Validation of a Mobile, Sensor-based Neurobehavioral Assessment With Digital Signal Processing and Machine-learning Analytics

Kelly L. Sloane, MD, Joel A. Mefford, PhD,† Zilong Zhao, PhD,‡ Man Xu, MA,‡ Guifeng Zhou, BS,‡ Rachel Fabian, MS,§ Amy E. Wright, BA,§ and Shenly Glenn, BS‡*

Background: The Miro Health Mobile Assessment Platform consists of self-administered neurobehavioral and cognitive assessments that measure behaviors typically measured by specialized clinicians.

Objective: To evaluate Miro Health Mobile Assessment Platform's concurrent validity, test–retest reliability, and mild cognitive impairment (MCI) classification performance.

Method: Sixty study participants were evaluated with Miro Health version V.2. Healthy controls (HC), amnestic MCI (aMCI), and nonamnestic MCI (naMCI) ages 64–85 were evaluated with version V3. Additional participants were recruited at Johns Hopkins Hospital to represent clinic patients, with wider ranges of age and diagnosis. In all, 90 HC, 21 aMCI, 17 naMCI, and 15 other cases were evaluated with V.3. Concurrent validity of the Miro Health variables and legacy neuropsychological test scores was assessed with Spearman correlations. Reliability was quantified with the scores' intraclass correlations. A machine-learning algorithm combined Miro Health variable scores into a Risk score to differentiate HC from MCI or MCI subtypes.

Results: In HC, correlations of Miro Health variables with legacy test scores ranged 0.27–0.68. Test–retest reliabilities ranged 0.25–0.79, with minimal learning effects. The Risk score differentiated individuals with aMCI from HC with an area under the receiver operator curve (AUROC) of 0.97; naMCI from HC with an AUROC of 0.80; combined MCI from HC with an AUROC of 0.89; and aMCI from naMCI with an AUROC of 0.83.

Conclusion: The Miro Health Mobile Assessment Platform provides valid and reliable assessment of neurobehavioral and cognitive status, effectively distinguishes between HC and MCI, and differentiates aMCI from naMCI.

Key Words: mobile technology, machine learning, mild cognitive impairment

(*Cogn Behav Neurol* 2022;00:000–000)

aMCI = amnestic mild cognitive impairment. **AUROC** = area under the receiver operator curve. **HC** = healthy controls. **MCI** = mild cognitive impairment. **MMSE** = Mini-Mental State Examination. **naMCI** = nonamnestic mild cognitive impairment.

Mild cognitive impairment (MCI) is a complex clinical disorder with many possible causes, including neurodegenerative diseases, traumatic head injury, stroke or other vascular disorders, cancer, medication effect, and other medical issues. It is estimated that 20–40% of MCI cases will progress to dementia (Roberts and Knopman, 2013); other cases of MCI may remain static or resolve. For progressive cases, the early identification of MCI and timely support of patients and their care communities may help prolong patient independence, reduce health care costs, and advance research toward effective therapies (Lee and Chan, 2020; Wittenberg et al, 2019).

To aid in the early identification and continual monitoring of individuals with MCI, diagnostic tools are needed that can detect subtle deviations from healthy neurobehavioral and cognitive function and be administered repeatedly and reliably (Sabbagh et al, 2020). Though cognitive assessments that are performed and scored by licensed clinical neuropsychologists are the gold standard for the diagnosis of MCI (Petersen et al, 2018), these clinician-administered assessments are impractical for population-based care due to expense and lack of availability.

Brief screens like the Montreal Cognitive Assessment (Nasreddine et al, 2005) and the Mini-Mental State Examination (MMSE; Folstein et al, 1975) can be rapidly administered and interpreted by individuals without

Received for publication March 2, 2021; accepted August 7, 2021.
From the *Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania; †Department of Neurology, University of California, Los Angeles, California; ‡Miro Health Inc., San Francisco, California; and §Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland.
Associate Editor Victor W. Henderson oversaw the review process for this article.
Supported in part by Miro Health Inc.
The Miro Health Mobile Assessment Platform is a commercial product of Miro Health Inc. Z.Z., M.X., G.Z., and S.G. are employed by Miro Health Inc. J.A.M. is a consultant for and holds equity in Miro Health Inc. The remaining authors declare no conflict of interest.
Correspondence: Shenly Glenn, BS, 260 King St., Unit 1505, San Francisco, California 94107 (email: shenly@mirohealth.com).
Supplemental digital content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.cogbehavneurol.com.

neuropsychology training (Seo et al, 2011), but their ability to detect early MCI varies depending on patient population and is limited by practice effect. Other technologies, such as gaze trackers, have emerged as a potential rapid assessment (Oyama et al, 2019), but they require external hardware, and their diagnostic utility and reliability remain unknown.

Here, we present validity studies for the *Miro Health Mobile Assessment Platform*—a new sensor-based neurobehavioral and cognitive platform that automatically measures and quantifies neurobehavioral and cognitive functions for the detection and characterization of MCI. This platform offers self-administered assessments in the form of interactive modules and self-report questionnaires. The Miro Health Mobile Assessment Platform consists of multiple components: Study HQ automates administrative and participant management, Miro Mobile provides self-administered assessments and questionnaires, and Open Insights provides automated data processing and analyses.

This paper presents two versions of the Miro Health Mobile Assessment Platform that were sequentially validated: Version 2 (V.2) and Version 3 (V.3). Improvements made to V.2 assessment module tutorials, stimuli display times, and guided prompts changed the characteristics of the data collected, necessitating a new mobile version (V.3), an updated V.3-matched reference data set, and separate V.3 validity analyses. The aims of this study were to:

- demonstrate the validity of the Miro Health Mobile Assessment Platform V.2 and V.3 by measuring the correlation of scores between legacy neuropsychological tests and the interactive modules of the Miro Health Mobile Assessment Platform;
- evaluate the stability of the Miro Health Mobile Assessment Platform's measurement properties by test–retest reliability; and
- investigate the predictive validity of the Miro Health Mobile Assessment Platform's machine-generated amnestic MCI (aMCI) Risk score to differentiate healthy performance from aMCI performance and aMCI performance from nonamnestic MCI (naMCI) performance.

METHOD

V.2 Recruitment

Study participants were recruited from neurology clinics at Johns Hopkins Hospital in Baltimore, Maryland; from participant pools at contract research organization sites; and from the general public through advertisements in newspapers. Screening measures included demographics, medical history (self-reported), the Telephone Interview for Cognitive Status (Knopman et al, 2010; Seo et al, 2011), the Geriatric Depression Scale (Yesavage and Sheikh, 1986), the Mayo-Portland Modified Inventory (Lezak, 1987; Malec and Lezak, 2008), and the MMSE (Chapman et al, 2016; O'Bryant et al, 2008).

Inclusion criteria for the healthy controls (HC) were Telephone Interview for Cognitive Status score ≥33, English speaker before age 5 years, and high school or equivalent education. Inclusion criteria for the individuals with MCI were MCI diagnosis per the American Academy of Neurology clinical criteria (Petersen et al, 2018); and MMSE score ≥20 or medical records with a history of diagnosis of MCI, neurodegenerative disorder, or vascular disorder with cognitive impairment. Exclusion criteria for the HC and MCI groups were evidence of a comorbid neurologic disease; use of drugs known to affect cognition; uncorrected vision or hearing impairment; and history of cancer, substance abuse, or axis 1 disorder.

Sixty study participants were evaluated using Miro Health Mobile Assessment Platform version V.2: 33 HC and 27 participants affected by neurologic conditions, including 17 with MCI. Participants with neurologic conditions who were affected by a neurologic condition other than, or in addition to, MCI were evaluated and classified as *Affected* in our analysis. After updates to the Miro Health Mobile Assessment Platform to version V.3, new data sets were collected and analyzed separately from the Miro Health Mobile Assessment Platform V.2 data.

V.3 Recruitment

Study participants for evaluation with Miro Health Mobile Assessment Platform version V.3 were recruited from three contract research organization sites and from the Stroke Cognitive Outcomes & Recovery Laboratory at Johns Hopkins University. (See Table S1 of the supplementary digital content [SDC; Supplemental Digital Content 1, http://links.lww.com/CBN/A116] for details.) Study participants at the research sites were recruited to represent HC, aMCI, and naMCI in the age range of 64–85 years. The test–retest analysis was restricted to HC, and the predictive validity analysis was restricted to HC, aMCI, and naMCI participants. As detailed in Table S1, 89 HC, 12 aMCI, and 13 naMCI were recruited from the three research sites.

Study participants at Johns Hopkins were recruited to represent the patient population, research cohorts, and research questions at the Stroke Cognitive Outcomes & Recovery Lab, with a wider range of ages as low as 48 years and a wide range of diagnoses and medical histories: nine individuals with aMCI, four with naMCI, six with a history of left-hemisphere stroke, three with primary progressive aphasia, three with primary progressive aphasia and frontotemporal dementia, two with Parkinson disease, and one with Alzheimer disease. All of the study participants from Johns Hopkins were included in the concurrent validity analysis. The aMCI and naMCI participants who were recruited from Johns Hopkins were included in the predictive validity analyses.

The data from Johns Hopkins includes eight study participants with age <64: aMCI (49 years); naMCI (60 and 63 years); and other diagnoses (48, 53, 57, 59, and 61 years). To increase sample sizes, we included these study participants who were <64 years in the analyses, as

well as an additional 61-year-old control who was evaluated at study site 2.

The study protocol was approved by the Johns Hopkins University Institutional Review Board (protocol 00088299) and the New England Institutional Review Board (protocols 120180208, 120180211, 120180209, and 12080253) and was performed according to the ethical guidelines of the Declaration of Helsinki and its later amendments. All individuals provided informed written consent before enrolling in the study.

Computerized Cognitive Assessment

The Miro Health Mobile Assessment Platform can be downloaded as an application to personal devices such as tablets and smartphones from the App Store or Google Play and features neurobehavioral and cognitive assessments that can be self-administered. Assessment batteries can be tailored from more than 40 interactive modules and hundreds of questionnaires.

This study limited its scope to Miro Health modules that are analogous with legacy neuropsychological tests (Table 1) but were redesigned to capture high-fidelity data such as movement, speech, and language (Glenn and Mefford, 2019b) through sensors that are built in to personal devices (eg, touch screen, camera, or microphone of a tablet or smartphone). This approach allows patient neurobehaviors to be recorded and saved for the machine-based processing and analysis that is required to quantify brain behaviors.

Each module within the assessment battery of this study was self-administered and scored automatically with the Miro Health Mobile Assessment Platform data processing pipeline and machine-learning engine. Modules adapt to patient performance and are equipped with patented and proprietary anti-cheat mechanisms (Glenn and Mefford, 2019a; Glenn et al, 2017). The Miro Health Mobile Assessment Platform has been designated as a *Breakthrough Device* by the US Food and Drug Administration.

During the assessments measuring verbal learning and memory and digit span forward and backward, the Miro Health Mobile Assessment Platform recorded each user's touchscreen interactions and spoken responses. Then, proprietary digital signal processing (artificial intelligence) extracted features or variables from the user recordings (Glenn and Mefford, 2019b). In addition to familiar variables like out-of-set and repeat errors, commission and omission errors, and total test time, hundreds of other categorical and continuous variables were extracted and calculated, such as initiation latency, variability in fine motor movement, language, and vocal acoustic features. In this study, we focused on variables from tests that were delivered through the Miro Health Mobile Assessment Platform that have analogous scores from legacy tests (Table 1) for use in a concurrent validity analysis, but we did consider some novel Miro Health variable scores for the test–retest analysis and for evaluation of differences in the variable scores between HC and MCI cases.

Data analysis also included the weighted combination of Miro Health variable scores to make an aMCI Risk score that accurately differentiates MCI cases from HC. Analysis included comparison of an individual's Miro Health module scores to existing reference sets of HC, aMCI, and naMCI users of the Miro Health Mobile Assessment Platform.

Statistical Analysis

All of the statistical analyses were performed using R software, R 4.0.3 (www.R-project.org).

Concurrent Validity

Legacy testing was conducted with two test batteries that consisted of the same assessments but were ordered differently. The battery was alternated upon odd-number enrollee triggers (eg, enrollees 1 and 2 received battery order A, enrollees 3 and 4 received battery order B, enrollees 5 and 6 received battery order A, etc.). Miro Health Mobile Assessment Platform testing was conducted with a single, static test battery.

Miro Health Mobile Assessment Platform modules generate a unique instance at each user session in order to deter cheating and learning effects. Each unique instance refers to a version of the test that is newly generated and captures equivalent variables to other versions in order to enable comparative analysis. The exceptions to this patented design are Category Fluency, Letter Fluency, and Picture Description, which deliver nonequivalent and noninterchangeable instances such as Letter Fluency's cycling letters (*a*, *s*, *d*, *t*, etc.).

Missing data, either Miro Health module scores or comparator test scores (or both) for a particular variable, were excluded from the concurrent validity study. Standardization of scores occurred via a two-step process to ensure that the standardized scores for the HC had mean values of 0 and standard deviations of 1 and an overall normal distribution:

- Scores were quantile normalized or had quantiles in the empirical distribution of scores mapped to quantiles of a standard normal distribution.
- The mean (normalized) score for the HC was subtracted from each score, and the resulting centered scores were rescaled by the standard deviation of the scores among the HC.

For modules measuring Category Fluency and Letter Fluency, scores based on test sessions with different stimuli were standardized separately.

Spearman correlations were used to analyze the concurrent validity between the individual variable scores that were produced by the Miro Health (V.3) Mobile Assessment Platform and the legacy neuropsychological test scores that were administered and scored by licensed clinical neuropsychologists (Signorell et al, 2020). A smaller set of Miro Health scores from the earlier Miro Health Mobile Assessment Platform V.2 assessment were compared to legacy test scores with Spearman correlations as well, with results shown in Table S2 of the SDC

TABLE 1. Examples of Miro Health Variables Derived From Miro Health Modules and Comparator Neuropsychological Tests

Comparator Tests and Tasks	Miro Health Module	Compared Variable
D-KEFS Category Fluency[1]	Categories	Category Fluency: total correct
D-KEFS Design	Chart A Course	Design Fluency A: total correct
D-KEFS Design	Chart A Course	Design Fluency B Switching: total correct
D-KEFS Phonemic fluency	Lucky Letters	Phonemic Fluency: total correct
Finger Tapping Test[2]	Take Flight	Finger Tapping: mean taps dominant
Finger Tapping Test	Take Flight	Finger Tapping: mean taps nondominant
Finger Tapping Test	Take Flight	Finger Tapping: greatest taps dominant
HVLT–R Immediate Recall[3]	Spy Games	VLM: largest correct set spoken
HVLT–R Delayed Cued Recall	Spy Report	DVLM cued recall: total correct
HVLT–R Delayed Cued Recall	Spy Report	DVLM cued recall: total errors
HVLT–R Delayed Free Recall	Spy Report	DVLM free recall: largest correct set spoken
HVLT–R Delayed Free Recall	Spy Report	DVLM free recall: total errors
Iowa Trail-Making Test[4]	Bolt Bot	Iowa Trail-Making Test Part A: total time
Iowa Trail-Making Test	Bolt Bot	Iowa Trail-Making Test Part B: total time
Iowa Trail-Making Test	Bolt Bot	Iowa Trail-Making Test Part B: total correct
WAIS–IV Coding[5]	Treasure Tomb	Coding: total correct
WAIS–IV Digit Span	Hungry Bees	Digit Span Backward: longest correct span, spoken
WAIS–IV Digit Span	Hungry Bees	Digit Span Backward: longest correct span, touch
WAIS–IV Digit Span	Hungry Bees	Digit Span Forward: longest correct span, spoken
WAIS–IV Digit Span	Hungry Bees	Digit Span Forward: longest correct span, touch
WMS–II Spatial Span[6]	Follow the Glow	Spatial Span Forward: total correct
WMS–II Spatial Span	Follow the Glow	Spatial Span Backward: total correct
Choice reaction time	Two of A Kind	No comparator test used
Picture Description[7]	Speak the Scene	Picture Description: fundamental frequency
Picture Description	Speak the Scene	Picture Description: total number of content units
Picture Description	Speak the Scene	Picture Description: total prepositions

Comparator tests and tasks refer to legacy neuropsychological tests and tasks that are administered in person by a licensed clinical neuropsychologist. Miro Health modules are interactive assessments that are self-administered on mobile devices such as an iPhone or iPad on which the Miro Health Mobile Assessment Platform has been installed. Compared variables are examples of numerical scores that are calculated from Miro Health modules.

D-KEFS = Delis-Kaplan Executive Function System. **DVLM** = delayed verbal learning and memory. **HVLT–R** = Hopkins Verbal Learning Test—Revised. **VLM** = verbal learning and memory. **WAIS–IV** = Wechsler Adult Intelligence Scale—Fourth Edition. **WMS–III** = Wechsler Memory Scale—Third Edition.

[1]Delis DC, Kaplan E, Kramer JH. 2001. *The Delis-Kaplan Executive Function System.* San Antonio, Texas: Psychological.

[2]Ashendorf L, Horwitz JE, Gavett BE. 2015. Abbreviating the Finger Tapping Test. *Arch Clin Neuropsychol.* 30:99–104. doi:10.1093/arclin/acu091

[3]Brandt J, Benedict RHB. 1997. *Hopkins Verbal Learning Test—Revised.* Odessa, Florida: Psychological Assessment Resource.

[4]Partington JE, Leiter RG. 1949. Partington's Pathways Test. *Psychological Service Center Journal.* 1:11–20.

[5]Wechsler D. 2008. *WAIS–IV Technical and Interpretive Manual.* San Antonio, Texas: Psychological.

[6]Wechsler D. 1997. *WMS–III: Wechsler Memory Scale Administration and Scoring Manual.* San Antonio, Texas: Psychological.

[7]Kaplan EF, Goodglass H, Weintraub S. 1983. The Boston Naming Test. 2nd ed. Philadelphia, Pennsylvania: Lea & Febiger.

(http://links.lww.com/CBN/A116). For comparison, Pearson correlations were also calculated for the Miro Health Mobile Assessment Platform V.3 variables and their analogous variables on legacy testing and are shown with the Spearman correlations in Table S3 of the SDC (http://links.lww.com/CBN/A116).

Test–Retest Reliability Across Three Time Points

Participant performance with the Miro Health Mobile Assessment Platform was assessed over three time points at 1-week intervals. The same battery of tests was administered on the Miro Health Mobile Assessment Platform in the same order at each time point. A unique instance of each module was generated at each participant session (Glenn and Mefford, 2020) with the exception of Category Fluency, Letter Fluency, and Picture Description. Three versions of each of these modules were presented in the same order across time points to all of the participants.

The test–retest reliability of the scores over three time points was quantified by intraclass correlation, which is an estimate of the variance in the distribution of scores that can be explained by individual participant random effects in a mixed effect model (Gamer et al, 2019). In addition to the reliability of measurements, we analyzed learning effects or trends across the three time points using two measures named trends and progression. Estimated *trends* were calculated for each variable score as the slope of a linear model for score versus observation number. *Progression* was calculated as the difference in slopes between the first and second assessments and between the second and third assessments. The statistical significance of the trends and progression was tested with Wald tests.

Predictive Validity Analyses

Data collected from the concurrent validity study and time point 1 of the test–retest reliability study using V.3 of the Miro Health Mobile Assessment Platform were included in the predictive validity study. The HC had to have completed >85% of the assessment.

Diagnostic group assignments were made by an unaffiliated licensed clinical neuropsychologist in private practice who had access to the participants' legacy neuropsychological test results (Table 1), Mayo-Portland Modified Inventory scores, Geriatric Depression Scale scores, and demographics and medical history. The 38 individuals with MCI were then assigned by the neuropsychologist to an aMCI subtype or naMCI subtype depending on their performance in the domains of learning and memory. MCI subtype analyses were not conducted for V.2 participants due to low total MCI numbers.

Miro Health Mobile Assessment Platform data scientists developed diagnostic classification models using sparse ordinal logistic regression (Wurm et al, 2021) to find an optimal weighted combination of scores for the assignment of study participants to their neuropsychologist-designated groups. Inputs to the classification model came from a broad set of Miro Health modules, many of which lack legacy neuropsychological test comparators, such as continuous timing variables, movement variables, vocal characteristics, speech production, and language (Table 1). Variables were quantile normalized, and missing values were imputed (Hastie and Mazumder, 2021).

The regression algorithm in the OrdinalNet package v.2.9 (Wurm et al, 2021) was used to (a) fit an ordinal logistic regression model with elastic net penalties; (b) identify a subset of variables that are most relevant for classification; and (c) find a weighted combination of the selected scores that differentiates the aMCI, naMCI, and HC groups. The OrdinalNet parameters, α (the ratio of L1 to L2 penalties) and λ (the scaling factor for the penalties), were optimized with five-fold cross-validation for classification performance as measured by area under the receiver operator curve (AUROC) for differentiation of HC from aMCI. Classification models were not adjusted for age, sex, or education due to the small sample size and the risk of overfitting the model. The resulting weighted combination of input scores is the aMCI Risk score, or Risk score.

The algorithm was evaluated by leave-one-out cross-validation. The resulting aMCI Risk score formula was applied to the left-out individual's data set in order to obtain an out-of-sample predicted Risk score. This procedure was followed for each observation in turn, and the AUROC was calculated using the set of out-of-sample Risk scores as a classifier to differentiate the groups. For comparison, individual variable AUROCs for group separation were also calculated. For data collected with Miro Health Mobile Assessment Platform V.2, MCI Risk scores were generated using the same approach of elastic net (Zou and Hastie, 2005) but with only two groups, HC and MCI, whereas three groups were analyzed in the Miro Health Mobile Assessment Platform V.3 analysis.

RESULTS

Concurrent Validity

A total of 160 participants were included in the concurrent validity analyses. Table 2 shows the numbers of study participants with different diagnoses. Forty-six participants were evaluated using Miro Health Mobile Assessment Platform V.2 (19 HC, 27 Affected), and 114 were evaluated using Miro Health Mobile Assessment Platform V.3 (64 HC, 18 aMCI, 17 naMCI, 15 Affected). Half of the participants received the Miro Health modules first; the other half received the legacy neuropsychological tests first, alternating by even and odd enrollee numbers. All 160 participants were assessed with the comparator neuropsychological tests.

Data sets from the two versions of Miro Health Mobile Assessment Platform were analyzed separately. Study participants with different diagnoses were analyzed jointly. Twenty-eight participants were excluded for incomplete data sets (22 incomplete legacy neuropsychological data sets, six incomplete Miro Health data sets).

Table 3 shows the comparator neuropsychological tests and the paired Miro modules that were considered for the concurrent validity analysis. Spearman correlations from Miro Health Mobile Assessment Platform V.2 (SDC Table S2, http://links.lww.com/CBN/A116) and Miro Health Mobile Assessment Platform V.3 (Table 3) ranged from 0.36 to 0.60 and 0.27 to 0.68, respectively. SDC Table S3 (http://links.lww.com/CBN/A116) includes the results of both the Spearman and Pearson correlations for the concurrent validity analysis of Miro Health Mobile Assessment Platform V.3 and demonstrates that the two results are strongly in agreement. Data for V.3 concurrent validity are included in SDC Data Set 1 (http://links.lww.com/CBN/A117).

Test–Retest Reliability Across Three Time Points

A total of 59 HC were included in the test–retest reliability analysis. Six participants were excluded from the study for failing Miro Health Mobile Assessment Platform's proprietary anti-cheat mechanism. Thirty-three HC completed the test–retest reliability study using Miro Health Mobile Assessment Platform V.2, and 26 HC completed the study using Miro Health Mobile Assessment Platform V.3 (Table 2). Half of the HC were unsupervised and half were supervised during their evaluation with the Miro Health Mobile Assessment Platform, alternating by even and odd enrollee numbers.

Intraclass correlation estimates between Miro Health Mobile Assessment Platform V.3 variables with comparator clinical neuropsychological test variables ranged from 0.25 to 0.79 (Table 4). SDC Table S4 (http://links.lww.com/CBN/A116) shows the estimated trends and progression for these variables. No variable

TABLE 2. Participant Demographics by Analysis and Miro Health Version

Miro Health Version	Analysis	Group	Number of Participants	Female/Male	Mean Age (Range) in Years
V.2	Concurrent validity	HC	19	15/4	73.2 (65–89)
		Affected†	27	16/11	78.5 (65–95)
	Test–retest	HC	33	25/8	74.2 (65–81)
	Predictive validity	HC	32	24/8	73.2 (65–89)
		MCI	17	8/9	70.4 (65–92)
	All participants	HC	33	25/8	
		Affected	27	16/11	
	All participants	Total	60	41/19	
V.3	Concurrent validity	HC	64	36/28	70.3 (61–82)
		aMCI	18	5/13	70.0 (49-83)
		naMCI	17	6/11	71.1 (60-82)
		Affected‡	15	5/10	66.1 (48–79)
	Test–retest	HC	26	16/10	72.0 (64–82)
	Predictive validity	HC	65	36/29	70.2 (61–82)
		aMCI	21	6/15	71.1 (49–83)
		naMCI	17	6/11	71.1 (60–82)
	All participants	HC	90	51/39	
		aMCI	21	6/15	
		naMCI	17	6/11	
		Affected	15	5/10	
		Total	143	68/75	

The characteristics of the study participants whose data were used in each analysis are shown. Data were collected using two versions of the Miro Heath Mobile Assessment Platform, Version 2 (V.2) before 2019 and Version 3 (V.3) since 2019. The data collected by different module versions cannot be mixed. Overlapping but not identical sets of study participants were used for each analysis: concurrent validity, test–retest, and predictive validity. The concurrent validity analysis required a complete battery of comparator assessments and the Miro Health modules, whereas for the test–retest analysis, participants needed to complete three assessments with the same Miro Health module at weekly intervals.

†To indicate the functional status of the study participants used in the concurrent validity analysis with Miro Health Mobile Assessment Platform V.2, study participants were classified as either HC or Affected, and counts and age ranges for each group are shown. Individuals were designated *Affected* to indicate that the study participant had another neurologic diagnosis. The HC and Affected study participants were analyzed jointly.

‡To indicate the functional status of the study participants used in the concurrent validity analysis with Miro Health Mobile Assessment Platform V.3, study participants were classified as either HC, aMCI, naMCI, or Affected, and counts and age ranges for each group are shown. Individuals were designated *Affected* to indicate that the study participant had another neurologic diagnosis. The Affected group contained six participants with a history of left-hemisphere stroke, three with primary progressive aphasia, three with primary progressive aphasia and frontotemporal dementia, two with Parkinson disease, and one with Alzheimer disease. In the supplementary data set for the concurrent validity analysis of Miro Health V.3, participants in each of the case groups are labeled "Affected." Study participants in the HC and case groups were analyzed jointly.

aMCI = amnestic mild cognitive impairment. **HC** = healthy controls. **MCI** = mild cognitive impairment. **naMCI** = nonamnestic mild cognitive impairment.

demonstrated significant trends or progression. Data for V.3 test–retest reliability are included in SDC Data Set 1 (http://links.lww.com/CBN/A117).

Predictive Validity

A total of 152 participants were included in the predictive validity study: 49 V.2 (32 HC and 17 MCI) and 103 V.3 (65 HC, 21 aMCI, and 17 naMCI). In the V.3 data set, when the HC, aMCI, and naMCI labels were assigned according to the individuals' scores on the MMSE, the Miro Health Mobile Assessment Platform's aMCI Risk score yielded a classification performance measured as an AUROC of 0.83. When the licensed neuropsychologist assigned the HC, aMCI, and naMCI labels, the Miro Health Mobile Assessment Platform's aMCI Risk score yielded an AUROC of 0.97 for the classification of aMCI versus HC and an AUROC of 0.89 for the classification of MCI (aMCI + naMCI) versus HC. The Risk score classified aMCI versus naMCI study participants with an AUROC of 0.83. The V.2 MCI Risk score classified the MCI participants and HC with an AUROC of 0.94.

For comparison, individual variable AUROCs for group separation (Table 5) were also calculated. SDC Table S5 (http://links.lww.com/CBN/A116) shows the means and standard deviations of non-normalized scores in the HC and aMCI groups, along with *P* values from the Wilcoxon rank-sum tests and *t* tests.

DISCUSSION

We reported the results of three analyses that examined the validity and reliability of the Miro Health Mobile Assessment Platform to classify an individual's neurobehavioral

TABLE 3. Concurrent Validity of Comparator Legacy Neuropsychological Assessment Variables Versus Miro Health Mobile Assessment Platform Variables (Version 3)

Variable	Spearman (95% CI)	*P*
Digit Span Backward: longest correct span, touch[1]	0.68 (0.54, 0.79)	< 0.001**
Phonemic Fluency: total correct[2]	0.66 (0.53, 0.75)	< 0.001**
Category Fluency: total correct[2]	0.66 (0.54, 0.76)	< 0.001**
Digit Span Forward: longest correct span, touch[1]	0.57 (0.43, 0.69)	< 0.001**
Trail-Making Test Part B: total time[3]	0.59 (0.45, 0.71)	< 0.001**
Design Fluency A: total correct[2]	0.48 (0.31, 0.63)	< 0.001**
Trail-Making Test Part A: total time[3]	0.52 (0.37, 0.65)	< 0.001**
Coding: total correct[1]	0.40 (0.21, 0.57)	< 0.001**
Finger Tapping: mean taps nondominant[4]	0.35 (0.15, 0.52)	< 0.001**
Spatial Span Backward: total correct[5]	0.33 (0.15, 0.50)	< 0.001**
Finger Tapping: mean taps dominant[4]	0.32 (0.11, 0.49)	0.003*
Spatial Span Forward: total correct[5]	0.27 (0.08, 0.45)	0.005*
Design Fluency B Switching: total correct[2]	0.28 (–0.02, 0.54)	0.064
DVLM free recall: largest correct set spoken[6]	0.36 (0.14, 0.55)	0.002*
DVLM free recall: total errors[6]	0.30 (0.07, 0.51)	0.012
VLM: largest correct set spoken[6]	0.30 (0.10, 0.47)	0.004*

Spearman correlation estimates and *P* values for tests to reject the assumption that the true correlations are zero.

*Significant at $P < 0.01$.

**Significant at $P < 0.001$.

DVLM = Delayed Verbal Learning and Memory. **VLM** = Verbal Learning and Memory.

[1]Wechsler D. 2008. *WAIS–IV Technical and Interpretive Manual.* San Antonio, Texas: Psychological.

[2]Delis DC, Kaplan E, Kramer JH. 2001. *The Delis-Kaplan Executive Function System.* San Antonio, Texas: Psychological.

[3]Partington JE, Leiter RG. 1949. Partington's Pathways Test. *Psychol Serv Center J.* 1:11–20.

[4]Ashendorf L, Horwitz JE, Gavett BE. 2015. Abbreviating the Finger Tapping Test. *Arch Clin Neuropsychol.* 30:99–104. doi:10.1093/arclin/acu091

[5]Wechsler D. 1997. *WMS–III: Wechsler Memory Scale Administration and Scoring Manual.* San Antonio, Texas: Psychological.

[6]Brandt J, Benedict RHB. 1997. *Hopkins Verbal Learning Test—Revised.* Odessa, Florida: Psychological Assessment Resource.

and cognitive status. The Miro Health Mobile Assessment Platform was able to differentiate HC from individuals with MCI within this data set. This study demonstrated substantial differences in AUROCs when labels were assigned using scores from the MMSE (AUROC 0.83) versus from a licensed clinical neuropsychologist (AUROC 0.97), which is a consequence of the limited measurement properties of brief screening tools. The modest concurrent validity correlations evidenced in this study were expected given the modest test–retest reliability of the comparator tests themselves (Iverson, 2001; Snow et al, 1989).

Unlike typical studies for evaluating test–retest reliability that include only two time points, our study analyzed data that were collected at three time points. We chose this design in order to examine the effects that increasing familiarity with mobile interfaces and assessment paradigms may exhibit on performance scores. Familiarity effects were evaluated by testing for trends, or average changes in scores, on successive assessments and by testing for progression, or differences, in the changes in performance between the first and second assessments and between the second and third assessments. Neither type of familiarity effect was significant for the scores that we evaluated in this study.

A machine-learning–derived Risk score based on mobile, self-administered neurobehavioral and cognitive modules distinguished individuals with aMCI from HC with an AUROC of 0.97; individuals with naMCI and HC were separated with an AUROC of 0.80, and the combined MCI group (aMCI + naMCI) was separated from HC with an AUROC of 0.89. The Risk score distinguished aMCI from naMCI cases with an AUROC of 0.83.

The strong performance of the Miro Health Mobile Assessment Platform's machine-learning–derived Risk score is the result of a combination of factors within the platform. First, in addition to cognitive data, behavioral data such as movement, speech, language, and voice are quantified and incorporated into the Miro Health Risk score. Second, machine-driven scoring removes the variability, errors, and subjectivity that can be found in human administration and scoring of neuropsychological assessments. Third, the platform has a built-in quality management system to ensure data validity at all stages of assessment, processing, and scoring.

TABLE 4. Intraclass Correlations for a Sample of Comparator Legacy Neuropsychological Assessment Variables and Miro Health Mobile Assessment Platform Variables

Variable	ICC (95% CI)
Category Fluency: total correct[1]	0.30 (0.09, 0.53)
Choice reaction time: mean RT, complex†	0.60 (0.41, 0.75)
Choice reaction time: mean RT, simple†	0.41 (0.19, 0.61)
Coding: total correct[2]	0.61 (0.42, 0.76)
Design Fluency A: total correct[1]	0.53 (0.33, 0.70)
Design Fluency B Switching: total correct	0.79 (0.65, 0.89)
Digit Span Forward: longest correct span, spoken[2]	0.52 (0.28, 0.72)
Digit Span Forward: longest correct span, touch	0.67 (0.50, 0.80)
Digit Span Backward: longest correct span, spoken spoken	0.35 (0.10, 0.58)
Digit Span Backward: longest correct span, touch	0.58 (0.23, 0.78)
VLM: largest correct set spoken[3]	0.56 (0.34, 0.74)
DVLM cued recall: total correct	0.65 (0.35, 0.82)
DVLM cued recall: total errors	0.50 (0.14, 0.73)
DVLM free recall: largest correct set spoken	0.57 (0.24, 0.78)
DVLM free recall: total errors	0.54 (0.19, 0.76)
Finger Tapping: mean taps dominant[4]	0.42 (0.21, 0.63)
Finger Tapping: mean taps nondominant	0.41 (0.19, 0.61)
Finger Tapping: greatest taps dominant	0.25 (0.03, 0.48)
Trail-Making Test Part A: total time[5]	0.74 (0.60, 0.85)
Trai-Making Test Part B: total time[5]	0.76 (0.62, 0.86)
Phonemic Fluency: total correct[1]	0.75 (0.61, 0.85)
Picture Description: fundamental frequency[6]	0.77 (0.64, 0.87)
Picture Description: total number of content units	0.68 (0.51, 0.81)
Picture Description: total prepositions	0.41 (0.19, 0.61)
Spatial Span Forward: total correct[7]	0.40 (0.19, 0.61)
Spatial Span Backward: total correct	0.25 (0.04, 0.48)

All ICCs were significantly different than 0 at the 0.05 level.
†No comparator assessment.
DVLM = Delayed Verbal Learning and Memory. **ICC** = intraclass correlation. **VLM** = Verbal Learning and Memory.
[1]Delis DC, Kaplan E, Kramer JH. 2001. *The Delis-Kaplan Executive Function System.* San Antonio, Texas: Psychological.
[2]Wechsler D. 2008. *WAIS–IV Technical and Interpretive Manual.* San Antonio, Texas: Psychological.
[3]Brandt J, Benedict RHB. 1997. *Hopkins Verbal Learning Test—Revised.* Odessa, Florida: Psychological Assessment Resource.
[4]Ashendorf L, Horwitz JE, Gavett BE. 2015. Abbreviating the Finger Tapping Test. *Arch Clin Neuropsychol.* 30:99–104. doi:10.1093/arclin/acu091
[5]Partington JE, Leiter RG. 1949. Partington's Pathways Test. *Psychol Serv Center J.* 1:11–20.
[6]Kaplan EF, Goodglass H, Weintraub S. 1983. The Boston Naming Test. 2nd ed. Philadelphia, Pennsylvania: Lea & Febiger.
[7]Wechsler D. 1997. *WMS–III: Wechsler Memory Scale Administration and Scoring Manual.* San Antonio, Texas: Psychological.

This paper did not present analyses of supervised versus unsupervised assessment results, the effectiveness of Miro Health's anti-cheat detection, the capture and creation of new variable types, the value of categorical versus continuous data in MCI detection, or identification of specific modules for their ability to detect and subtype MCI. Additional data collected in Miro Health's modular platform has the potential to enhance our ability to detect and monitor MCI through quantification of nuanced neurobehavioral factors.

Study Limitations

Study limitations include small sample sizes, the lack of diverse diseases and symptom severities, and the lack of longitudinal data. Larger sample sizes would allow more stable estimates of the weights for combined scores and the evaluation of score performance as a classifier by cross-validation and hold-out sets. Data from diverse diseases and symptom severities would help demonstrate the Miro Health Mobile Assessment Platform's real-world ability to differentiate aMCI from

TABLE 5. AUROC of the MCI Risk Score and 27 Miro Health Mobile Assessment Platform Variables on the Differentiation of HC From MCI Participants By Diagnosis

Variable	aMCI vs HC	naMCI vs HC	MCI vs HC	aMCI vs naMCI
Category Fluency: total correct[1]	0.92	0.69	0.82	0.72
Choice reaction time: mean RT, complex†	0.75	0.62	0.69	0.65
Choice reaction time: mean RT, simple†	0.74	0.56	0.66	0.73
Coding: total correct[2]	0.91	0.68	0.80	0.75
Design Fluency A: total correct[1]	0.81	0.63	0.72	0.69
Design Fluency B Switching: total correct	0.81	0.64	0.70	0.77
Trail-Making Test Part A: total time[3]	0.94	0.69	0.83	0.84
Trail-Making Test Part B: total correct	0.89	0.69	0.80	0.75
Trail-Making Test Part B: total time	0.85	0.72	0.80	0.70
Digit Span Forward: longest correct span, spoken[2]	0.85	0.55	0.70	0.83
Digit Span Forward: longest correct span, touch	0.78	0.68	0.73	0.65
Digit Span Backward: longest correct span, spoken	0.82	0.57	0.70	0.81
Digit Span Backward: longest correct span, touch	0.95	0.77	0.85	0.73
VLM: largest correct set spoken[4]	0.75	0.60	0.68	0.74
DVLM cued recall: total correct	0.70	0.57	0.65	0.65
DVLM cued recall: total errors	0.52	0.74	0.61	0.69
DVLM free recall: largest correct set spoken	0.71	0.67	0.69	0.56
DVLM free recall: total errors	0.51	0.57	0.52	0.58
Finger Tapping: mean taps dominant[5]	0.73	0.58	0.66	0.67
Finger Tapping: mean taps nondominant	0.67	0.64	0.66	0.57
Finger Tapping: greatest taps dominant	0.67	0.57	0.63	0.63
Phonemic Fluency: total correct[1]	0.88	0.79	0.84	0.67
Picture Description: fundamental frequency[6]	0.52	0.62	0.57	0.56
Picture Description: total number of content units	0.82	0.63	0.74	0.74
Picture Description: total prepositions	0.84	0.66	0.76	0.75
Spatial Span Forward: total correct[7]	0.63	0.55	0.59	0.58
Spatial Span Backward: total correct	0.70	0.54	0.62	0.68
MCI Risk score	0.97	0.80	0.89	0.83

†No comparator assessment.

aMCI = amnestic mild cognitive impairment. **AUROC** = area under the receiver operator curve. **DVLM** = Delayed Verbal Learning and Memory. **HC** = healthy controls. **MCI** = mild cognitive impairment. **naMCI** = nonamnestic mild cognitive impairment. **VLM** = Verbal Learning and Memory.

[1]Delis DC, Kaplan E, Kramer JH. 2001. *The Delis-Kaplan Executive Function System.* San Antonio, Texas: Psychological.

[2]Wechsler D. 2008. *WAIS–IV Technical and Interpretive Manual.* San Antonio, Texas: Psychological.

[3]Partington JE, Leiter RG. 1949. Partington's Pathways Test. *Psychol Serv Center J.* 1:11–20.

[4]Brandt J, Benedict RHB. 1997. *Hopkins Verbal Learning Test—Revised.* Odessa, Florida: Psychological Assessment Resource.

[5]Ashendorf L, Horwitz JE, Gavett BE. 2015. Abbreviating the Finger Tapping Test. *Arch Clin Neuropsychol.* 30:99–104. doi:10.1093/arclin/acu091

[6]Kaplan EF, Goodglass H, Weintraub S. 1983. The Boston Naming Test. 2nd ed. Philadelphia, Pennsylvania: Lea & Febiger.

[7]Wechsler D. 1997. *WMS–III: Wechsler Memory Scale Administration and Scoring Manual.* San Antonio, Texas: Psychological.

other similar conditions. Longitudinal data collection would allow retrospective relabeling of data once symptoms had progressed and the accuracy of diagnosing MCI's underlying pathology improved. Longitudinal data would also facilitate a move away from the use of threshold scores for the diagnosis of MCI and toward more personalized measures of change over time.

CONCLUSION

This study evaluated the ability of a self-administered mobile neurobehavioral and cognitive assessment platform combined with automated machine-learning scoring to evaluate neuropsychological status compared with legacy neuropsychological testing and to differentiate individuals with MCI from HC. Mobile, self-administered, and auto-scored platforms such as the Miro Health Mobile Assessment Platform have the potential to support parallel data collection and analyses so as to improve health care, accelerate research, and reduce costs.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful feedback. Their contributions have greatly improved this paper.

REFERENCES

Ashendorf L, Horwitz JE, Gavett BE. 2015. Abbreviating the Finger Tapping Test. *Arch Clin Neuropsychol.* 30:99–104. doi:10.1093/arclin/acu091

Brandt J, Benedict RHB. 1997. *Hopkins Verbal Learning Test—Revised.* Odessa, Florida: Psychological Assessment Resource.

Chapman KR, Bing-Canar H, Alosco ML, et al. 2016. Mini Mental State Examination and Logical Memory scores for entry into Alzheimer's disease trials. *Alzheimers Res Ther*. 8:1–11. doi:10.1186/s13195-016-0176-z

Delis DC, Kaplan E, Kramer JH. 2001. *The Delis-Kaplan Executive Function System*. San Antonio, Texas: Psychological.

Folstein MF, Folstein SE, McHugh PR. 1975. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 12:189–198. doi:10.1016/0022-3956(75)90026-6

Gamer M, Lemon J, Fellows I, et al. 2019. Package "irr". Version 0.84.1. Available at: https://cran.r-project.org/web/packages/irr/irr.pdf. Accessed October 2020.

Glenn S, Mefford J, Pieters A, The Cognitive Healthcare Company. 2017. Motion restriction and measurement for self-administered cognitive tests. US Patent 9,703,407.

Glenn S, Mefford J, The Cognitive Healthcare Company, assignee. 2019a. Biomechanical motion measurement and analysis for self-administered tests. US Patent 10,444,980.

Glenn S, Mefford J, The Cognitive Healthcare Company. 2019b. Data collection and analysis for self-administered cognitive tests characterizing fine motor functions. US Patent 10,383,553.

Glenn S, Mefford J, The Cognitive Healthcare Company. 2020. Automated delivery of unique, equivalent task versions for computer delivered testing environments. US Patent 10,748,439.

Hastie T, Mazumder R. 2021. Package "softImpute". Available at: https://cran.r-project.org/web/packages/softImpute/index.html. Accessed October 2020.

Iverson GL. 2001. Interpreting change on the WAIS–III/WMS–III in clinical samples. *Arch Clin Neuropsychol*. 16:183–191.

Kaplan EF, Goodglass H, Weintraub S. 1983. *The Boston Naming Test*, 2nd ed. Philadelphia, Pennsylvania: Lea & Febiger.

Knopman DS, Roberts RO, Geda YE, et al. 2010. Validation of the Telephone Interview for Cognitive Status—Modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology*. 34:34–42. doi:10.1159/000255464

Lee TM, Chan CC. 2020. The significance of early identification and timely intervention for people at risk of developing Alzheimer's disease. *J Alzheimers Neurodegener Dis*. 6:034. doi:10.24966/AND-9608/100034

Lezak MD. 1987. Relationships between personality disorders, social disturbances, and physical disability following traumatic brain injury. *J Head Trauma Rehabil*. 2:57–69. doi:10.1097/00001199-198703000-00009

Malec JF, Lezak MD. 2008. Manual for the Mayo-Portland Adaptability Inventory (MPAI-4) for Adults, Children and Adolescents. Available at: http://www.tbims.org/mpai/manual.pdf. Accessed October 10, 2019.

Nasreddine ZS, Phillips NA, Bédirian V, et al. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 53:695–699. doi:10.1111/j.1532-5415.2005.53221.x

O'Bryant SE, Humphreys JD, Smith GE, et al. 2008. Detecting dementia with the Mini-Mental State Examination in highly educated individuals. *Arch Neurol*. 65:963–967. doi:10.1001/archneur.65.7.963

Oyama A, Takeda S, Ito Y, et al. 2019. Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. *Sci Rep*. 9:12932. doi:10.1038/s41598-019-49275-x

Partington JE, Leiter RG. 1949. Partington's Pathways Test. *Psychol Serv Center J*. 1:11–20.

Petersen RC, Lopez O, Armstrong MJ, et al. 2018. Practice guideline update summary: mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology*. 90:126–135. doi:10.1212/WNL.0000000000004826

Roberts R, Knopman DS. 2013. Classification and epidemiology of MCI. *Clin Geriatr Med*. 29:753–772. doi:10.1016/j.cger.2013.07.003

Sabbagh MN, Boada M, Borson S, et al. 2020. Early detection of mild cognitive impairment (MCI) in primary care. *J Prev Alzheimers Dis*. 7:165–170. doi:10.14283/jpad.2020.21

Seo EH, Lee DY, Kim SG, et al. 2011. Validity of the Telephone Interview for Cognitive Status (TICS) and modified TICS (TICSm) for mild cognitive impairment (MCI) and dementia screening. *Arch Gerontol Geriatr*. 52:e26–e30. doi:10.1016/j.archger.2010.04.008

Signorell A, Aho K, Alfons A, et al. 2020. DescTools: Tools for Descriptive Statistics. R package version 0.99.34. Available at: https://cran.r-project.org/web/packages/DescTools/index.html. Accessed November 10, 2020.

Snow WG, Tierney MC, Zorzitto ML, et al. 1989. WAIS–R test–retest reliability in a normal elderly sample. *J Clin Exp Neuropsychol*. 11:423–428. doi:10.1080/01688638908400903

Wechsler D. 1997. *WMS–III: Wechsler Memory Scale Administration and Scoring Manual*. San Antonio, Texas: Psychological.

Wechsler D. 2008. *WAIS–IV Technical and Interpretive Manual*. San Antonio, Texas: Psychological.

Wittenberg R, Knapp M, Karagiannidou M, et al. 2019. Economic impacts of introducing diagnostics for mild cognitive impairment Alzheimer's disease patients. *Alzheimers Dement (N Y)*. 5:382–387. doi:10.1016/j.trci.2019.06.001

Wurm MJ, Rathouz PJ, Hanlon BM. 2021. Regularized ordinal regression and the ordinalNet R package. *J Stat Softw*. 99:1–42. doi:10.18637/jss.v099.i06

Yesavage JA, Sheikh JI. 1986. Geriatric Depression Scale (GDS) recent evidence and development of a shorter version. *Clin Gerontol*. 5:165–173. doi:org/10.1300/J018v05n01_09

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *JR Stat Soc Series B Stat Methodol*. 67:301–320. doi:10.1111/j.1467-9868.2005.00503.x