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

Original Papers

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

[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.

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

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).

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

[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].

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]

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]

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]

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

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

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

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

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

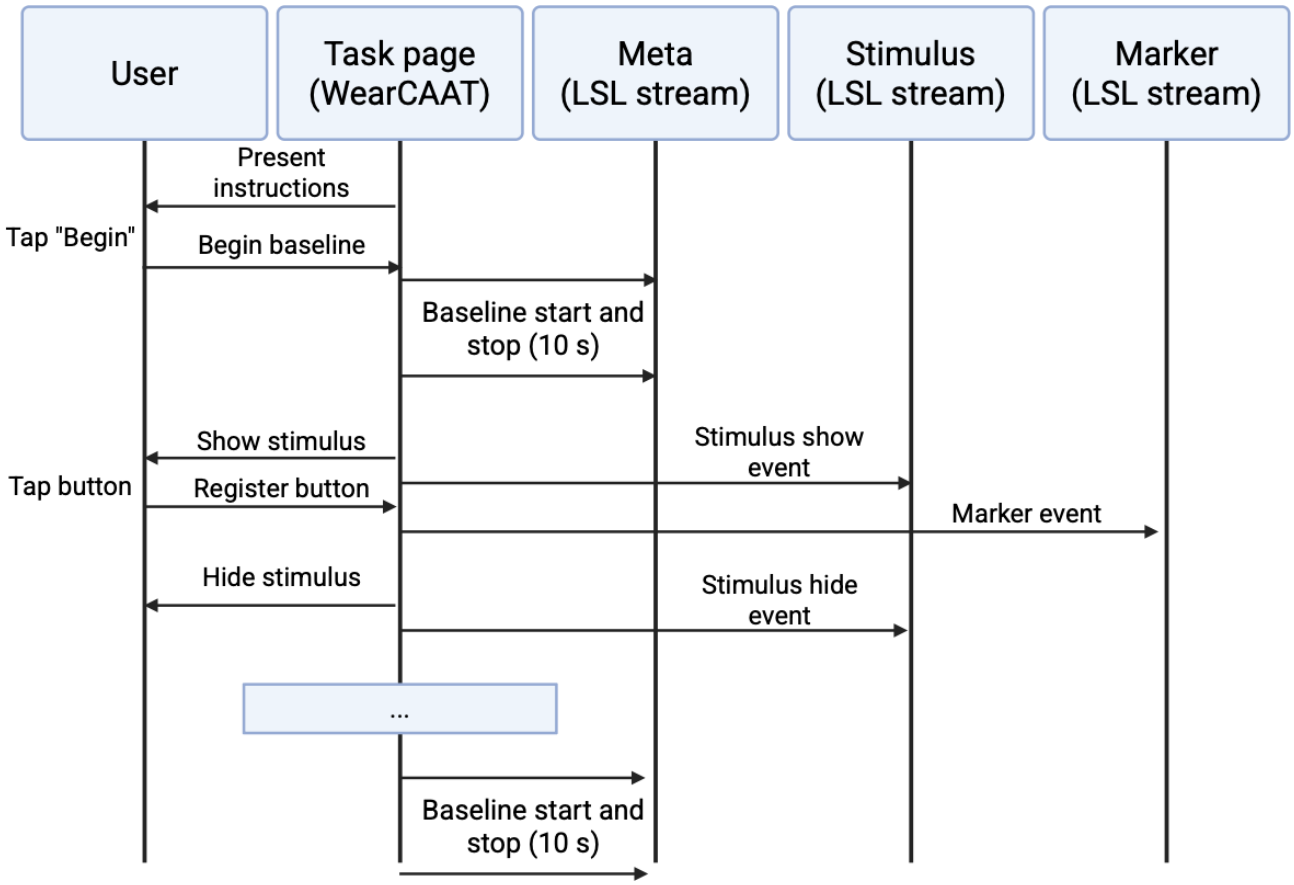
WearCAAT implements task-based neurocognitive monitoring, wherein participants perform 1 of 11 built-in tasks to elicit known cognitive effects in different domains, including attention, vigilance, working and episodic memory, response inhibition, set shifting, and conflict resolution. Currently implemented 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.

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)

Data Collection

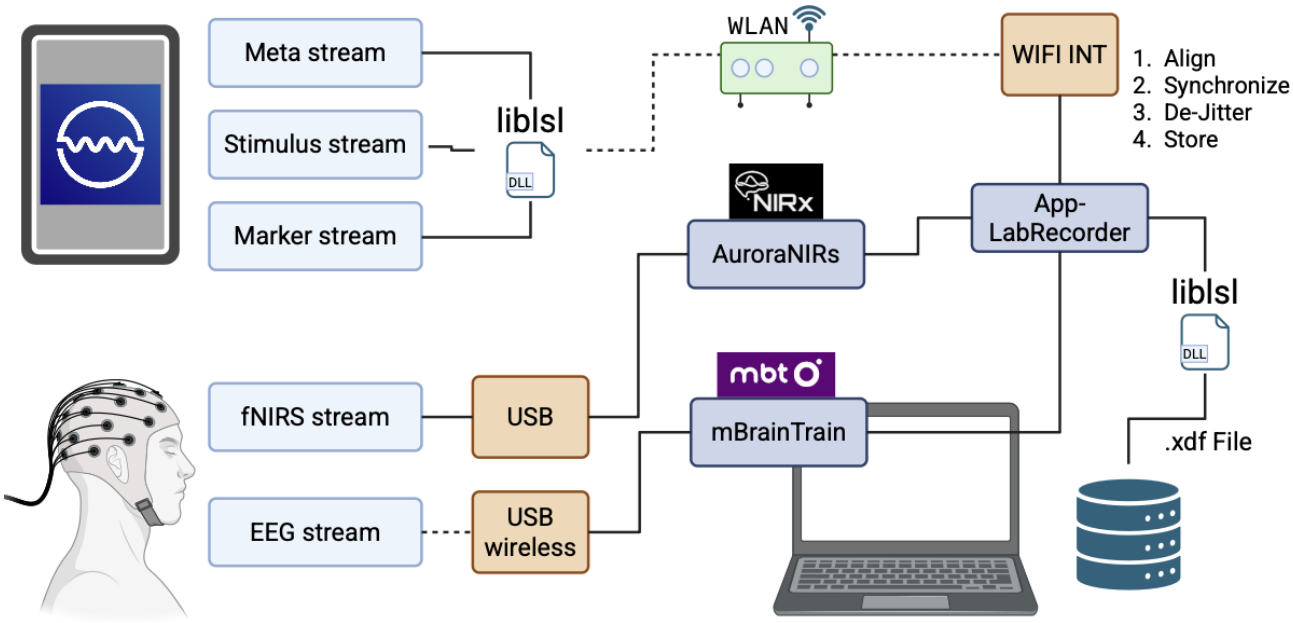
We administered all 11 abbreviated tasks to participants via WearCAAT in sequence, where the first 2 were the resting tasks (eyes opened and eyes closed) and the remaining 9 tasks were presented in randomized order (Table 3). All participants completed the entire task battery built in WearCAAT in 1 sitting 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.

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.

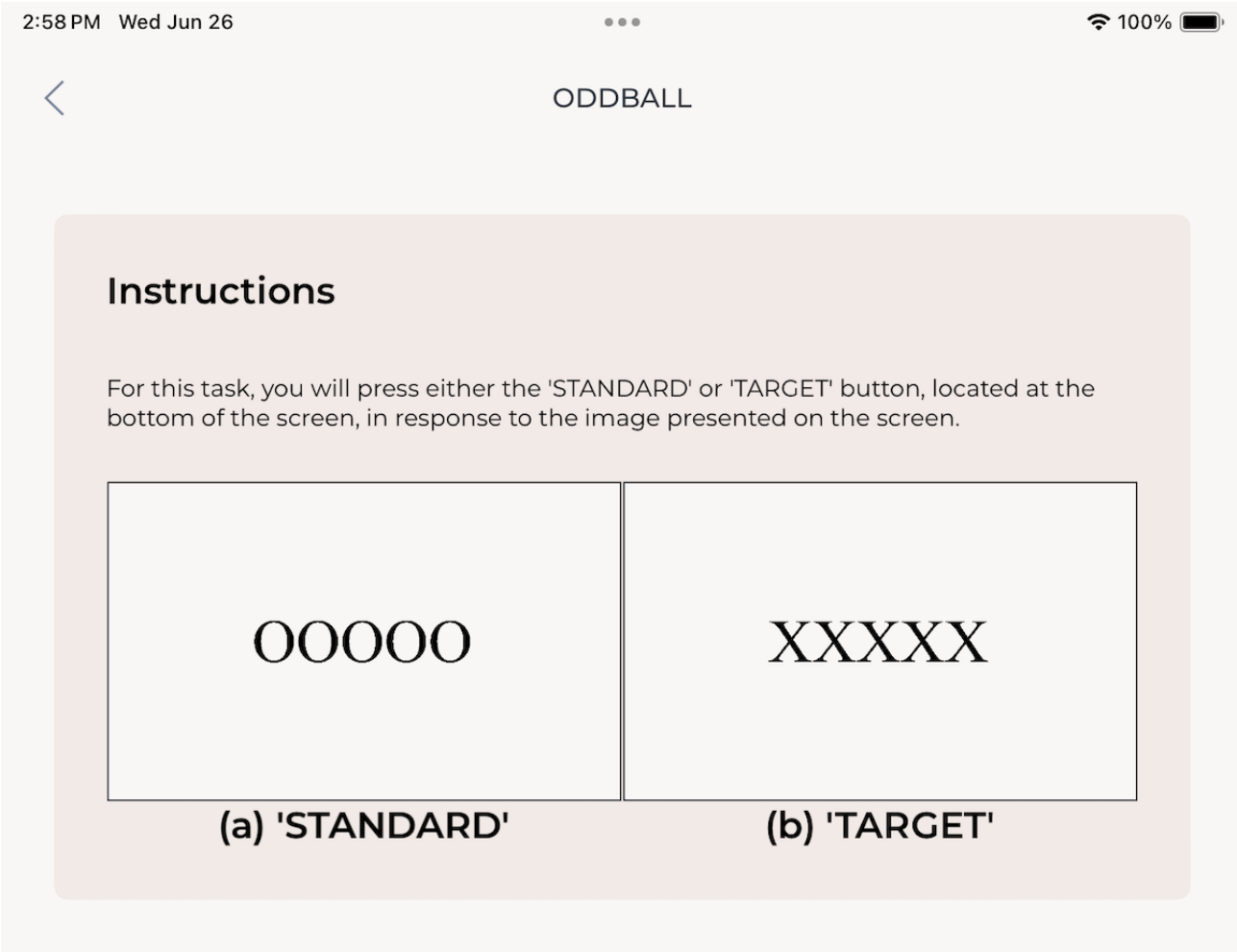


Behavioral responses were recorded in WearCAAT where task events (eg, stimulus presentation times, user responses via button presses, etc) were time stamped in WearCAAT and wirelessly streamed through LSL to a laptop (Windows 10) running App-LabRecorder [41]. Concurrent EEG and fNIRS data were also streamed 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.

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.

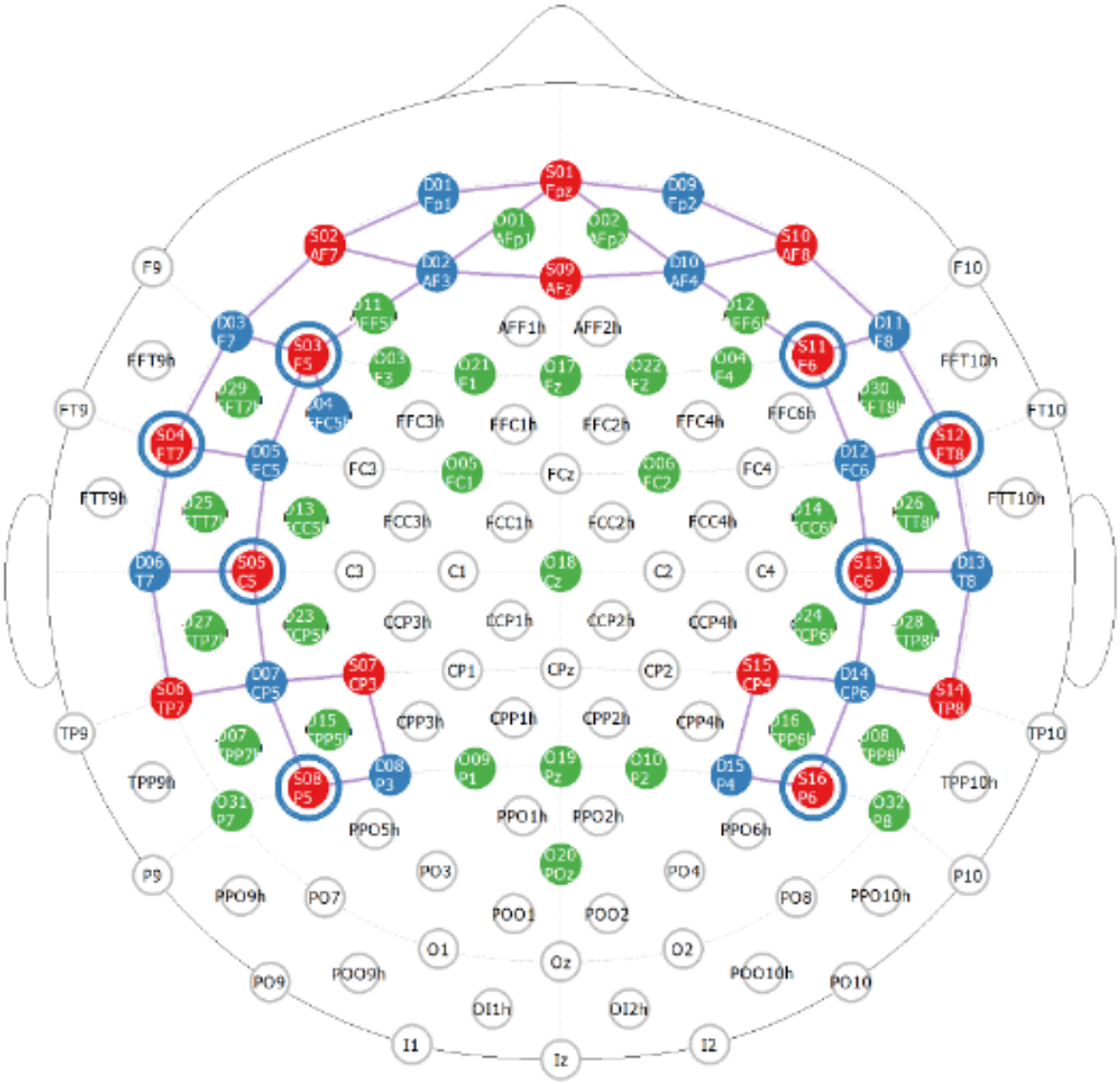


Signal Processing

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.

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

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

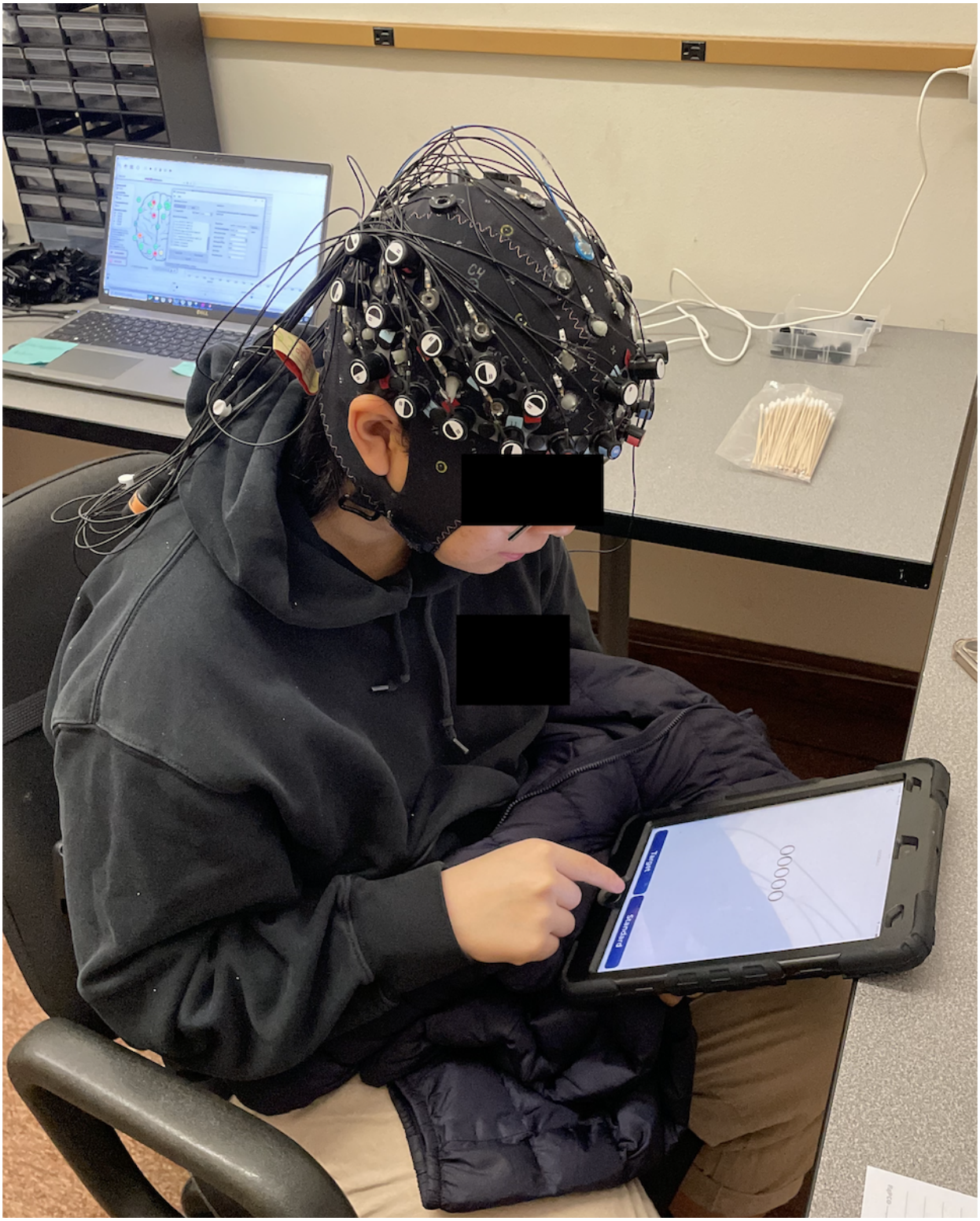
We processed our fNIRS data by removing motion artifacts and baseline shifts through wavelet and spline filters [44]. We further removed physiological signals, cardiac, respiratory, and Mayer waves, with a finite impulse response low-pass filter using the cutoff frequency 0.08 Hz [51]. Finally, we converted 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.

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



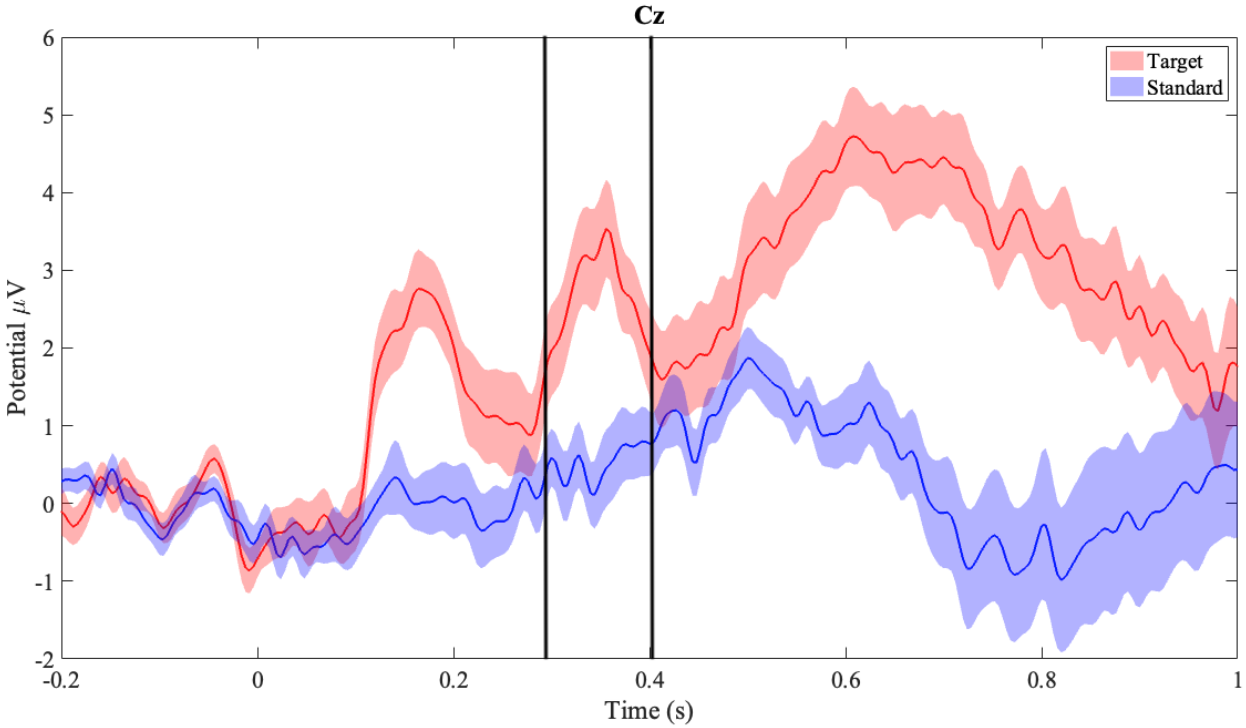
User Response Times

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.

EEG: Extracted P300

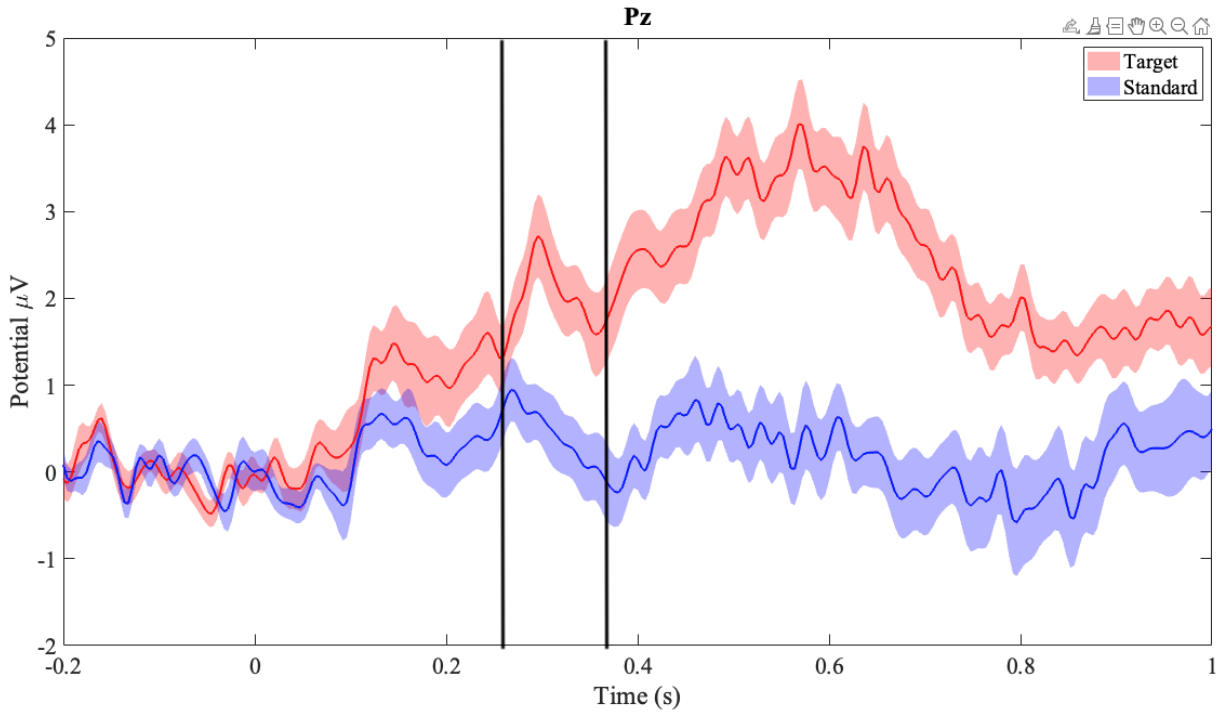
All participant-averaged (SEM) ERP waveforms obtained using WearCAAT in an iPad for the abbreviated visual oddball task for target (red) and standard (blue) stimulus in the parietal midline (Pz) and central midline (Cz) regions are presented in Figure 6, separately. Our results showed higher P300 amplitude 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.

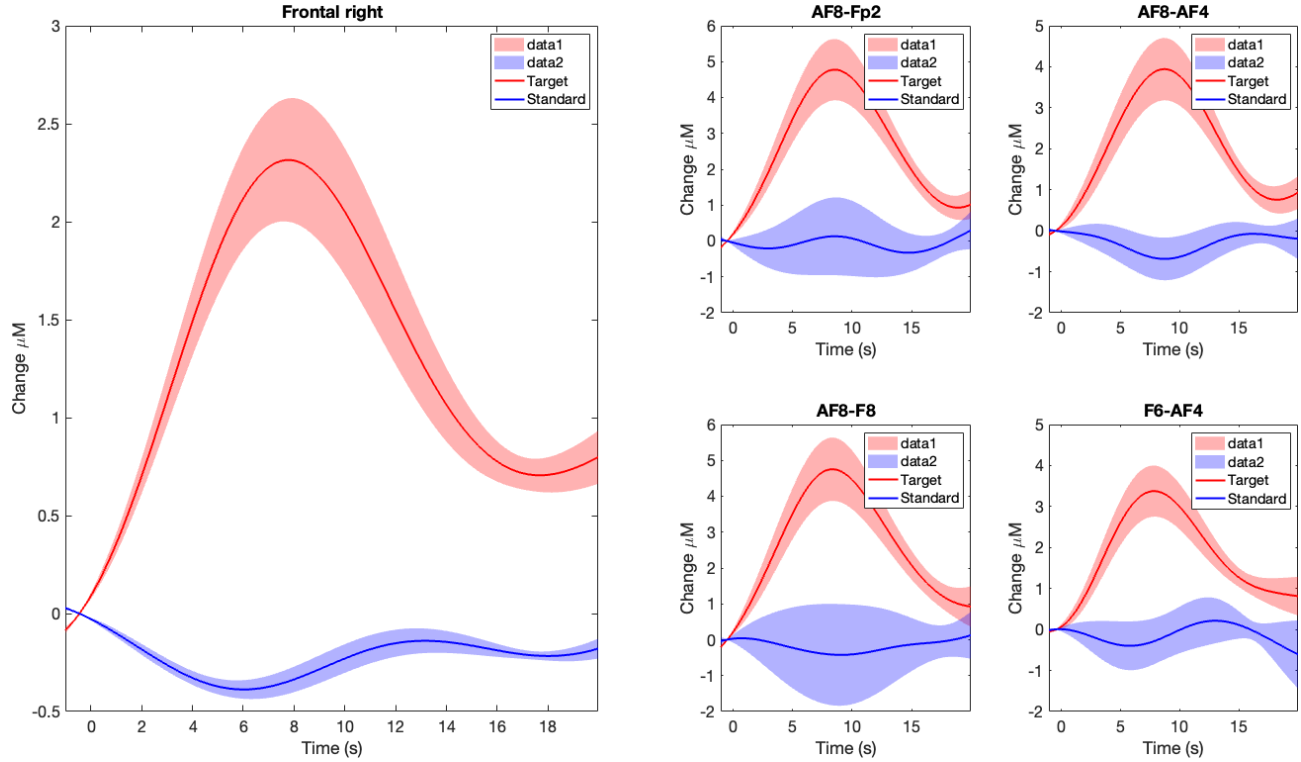
Figure 7. Mean amplitudes across participants in the parietal midline (Pz). Clouded regions represent the standard error of the mean. Red represents target (infrequent) stimulus and blue represents (frequent) stimulus responses.



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.



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

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]

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]

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

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

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

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

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.

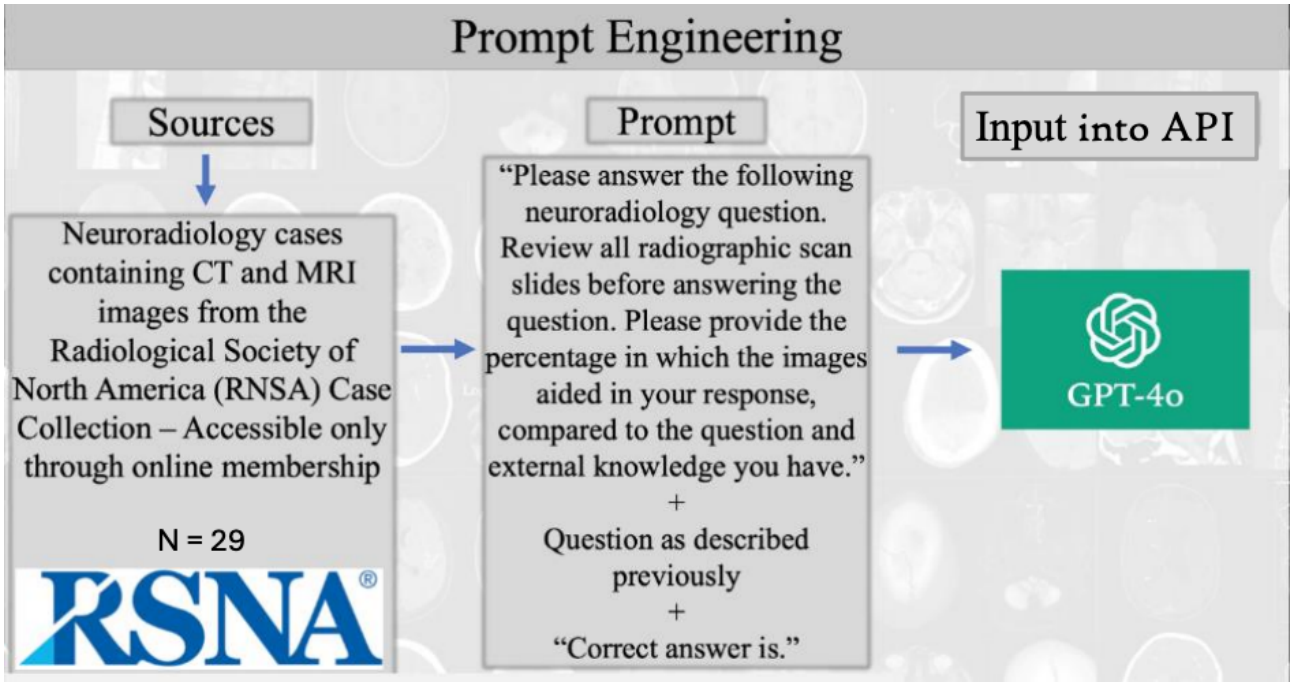


Figure 2 is a representative infectious disease case (Table S6.1 in Multimedia Appendix 3, Case #14: Neurocysticercosis) from the RSNA Case Collection, 1 of 29 adult brain and central nervous system pathology cases (Figure S5.1 in Multimedia Appendix 2, Table S6.1 in Multimedia Appendix 3) used in this 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).

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.

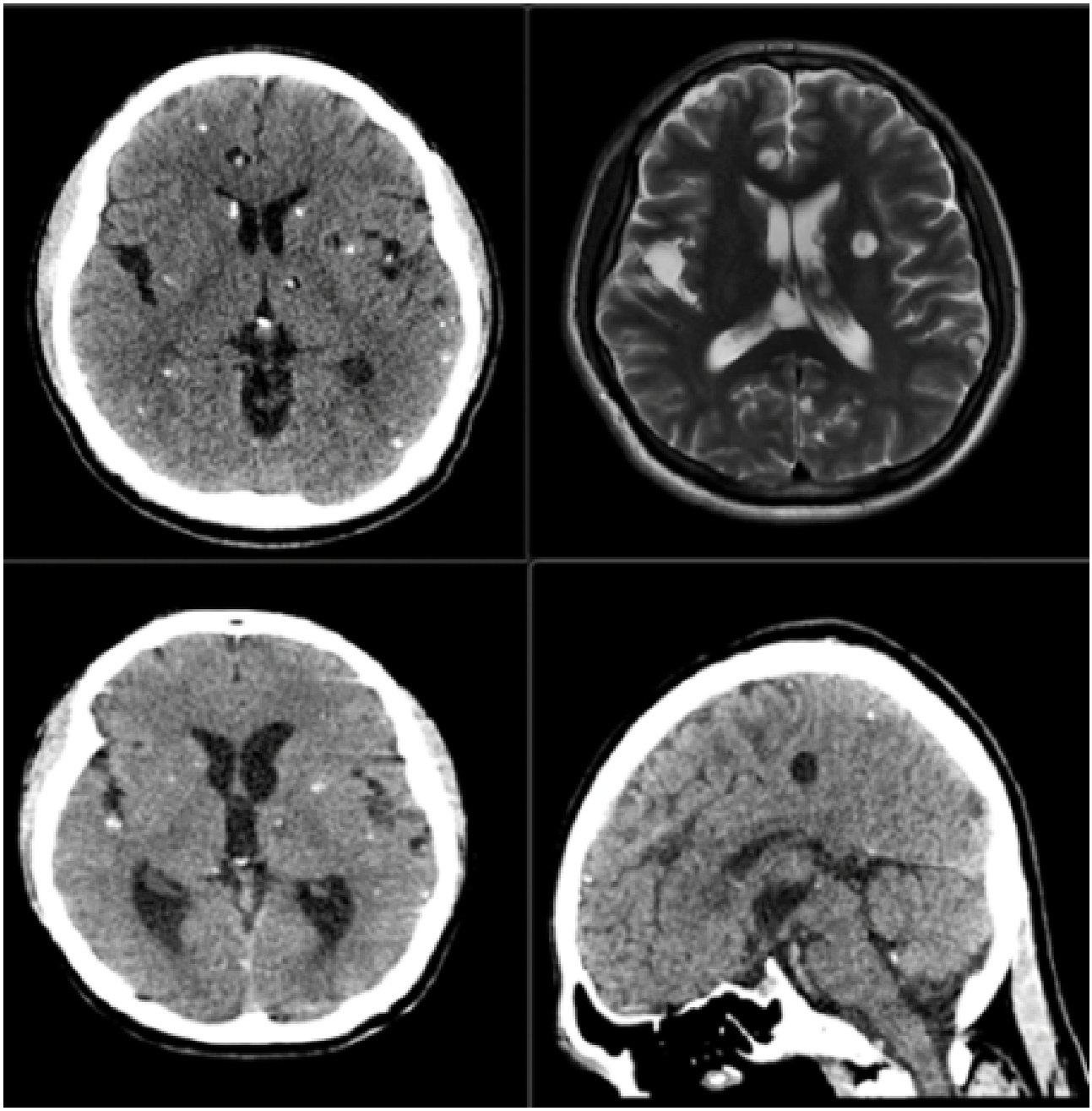
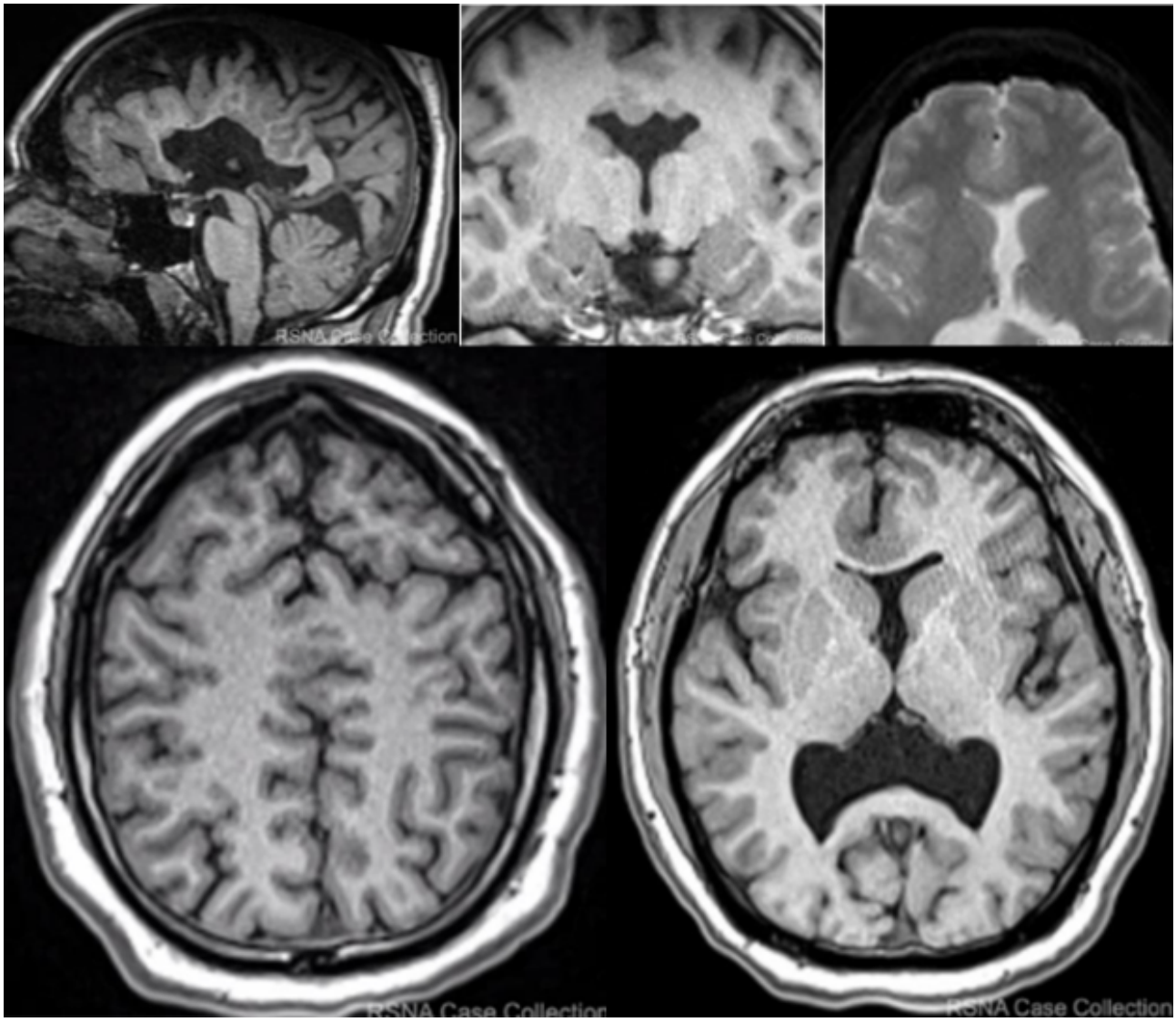


Figure 3 is a complex neuroradiology case (Table S6.1 in Multimedia Appendix 3, Case #19: Syntelencephaly) from this July 2024 cross-sectional study (Figure S5.1 in Multimedia Appendix 2) evaluating GPT-4V's diagnostic performance on 29 adult brain and central nervous system pathology cases from the 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%),

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

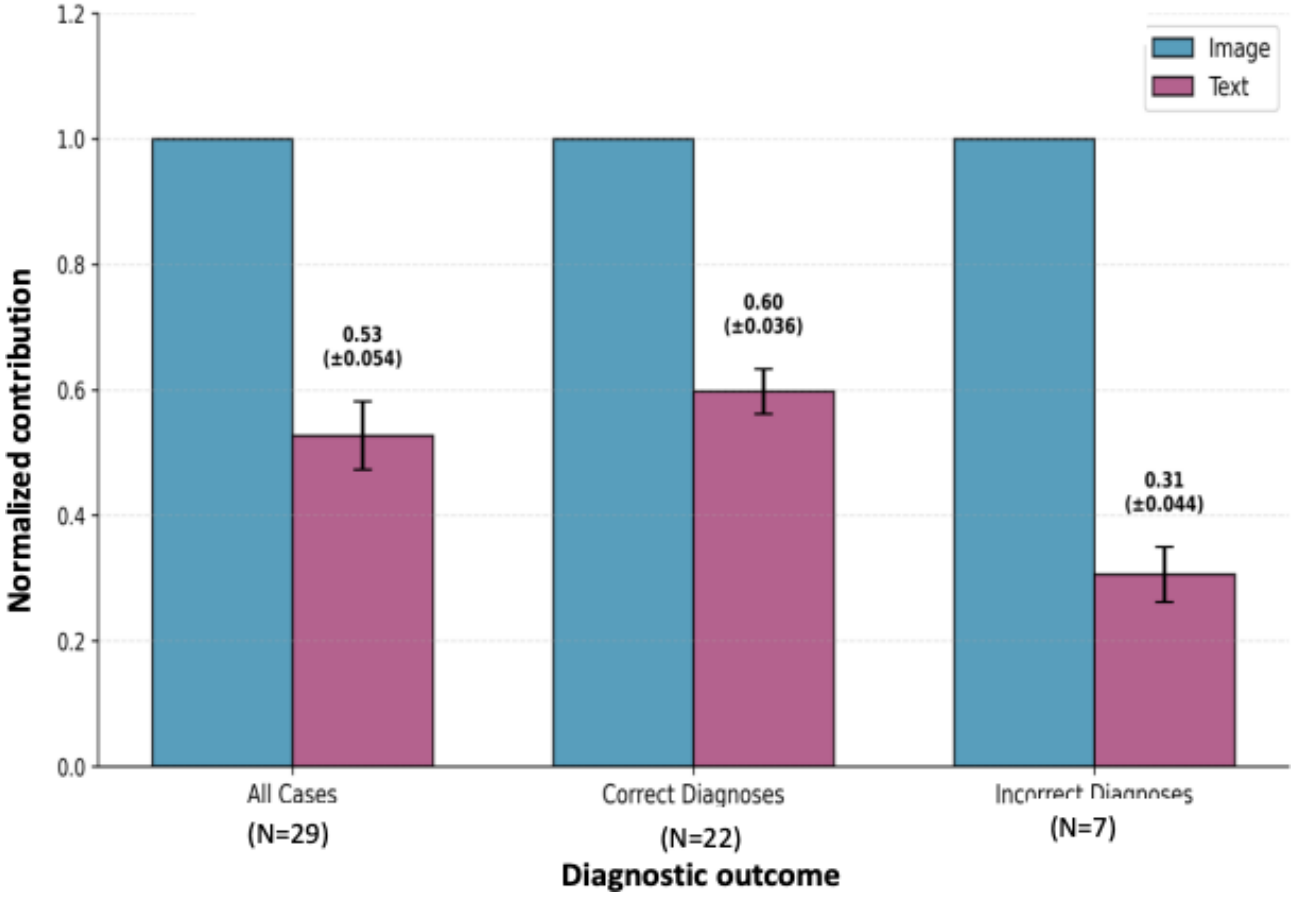
GPT-4V correctly diagnosed 22 out of 29 neuroradiology cases, yielding 76% accuracy (95% CI 57.9% - 87.8% by Wilson method), significantly above the 25% expected by chance (z=6.33; P<.001; Table S1 in Multimedia Appendix 1, Table S4.1 in Multimedia Appendix 9, and Table S6.1 in Multimedia Appendix 3). 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

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

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

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.

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.

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

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]

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]

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

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

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