's log-scaled arbitrary-unit scale. SUS item-level analyses were exploratory. The global item-level permutation test indicated a nonrandom item-level association pattern (global permutation *P*=.02). The strongest item-level association was observed for SUS item 5, perceived functional integration (Spearman ρ=−0.43, 95% CI −0.63 to −0.18, raw *P*=.001), which remained significant after correction across the 10 SUS items (Holm-adjusted *P*=.014; max-item permutation *P*=.02). Panel C uses post hoc SUS categories based on the conventional SUS reference value of 68 and is shown only for descriptive visualization. Panel C shows SDK-derived absolute beta-band estimates, not the baseline-referenced relative beta signal that drove the visual feedback display. The figure should not be interpreted as evidence of prediction, causality, beta suppression, or learned neurofeedback regulation.



Harms

Harms and minor transient discomfort were not assessed systematically. This limitation is also relevant to the broader consumer-NFB literature, in which adverse effects have often not been systematically evaluated [7,40]. The absence of reported adverse events in the present trial should therefore not be interpreted as evidence that no discomfort or other unintended effects occurred.

Discussion

Interpretation

This study examined 3 questions: whether first-contact users could complete a fully autonomous consumer-EEG NFB workflow in a public setting, whether adding a fixed paced-breathing guide improved usability or subjective experience, and whether pretask resting beta activity was associated with perceived usability. Three main findings emerged. First, autonomous use was feasible for many participants: most session initiators completed the NFB task, and participants retained in the analytic sample rated the system in the acceptable-to-good usability range. At the same time, the full initiation denominator showed that autonomous task completion was feasible for a majority, although not all initiated sessions progressed to a complete analyzable record. Second, adding a fixed, open-loop breathing pacer did not improve perceived usability, subjective relaxation, state mindfulness, or descriptive EEG band-power estimates compared with visual NFB alone. Third, exploratory analyses indicated a negative association between pre-task resting beta activity and perceived usability, with the clearest item-level association observed for perceived functional integration. However, as the total-score association did not remain significant after correction across the broader EEG-outcome family, these neural findings should be interpreted as hypothesis-generating rather than confirmatory.

Autonomous First-Contact Feasibility and Usability

The principal contribution of this trial was the evaluation of a consumer-EEG NFB system as a complete autonomous first-contact pathway rather than only as a portable feedback interface. NFB research has traditionally relied on structured laboratory or clinical protocols in which trained personnel support EEG setup, signal-quality assessment, task delivery,

XSL•FO
RenderX

and interpretation of the feedback [1,41,42]. Remote and consumer-facing approaches have expanded NFB beyond conventional laboratories, but many retain repeated-session structures, remote support, or some degree of professional involvement [12,40]. Consumer EEG hardware has nevertheless made lower-burden and more mobile deployment increasingly feasible, with recent studies demonstrating applications of wearable EEG and Muse-type systems outside conventional research-grade setups [21,43-45]. The present system extended this approach by integrating consent, headset fitting, signal-quality checking, baseline calibration, randomization, real-time feedback, and transition to postsession questionnaires within a workflow that participants completed without direct experimenter assistance. The observed completion pattern provides a more nuanced picture than usability scores alone. Most visitors who initiated the procedure completed the NFB task, suggesting that the workflow was manageable for many first-contact users in a low-support public setting. However, 31 of 110 initiated sessions did not yield a valid task recording. Insufficient EEG contact quality was the most frequent reason (17 sessions), exceeding sessions that ended before allocation (14 sessions), which indicates that achieving adequate electrode contact without assistance was the main barrier in the autonomous pathway. In addition, 4 participants did not complete the postsession questionnaire, and post-task exclusions reflected prespecified analytic and data-quality criteria, namely recent alcohol or stimulant use and implausibly rapid questionnaire completion; these exclusions should be distinguished from failures of the interface itself. This distinction is important for autonomous neurotechnology because feasibility encompasses several stages: initiating the system, establishing usable physiological signals, completing the intervention, and ultimately providing analyzable data. Contemporary brain-computer interface (BCI) and digital-health usability research similarly emphasizes that successful translation depends not only on algorithmic performance but also on setup burden, reliability, usability, satisfaction, and the ability of users to operate sensor-based systems independently [5,25].

Among retained participants, perceived usability fell within the conventionally interpreted acceptable-to-good range, based on established SUS interpretation and more recent digital-health benchmarking [31,32]. Setup was generally brief, data completeness was high, and no major app or connectivity problems were observed in the analytic sample. These findings complement previous research suggesting that consumer EEG can support NFB-related and psychophysiological applications outside specialized laboratories [21,43,44,46]. They also contribute to a growing literature on consumer-grade NFB, although recent meta-analytic evidence emphasizes that successful deployment and favorable user experience should be distinguished from demonstrated modulation of the intended neural target [40]. More broadly, usability and acceptance are established early outcomes in proof-of-concept evaluations of adaptive digital-health systems, while BCI usability research has highlighted workload, satisfaction, reliability, and ease of access as central requirements for practical translation [5,25]. This study extends this work to the technically demanding context of autonomous NFB, in which users must independently fit EEG hardware, establish adequate signal quality, complete baseline calibration, and interpret real-time physiological feedback. For autonomous NFB, usability has a particularly direct functional role. BCI research has emphasized that practical deployment depends not only on signal-processing performance but also on ergonomic factors, sensor setup, reliability, and the overall user experience [5,25,27]. Consumer EEG systems offer greater portability and accessibility, but wearable and dry-electrode recordings remain sensitive to electrode contact, movement, and other sources of signal degradation [43-45]. In supervised settings, an expert can assist with headset positioning, poor signal quality, task comprehension, and technical problems; in an autonomous workflow, these functions must instead be supported by the instructions, interface, and system logic [5,47]. This study therefore contributes evidence about the feasibility of the complete first-contact pathway rather than only the feedback display or the experience of users who had already completed setup. The favorable usability ratings should not, however, be interpreted as evidence of therapeutic benefit or successful NFB learning. This was a brief, single-session study without a sham-feedback, no-feedback, or no-intervention control condition. The results show that many first-time users could operate and evaluate the system autonomously; they do not establish that the task reduced symptoms, improved self-regulation, or produced learned control over beta activity. This distinction is important when positioning the present findings within the broader BCI literature; evidence for therapeutic BCI effects generally comes from repeated, task-specific clinical interventions, such as BCI-supported upper-limb rehabilitation after stroke [27,48]. This study instead evaluated a brief, autonomous first-contact experience in a nonclinical setting and should not be interpreted as a test of therapeutic BCI efficacy.

No Detectable Added Benefit of Fixed Breathing Guidance

The breathing pacer was included as a simple body-based cue that might support relaxation and make the visual feedback easier to follow. Slow breathing at approximately 6 breaths/minute has been associated with greater heart rate variability, increased parasympathetic influence, and enhanced baroreflex function [29,30]. Breath-focused practices may also influence attention and immediate psychological state, and recent work has shown that slow breathing can modulate EEG [28,49]. It was therefore plausible that adding paced-breathing guidance to the visual NFB task could improve the user experience. In this study, however, the fixed breathing pacer produced no detectable incremental improvement in perceived usability, subjective relaxation, state mindfulness, or descriptive EEG band-power estimates. This finding is specific to the brief, fixed, open-loop breathing overlay tested here and should not be interpreted as evidence that paced breathing is generally ineffective. Slow or resonance-frequency breathing has well-established effects on autonomic cardiovascular dynamics [29,30], and related biofeedback research has demonstrated effects on heart rate variability and user experience [50]. Nor does the present result establish that visual NFB itself caused the favorable subjective outcomes. Because both randomized conditions received the same beta-responsive visual NFB, the

XSL•FO
RenderX

comparison isolates only whether the additional breathing guide improved outcomes beyond visual NFB alone.

Several aspects of the implementation may have limited the added value of the breathing guide. The NFB task lasted only 120 seconds, which may have provided insufficient time for users to establish a stable slow-breathing rhythm [50]. The pacer used the same breathing rate for all participants and did not adapt to individual respiration, comfort, EEG activity, or autonomic state. Respiration and heart rate variability were not recorded, so it was not possible to determine whether participants followed the guide or whether it produced the intended physiological response. In addition, the breathing display may have been unnecessary for some participants, while for others it may have increased the amount of visual information competing for attention.

From a physiological-computing perspective, the relevant question may therefore be not simply whether breathing guidance is present, but whether its timing and intensity are appropriate for the individual user [50]. Recent work on human-centered digital health and adaptive closed-loop BCI systems supports greater attention to individualization and context-sensitive adaptation [25,33]. Future research could compare fixed with individualized breathing rates, measure respiratory adherence and autonomic responses, and test whether breathing guidance is more effective when introduced selectively; for example, when a user shows difficulty settling, unstable physiological signals, or reduced engagement. Such adaptive support remains a design hypothesis and was not evaluated in this study. The absence of differences in descriptive EEG measures also reinforces the need to distinguish a beta-responsive feedback task from demonstrated beta downregulation. The system continuously changed the visual environment according to each participant's baseline-referenced relative beta signal, but the study did not show that participants learned to reduce beta activity. This distinction is consistent with recent meta-analytic evidence showing that the use of consumer-grade NFB does not by itself establish reliable modulation of the intended neural target [40]. Demonstrating NFB learning requires evidence that the targeted neural feature changes systematically through interaction with the feedback, ideally across repeated training and against an appropriate control condition [42,51]. The current contribution lies in autonomous delivery and usability, not in evidence of successful beta modulation.

Exploratory Resting Beta-Usability Association

The exploratory EEG analyses suggest that the experience of an autonomous NFB interface may depend partly on the user's state before the task begins. Higher pretask resting beta activity was associated with lower SUS scores in the uncorrected analysis, and the direction of the association remained similar after adjustment for age and gender. However, the total-score association did not survive correction across the broader EEG-outcome family. The result should therefore not be considered confirmatory. The item-level analyses provided a more specific, although still exploratory, pattern. Most item correlations were negative, the overall item-level pattern was unlikely to have arisen by chance, and the clearest association concerned whether the system's functions felt well integrated. This suggests that resting beta may have been related more closely to the perceived coherence of the interaction than to subjective relaxation or state mindfulness [15]. Nevertheless, the individual item findings should be interpreted cautiously because some response categories contained relatively few observations and the analyses were not preregistered. Independent replication is needed before this pattern can be treated as reliable. Beta-band activity has been associated with cortical activation, vigilance, cognitive effort, and maintenance of the current neural or behavioral state [15,18,52]. Some NFB studies have also examined elevated high-beta activity in relation to anxious or hyperaroused states and have investigated beta downtraining in participants with anxiety or depressive symptoms [8,17]. More recently, resting-state EEG research has reported increased beta-band activity in patients with generalized anxiety disorder compared with healthy controls [20]. However, beta activity is not a specific measure of anxiety, and EEG biomarkers across anxiety and related psychiatric conditions remain heterogeneous [19]. This study did not assess baseline anxiety or clinical symptoms. It would therefore be inappropriate to conclude that participants with higher beta were more anxious, more in need of treatment, or less likely to benefit from NFB. One possible interpretation is that resting beta reflected a pretask neural state that influenced how an unfamiliar interface was experienced [15]. Although this study did not measure this process directly, a user entering the session in a more effortful or evaluative state might experience the relationship between headset setup, baseline calibration, visual feedback, and breathing guidance as less immediately coherent. This interpretation is consistent with physiological-computing and adaptive-BCI frameworks, in which physiological activity can function not only as a feedback target but also as a contextual variable related to interaction demands and user experience [5,26,33]. It remains, however, one possible explanation rather than a mechanism established by the data. If the association is replicated, baseline state could become relevant to the design of autonomous onboarding. Physiological and neural signals have increasingly been considered as potential inputs for adaptive system behavior in BCI and neuroergonomic research [5,27,33], while consumer-NFB research has identified calibration and individualization as priorities for future systems [40]. Building on these approaches, future studies could test whether users entering a session with higher pretask activation benefit from a longer settling phase, reduced instruction density, slower feedback transitions, or additional guidance, whereas users showing lower initial activation could proceed directly to the task. The present findings do not establish resting beta as a usability biomarker; rather, they provide a candidate state variable for prospectively testing such adaptive onboarding strategies. No formal progression criteria were prespecified, and the present trial was not designed to support a formal decision on whether to proceed to a definitive trial. Before a larger confirmatory trial, future studies should prespecify feasibility thresholds, retain and validate raw EEG, assess objective target engagement and breathing adherence, and evaluate repeated use with an appropriate control condition.

XSL•FO
RenderX

Limitations

This study has several limitations. First, it was a brief, single-session feasibility and usability study conducted at a public science exhibition. Participants were self-selected visitors interested in trying a neurotechnology demonstration and were not representative of the general population or a clinical group. The findings therefore concern autonomous first-contact use and cannot establish repeated-session adherence, NFB learning, durable self-regulation, symptom improvement, or clinical efficacy. Second, the study did not include sham feedback, breathing-only, or no-intervention conditions. Consequently, the observed usability, relaxation, and mindfulness ratings cannot be attributed specifically to NFB. The randomized comparison tested only whether adding the breathing pacer changed outcomes relative to visual NFB alone. Third, EEG was recorded with a 4-channel consumer headband. This enabled a scalable autonomous procedure but limited spatial inference and increased reliance on manufacturer-provided signal processing. The app did not retain raw EEG time series, preventing independent offline artifact removal, spectral estimation, or validation against a research-grade system. Consumer EEG recordings may be affected by differences in electrode contact, eye movements, cranial muscle activity, movement-related artifacts, and device-specific signal-processing characteristics [43-45]. In 17 of 96 randomized sessions, contact quality was insufficient; these sessions were not recorded, and their allocation could not be retrieved. Because contact quality was assessed before allocation, these losses are unlikely to be related to the assigned condition, but they reduced the analyzable sample. The real-time and analytic EEG measures also differed. The display was driven by baseline-referenced relative beta provided by the Muse SDK, whereas the exploratory analyses used separately logged SDK-derived absolute band-power estimates. These signals serve different purposes and should not be treated as equivalent. In addition, the eyes-closed baseline and the visually engaging task differed in visual input and attentional demands, limiting direct interpretation of baseline-to-task band-power differences as NFB-related regulation [53,54]. Fourth, the beta-usability analyses were exploratory and based on a modest sample. The total SUS association did not remain significant after correction across the broader EEG-outcome family, while the item-level result requires replication because it decomposed a composite measure into individual responses. Future studies should preregister the beta-usability hypothesis, define multiplicity families in advance, retain raw EEG, and compare consumer-derived features with validated offline pipelines and research-grade recordings. Fifth, the study relied predominantly on self-report outcomes. The SUS is an established measure of perceived usability [5,31], but it does not assess whether participants understood the feedback accurately, achieved objective control over the target signal, or would continue using the system. Future work should combine perceived usability with behavioral completion measures, comprehension checks, objective NFB-control indices, respiration, heart rate variability, state anxiety, perceived stress, workload, and prior experience with digital or neurotechnology. Finally, the breathing pacer was fixed and open-loop. Its physiological effect and participant adherence were not measured. Future studies should compare fixed with individualized or adaptive breathing support, investigate longer and repeated exposure, and determine whether perceived usability, physiological regulation, and control over the NFB signal improve together or represent distinct outcomes [28,30,50].

Generalizability

The findings are most applicable to brief, first-contact use by self-selected adults in a public, nonclinical exhibition setting. They should not be generalized to clinical populations, repeated home use, long-term adherence, therapeutic NFB, or users requiring substantial assistance with digital technology or EEG hardware without further usability and feasibility testing, in line with recommendations to evaluate sensor-based digital health technologies across relevant user groups and contexts of use [25]. Participation also required the ability to complete the procedure in German or French and willingness to engage with a neurotechnology demonstration in a public exhibition. These requirements may have selected for participants with greater interest in technology or greater confidence in navigating an autonomous tablet-based procedure. Because digital health literacy is closely linked to the adoption and use of digital health technologies [55,56], the present findings may overestimate usability relative to less technology-engaged populations. Nevertheless, the study provides information that may be relevant to similar low-support consumer-EEG applications. In particular, distinguishing session initiation, allocation, contact-quality failure, task completion, posttask exclusion, and final analytic retention offers a useful framework for evaluating autonomous neurotechnology outside controlled laboratory settings. The findings also highlight that the feasibility of the complete user pathway should be evaluated separately from the efficacy of the NFB signal or intervention.

Conclusion

This study evaluated the complete operation of a fully autonomous, consumer-EEG, beta-responsive visual NFB system in a public first-contact setting. The workflow integrated consent, self-fitting, signal-quality checking, baseline calibration, randomized condition assignment, real-time feedback, and transition to postsession self-report without direct operator assistance. A majority of session initiators completed the NFB task, and participants retained in the analytic sample reported generally favorable usability. However, autonomous completion was not universal, and insufficient EEG contact quality and posttask analytic exclusions substantially reduced the final analytic sample. Adding a fixed, open-loop breathing pacer produced no detectable improvement in usability, subjective relaxation, state mindfulness, or descriptive EEG outcomes, although the trial was not designed to establish equivalence between conditions. Exploratory analyses identified a possible relationship between pretask resting beta activity and perceived usability, particularly perceived functional integration, but the total-score association did not survive correction across the broader EEG-outcome family. The findings therefore provide evidence about autonomous feasibility and usability rather than beta suppression, therapeutic efficacy, or learned self-regulation. Future preregistered studies should retain and validate raw EEG, define progression and feasibility criteria prospectively, assess

XSL•FO
RenderX

objective target engagement and breathing adherence, include appropriate control conditions, and examine repeated use. Such work is necessary before autonomous consumer-EEG NFB can be evaluated as a home-based or clinical support tool.

Acknowledgments

We thank the exhibition organizers and staff in Bern as well as all visitors who volunteered to take part in the study. We also thank colleagues at the Mobiliar Lab for Analytics for their feedback on the study design and prototype implementation.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: code optimization, proofreading and editing, translation, and reformatting. The GenAI tools used were GPT-5.5 and Claude Opus 4.7. GPT-5.5 was used for proofreading, German-to-English translation, and reformatting of manuscript text, whereas Claude was used to review and, where necessary, optimize analysis code. All AI-assisted content and code modifications were critically reviewed and verified by the authors, who take full responsibility for the final manuscript and analyses. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. The declaration was submitted by MB.

Funding

This work was conducted at the Mobiliar Lab for Analytics, Department of Management, Technology, and Economics, ETH Zurich. No dedicated external grant was obtained for this study. The institutional host had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Data Availability

Deidentified analytical data and analysis materials are available in the Open Science Framework repository [57]. The repository includes the deidentified software development kit (SDK)–derived EEG band-power dataset, condition assignment, session metadata, questionnaire responses, questionnaire scoring scripts, exclusion and data-quality scripts, randomization documentation, and the analysis code used to reproduce the reported tables and figures.

Raw 256-Hz electroencephalography (EEG) time series are not available because they were not retained by the deployed acquisition application. The available EEG data therefore consist of manufacturer-processed Muse SDK band-power streams and associated signal-quality metadata. No directly identifying personal information was collected. The repository excludes any information that could reasonably identify individual participants.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Short demo video of neurofeedback training with pacer.
[MOV File, 66430 KB - neuro_v5i1e89256_app1.mov]

Multimedia Appendix 2
Methods, analyses, and figures for feasibility and usability study of self-guided, beta-responsive visual neurofeedback using a consumer electroencephalography headband in a public exhibition.
[PDF File, 1973 KB - neuro_v5i1e89256_app2.pdf]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 114 KB - neuro_v5i1e89256_app3.pdf]

Checklist 2
CONSORT pilot checklist.
[PDF File, 106 KB - neuro_v5i1e89256_app4.pdf]

Checklist 3
TIDieR checklist.

XSL•FO
RenderX

[PDF File, 73 KB - neuro_v5i1e89256_app5.pdf]

Checklist 4
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1077 KB - neuro_v5i1e89256_app6.pdf]

References

1. Sitaram R, Ros T, Stoeckel L, et al. Closed-loop brain training: the science of neurofeedback. Nat Rev Neurosci 2017 Feb;18(2):86-100. [doi: 10.1038/nrn.2016.164] [Medline: 28003656]
2. Bandura A. Self-Efficacy: The Exercise of Control: W.H. Freeman; 1997.
3. Feldman MJ, Bliss-Moreau E, Lindquist KA. The neurobiology of interoception and affect. Trends Cogn Sci 2024 Jul;28(7):643-661. [doi: 10.1016/j.tics.2024.01.009] [Medline: 38395706]
4. Arpaia P, Coyle D, D'Errico G, et al. Virtual Reality Enhances EEG-Based Neurofeedback for Emotional Self-Regulation: Springer-Verlag; 2022:420-431. [doi: 10.1007/978-3-031-15553-6_29]
5. Pan H, Ding P, Wang F, et al. Comprehensive evaluation methods for translating BCI into practical applications: usability, user satisfaction and usage of online BCI systems. Front Hum Neurosci 2024;18:1429130. [doi: 10.3389/fnhum.2024.1429130] [Medline: 38903409]
6. Krause F, Linden DEJ, Hermans EJ. Getting stress-related disorders under control: the untapped potential of neurofeedback. Trends Neurosci 2024 Oct;47(10):766-776. [doi: 10.1016/j.tins.2024.08.007] [Medline: 39261131]
7. Onagawa R, Muraoka Y, Hagura N, Takemi M. An investigation of the effectiveness of neurofeedback training on motor performance in healthy adults: a systematic review and meta-analysis. Neuroimage 2023 Apr 15;270:120000. [doi: 10.1016/j.neuroimage.2023.120000] [Medline: 36870431]
8. Aristizabal JP, Pereira Jr A, Gomes BD, Goulart PRK, De Souza WC, Rocha FA. Neurofeedback beta down training in women with high state-trait anxiety and elevated beta patterns in temporal lobes: a pilot study. NeuroRegulation 2024;11(3):284. [doi: 10.15540/nr.11.3.284]
9. Xia Z, Yang PY, Chen SL, Zhou HY, Yan C. Uncovering the power of neurofeedback: a meta-analysis of its effectiveness in treating major depressive disorders. Cereb Cortex 2024 Jun 4;34(6):bhae252. [doi: 10.1093/cercor/bhae252] [Medline: 38889442]
10. Fernández-Alvarez J, Grassi M, Colombo D, et al. Efficacy of bio- and neurofeedback for depression: a meta-analysis. Psychol Med 2022 Jan;52(2):201-216. [doi: 10.1017/S0033291721004396] [Medline: 34776024]
11. Diotaiuti P, Valente G, Corrado S, et al. Enhancing working memory and reducing anxiety in university students: a neurofeedback approach. Brain Sci 2024 Jun 5;14(6):578. [doi: 10.3390/brainsci14060578] [Medline: 38928578]
12. Whitehead JC, Neeman R, Doniger GM. Preliminary real-world evidence supporting the efficacy of a remote neurofeedback system in improving mental health: retrospective single-group pretest-posttest study. JMIR Form Res 2022 Jul 8;6(7):e35636. [doi: 10.2196/35636] [Medline: 35802411]
13. Askovic M, Soh N, Elhindi J, Harris AWF. Neurofeedback for post-traumatic stress disorder: systematic review and meta-analysis of clinical and neurophysiological outcomes. Eur J Psychotraumatol 2023;14(2):2257435. [doi: 10.1080/20008066.2023.2257435] [Medline: 37732560]
14. Da Silva JC, De Souza ML. Neurofeedback training for cognitive performance improvement in healthy subjects: a systematic review. Psychol Neurosci 2021;14(3):262-279. [doi: 10.1037/pne0000261]
15. Lundqvist M, Miller EK, Nordmark J, Liljefors J, Herman P. Beta: bursts of cognition. Trends Cogn Sci 2024 Jul;28(7):662-676. [doi: 10.1016/j.tics.2024.03.010] [Medline: 38658218]
16. Chen TC, Lin IM. The learning effects and curves during high beta down-training neurofeedback for patients with major depressive disorder. J Affect Disord 2020 Apr 1;266:235-242. [doi: 10.1016/j.jad.2020.01.175] [Medline: 32056883]
17. Wang SY, Lin IM, Fan SY, et al. The effects of alpha asymmetry and high-beta down-training neurofeedback for patients with the major depressive disorder and anxiety symptoms. J Affect Disord 2019 Oct 1;257:287-296. [doi: 10.1016/j.jad.2019.07.026] [Medline: 31302517]
18. Al-Ezzi A, Kamel N, Faye I, Ebenezer EGM. EEG frontal theta-beta ratio and frontal midline theta for the assessment of social anxiety disorder. Presented at: 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE); Aug 21-22, 2020; Penang, Malaysia p. 107-112 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9203789 [accessed 2026-10-01] [doi: 10.1109/ICCSCE50387.2020.9204923]
19. Zhai L, Zhao M, Zhang J, Jamil M, Naz R, Li C. A systematic review of EEG-based biomarkers for depression, anxiety, and bipolar disorder: trends in explainable artificial intelligence (XAI). BMC Psychiatry 2026;26(1):90. [doi: 10.1186/s12888-025-07589-3]
20. Wang H, Mou S, Pei X, et al. The power spectrum and functional connectivity characteristics of resting-state EEG in patients with generalized anxiety disorder. Sci Rep 2025;15(1):5991. [doi: 10.1038/s41598-025-90362-z]
21. Hunkin H, King DL, Zajac IT. Evaluating the feasibility of a consumer-grade wearable EEG headband to aid assessment of state and trait mindfulness. J Clin Psychol 2021 Nov;77(11):2559-2575. [doi: 10.1002/jclp.23189] [Medline: 34061998]

XSL•FO
RenderX

22. Girivirya S, Subekti P, Indramayapanna R, Marjoko S, Sutadi S. Mind full or mindfulness! the effectiveness of mindfulness using brainwave entrainment portable EEG muse s first generation. Presented at: 4th International Conference on Innovation in Education (ICoIE 4); Dec 10-11, 2022. [doi: 10.5220/0012197100003738]
23. Alsayigh HKS, Khidhir ASM. Deep learning-based mental task classification using a muse 2 EEG headset. Presented at: 6th International Conference on Engineering Sciences (ICES2022); Dec 21-22, 2022. [doi: 10.1063/5.0204932]
24. Pantförder M, Hellwig A. Designing Neurogames to Support Patients under Psychotherapy Treatment: Opportunities and Challenges of the MUSE Headband: Science and Technology Publications; 2022:296-305. [doi: 10.5220/0010738500003123]
25. Tandon A, Cobb B, Centra J, et al. Human factors, human-centered design, and usability of sensor-based digital health technologies: scoping review. J Med Internet Res 2024 Nov 15;26:e57628. [doi: 10.2196/57628] [Medline: 39546781]
26. Fairclough SH, Lotte F. Grand challenges in neurotechnology and system neuroergonomics. Front Neuroergon 2020;1:602504. [doi: 10.3389/fnrgo.2020.602504] [Medline: 38234311]
27. Wascher E, Reiser J, Rinkenauer G, et al. Neuroergonomics on the go: an evaluation of the potential of mobile EEG for workplace assessment and design. Hum Factors 2023 Feb;65(1):86-106. [doi: 10.1177/00187208211007707] [Medline: 33861182]
28. Kumar G P, Adarsh A, Ramakrishnan AG. Modulation of EEG by slow-symmetric breathing incorporating breath-hold. IEEE Trans Biomed Eng 2025 Apr;72(4):1387-1396. [doi: 10.1109/TBME.2024.3505963] [Medline: 40030340]
29. Laborde S, Allen MS, Borges U, et al. Effects of voluntary slow breathing on heart rate and heart rate variability: a systematic review and a meta-analysis. Neurosci Biobehav Rev 2022 Jul;138:104711. [doi: 10.1016/j.neubiorev.2022.104711] [Medline: 35623448]
30. Schwerdtfeger AR, Schwarz G, Pfurtscheller K, Thayer JF, Jarczok MN, Pfurtscheller G. Heart rate variability (HRV): from brain death to resonance breathing at 6 breaths per minute. Clin Neurophysiol 2020 Mar;131(3):676-693. [doi: 10.1016/j.clinph.2019.11.013] [Medline: 31978852]
31. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry: Taylor & Francis; 1996:189-194.
32. Hyzy M, Bond R, Mulvenna M, et al. System usability scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022 Aug 18;10(8):e37290. [doi: 10.2196/37290] [Medline: 35980732]
33. Jin W, Zhu X, Qian L, et al. Electroencephalogram-based adaptive closed-loop brain-computer interface in neurorehabilitation: a review. Front Comput Neurosci 2024;18:1431815. [doi: 10.3389/fncom.2024.1431815] [Medline: 39371523]
34. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025 Apr 14;389:e081123. [doi: 10.1136/bmj-2024-081123] [Medline: 40228833]
35. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
36. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
37. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
38. Medvedeva A, Medvedev ON. Mindful Attention and Awareness Scale (MAAS). In: Handbook of Assessment in Mindfulness Research: Springer, Cham; 2025:395-427. [doi: 10.1007/978-3-031-47219-0_13]
39. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research Involving human participants. JAMA 2025;333(1):71-74. [doi: 10.1001/jama.2024.21972]
40. Treves I, Bajwa Z, Greene KD, et al. Consumer-grade neurofeedback with mindfulness meditation: meta-analysis. J Med Internet Res 2025 Apr 17;27:e68204. [doi: 10.2196/68204] [Medline: 40246295]
41. Kober SE, Wood G, Berger LM. Controlling virtual reality with brain signals: state of the art of using VR-based feedback in neurofeedback applications. Appl Psychophysiol Biofeedback 2025 Dec;50(4):593-612. [doi: 10.1007/s10484-024-09677-8] [Medline: 39542998]
42. Ros T, Enriquez-Geppert S, Zotev V, et al. Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). Brain 2020 Jun 1;143(6):1674-1685. [doi: 10.1093/brain/awaa009] [Medline: 32176800]
43. Garcia-Moreno FM, Badenes-Sastre M, Expósito F, Rodriguez-Fortiz MJ, Bermudez-Edo M. EEG headbands vs caps: how many electrodes do I need to detect emotions? The case of the MUSE headband. Comput Biol Med 2025 Jan;184:109463. [doi: 10.1016/j.compbiomed.2024.109463] [Medline: 39608032]
44. Kleeva D, Ninenko I, Lebedev MA. Resting-state EEG recorded with gel-based vs. consumer dry electrodes: spectral characteristics and across-device correlations. Front Neurosci 2024;18:1326139. [doi: 10.3389/fnins.2024.1326139] [Medline: 38370431]
45. Lee Y, Gwon D, Kim K, et al. A comprehensive evaluation framework for consumer-grade EEG devices: signal quality, robustness, and usability. Sci Rep 2026;16(1):8408. [doi: 10.1038/s41598-026-39056-8]
46. Levinton T, Gelech J, Golshan F, Mickleborough M. Experiences of a neurofeedback-based mindfulness meditation intervention for migraine: qualitative study. JMIR Neurotechnol 2025;4:e68369. [doi: 10.2196/68369] [Medline: 41341422]

XSL•FO
RenderX

47. Baek HJ, Chang MH, Heo J, Park KS. Enhancing the usability of brain-computer interface systems. Comput Intell Neurosci 2019;2019:5427154. [doi: 10.1155/2019/5427154] [Medline: 31316556]
48. Chen H, Yun G. Efficacy of brain-computer interface therapy for upper limb rehabilitation in chronic stroke: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2026 Jan 28;28:e79132. [doi: 10.2196/79132] [Medline: 41605490]
49. Brandmeyer T, Delorme A. Meditation and the wandering mind: a theoretical framework of underlying neurocognitive mechanisms. Perspect Psychol Sci 2021 Jan;16(1):39-66. [doi: 10.1177/1745691620917340] [Medline: 32598855]
50. Weibel RP, Kerr JI, Naegelin M, et al. Virtual reality-supported biofeedback for stress management: beneficial effects on heart rate variability and user experience. Comput Human Behav 2023 Apr;141:107607. [doi: 10.1016/j.chb.2022.107607]
51. Karvat G, Schneider A, Alyahyay M, Steenbergen F, Tangermann M, Diester I. Real-time detection of neural oscillation bursts allows behaviourally relevant neurofeedback. Commun Biol 2020 Feb 14;3(1):72. [doi: 10.1038/s42003-020-0801-z] [Medline: 32060396]
52. Tyas AE, Dharma Wibawa A, Purnomo MH. Theta, alpha and beta activity in the occipital based on EEG signals for mental fatigue in high school students. Presented at: 2020 International Conference on Smart Technology and Applications (ICoSTA); Feb 20, 2020 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9070070 [accessed 2026-10-01] [doi: 10.1109/ICoSTA48221.2020.1570614141]
53. Enz N, Schmidt J, Nolan K, et al. Self-regulation of the brain's right frontal beta rhythm using a brain-computer interface. Psychophysiology 2022 Nov;59(11):e14115. [doi: 10.1111/psyp.14115] [Medline: 35652562]
54. Martin T, Holliday E, Okhio C, et al. States, traits, and the resting state EEG task aftereffect. Int J Psychophysiol 2025 Apr;210:112523. [doi: 10.1016/j.ijpsycho.2025.112523] [Medline: 39933625]
55. Zaghloul H, Fanous K, Ahmed L, et al. Digital health literacy in patients with common chronic diseases: systematic review and meta-analysis. J Med Internet Res 2025 Aug 25;27:e56231. [doi: 10.2196/56231] [Medline: 40854231]
56. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun 16;8(2):e9. [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
57. Self-guided, beta-responsive visual neurofeedback with a consumer EEG headband in a public exhibition: data and analysis code. OSF. URL: https://osf.io/zfyd7 [accessed 2026-10-02]

Abbreviations

BCI: brain-computer interface
CONSORT: Consolidated Standards of Reporting Trials
EEG: electroencephalography
MAAS-State: Mindful Attention Awareness Scale–State
MCAR: missing completely at random
NFB: neurofeedback
PTSD: posttraumatic stress disorder
SDK: software development kit
SUS: System Usability Scale
TIDieR: Template for Intervention Description and Replication
VAS: Visual Analog Scale
VN: visual neurofeedback only
VN+BP: visual neurofeedback plus breathing pacer

Edited by S Brini; submitted 09.Dec.2025; peer-reviewed by G Guerrero-Contreras; revised version received 15.Sep.2026; accepted 16.Sep.2026; published 03.Nov.2026.

Please cite as:
Baumgartner M, Weibel R, Hoesli T, von Wangenheim F, Ollier J
Self-Guided Beta-Responsive Visual Neurofeedback in a Public Exhibition: Feasibility and Usability Randomized Controlled Trial
JMIR Neurotech 2026;5:e89256
URL: https://neuro.jmir.org/2026/1/e89256
doi:10.2196/89256

© Melanie Baumgartner, Raphael Weibel, Tobias Hoesli, Florian von Wangenheim, Joseph Ollier. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 3.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete

XSL•FO
RenderX

bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX