


RESEARCH ARTICLE

Reproducibility and replicability of high-frequency, in-home digital biomarkers in reducing sample sizes for clinical trials

Chao-Yi Wu^{1,2}  | Zachary Beattie^{1,2} | Nora Mattek^{1,2} | Nicole Sharma^{1,2} | Jeffrey Kaye^{1,2} | Hiroko H. Dodge^{1,2}

¹ Department of Neurology, Oregon Health & Science University (OHSU), Portland, Oregon, USA

² Oregon Center for Aging & Technology (ORCATECH), OHSU, Portland, Oregon, USA

Correspondence

Chao-Yi Wu, Department of Neurology, Oregon Health & Science University, School of Medicine, 3303 S Bond Avenue CH13B Portland, OR 97239, USA.

E-mail: wucha@ohsu.edu

Funding information

the National Institute on Aging, Grant/Award Numbers: P30AG024978, R01AG024059, P30-AG008017, P30AG066518, U2C AG054397; Department of Veterans Affairs Health Services Research and Development, Grant/Award Number: IIR 17-144; National Center for Advancing Translational Sciences, Grant/Award Number: UL1 TR002369

Abstract

Introduction: Reproducibility and replicability of results are rarely achieved for digital biomarkers analyses. We reproduced and replicated previously reported sample size estimates based on digital biomarker and neuropsychological test outcomes in a hypothetical 4-year early-phase Alzheimer's disease trial.

Methods: Original data and newly collected data (using a different motion sensor) came from the Oregon Center for Aging & Technology (ORCATECH). Given trajectories of those with incident mild cognitive impairment and normal cognition would represent trajectories of the control and experimental groups in a hypothetical trial, sample sizes to provide 80% power to detect effect sizes ranging from 20% to 50% were calculated.

Results: For the reproducibility, identical *P*-values and slope estimates were found with both digital biomarkers and neuropsychological test measures between the previous and current studies. As for the replicability, a greater correlation was found between original and replicated sample size estimates for digital biomarkers ($r = 0.87$, $P < .001$) than neuropsychological test outcomes ($r = 0.75$, $P < .001$).

Discussion: Reproducibility and replicability of digital biomarker analyses are feasible and encouraged to establish the reliability of findings.

KEYWORDS

early prevention, linear mixed-effect models, mild cognitive impairment, randomized controlled trials, technology assessment

1 | INTRODUCTION

Although billions of dollars are spent on preclinical or early-phase Alzheimer's disease (AD) drug development and hundreds of clinical trials, most drugs are still in the pipeline.¹ In this climate, the prospect of advancing new trials at tremendous additional cost is daunting. Much of this expense in AD trials lies in their inefficiency. Lengthy follow-up and large sample sizes are required to detect treatment effects with sufficient statistical power. A review concludes that existing 2-year AD trials often recruit 1200 to 2300 participants to reli-

ably determine the efficacy of a trial.² Approximately 30,000 participants were required for new drugs in the pipeline in 2020.¹ With the increases in AD cases and costs burdening the health-care system and caregivers, new methodologies are needed to facilitate the progress of AD research.

Digital biomarkers may offer a solution to improve the efficiency of AD trials. Various digital biomarkers have been shown to correlate with neuropsychological outcomes, imaging markers of neurodegeneration, and *post mortem* neuropathology.³ For example, in-home gait speed, sleep, and computer use collected from infrared passive

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* published by Wiley Periodicals, LLC on behalf of Alzheimer's Association

sensors (for gait or sleep analysis) or computer monitoring software (for computer use) were associated with global cognition, memory, attention, and processing.^{4–6} Computer usage collected from software was associated with hippocampal and medial temporal lobe volumes in cognitively intact older adults.⁷ A composite, as well as individual measures of digital biomarkers (mobility, cognition, socialization, and sleep) was correlated with greater neuritic plaque severity and Braak score.³ We previously demonstrated that these biomarkers can differentiate older community-dwelling adults with normal cognitive function from those with progression to mild cognitive impairment (MCI).^{8,9} Importantly, using these digital biomarkers and the subject-specific thresholds derived from data collected during a short duration of time at baseline, markedly lower sample sizes (compared to conventional cognitive tests) were projected to be needed in hypothetical preclinical AD trials.¹⁰ Fewer than 100 participants are needed to detect a 30% effect size with 80% power over 4 years using a single digital biomarker, compared to 1912 participants needed with a cognitive test. In other words, the digital markers derived from the in-home sensor platform yielded a > 10-fold reduction in sample size needed to obtain required power. These promising results may change the design of future interventions and drug trials. However, the reproducibility of the findings needs to be carefully evaluated before adopting this new paradigm.

Reproducibility and replicability in science establishes the reliability and validity of findings and experiments. According to the National Academies of Sciences, Engineering, and Medicine, the definition of reproducibility refers to “instances in which the original researcher’s data and computer codes are used to regenerate the results,” while replicability refers to “instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.”¹¹ A replicability crisis has been noted, suggesting that researchers are frequently not able to reproduce or replicate the findings generated from other researchers.¹² Even when results are replicated, most replication effects are smaller than the original study. Many studies point out that replication is susceptible to biases caused by the number of subjects, the selection of subjects, and publication bias.^{13,14} Thus, reproducing and replicating findings from previous studies has become important to increase the rigor of science and reduce study biases with a priori analytical plans.

The purpose of this work was to reproduce and replicate the study conducted by Dodge et al.,¹⁰ by estimating the sample sizes needed for a hypothetical 4-year preclinical AD trial with three measurement outcomes: (1) trajectories of in-home digital biomarkers using the data as observed (continuous outcomes), (2) trajectories of in-home digital biomarkers using the likelihood of experiencing deviations from subject-specific-thresholds defined at baseline as outcomes (subject-specific threshold model), and (3) conventional neuropsychological tests obtained using standard methods (i.e., annual assessment). The current analyses used the original data (collected between 2007 and 2012) from the above study to reproduce the results and also aimed to replicate the original results by using new digital biomarker data (collected between 2015 and 2018) collected with a new motion sensor. To evaluate the success of replication, we adopted the method of the

RESEARCH IN CONTEXT

- 1. Systematic review:** The authors reviewed the efficacy of existing Alzheimer’s disease (AD) trials (number of subjects and duration of trials) and the reproducibility and replicability of these studies. While many studies have reported the sample size needed for various assessments adopted in the AD trials (digital biomarkers; neuropsychological tests), few publications investigated the reproducibility and replicability of findings.
- 2. Interpretation:** Our study demonstrates that sample sizes needed with outcomes generated with high-resolution digital biomarkers (walking speed and computer use) are reproducible and replicable in early-phase, hypothetical AD trials, while the replicability of conventional neuropsychological tests is less robust.
- 3. Future directions:** The article proposes a promising paradigm of using digital biomarkers for AD research because of its reliability in detecting early behavioral changes before mild cognitive impairment. The replicability of digital biomarker analysis is encouraged to increase the rigor of future AD science.

Open Science Collaboration to compare the concordance between the original and replicated findings.¹⁵

2 | METHODS

2.1 | Participants

Data came from a longitudinal aging study, the Oregon Center for Aging & Technology (ORCATECH) Life Lab (OLL). Details of the protocol and in-home sensor platform have been published elsewhere.¹⁶ The study protocol was approved by the Oregon Health & Science University Institutional Review Board (IRB #2765). All participants provided written informed consent. In both the original and new studies, the inclusion criteria were participants aged 65 and above, living independently, and not demented. The original data were collected between 2007 and 2012. The new data used to examine replicability were collected between 2015 and 2018. Forty-one older adults (36%) included in the new study were also in the original study (Figure 1). The diagnosis of MCI was defined as a Clinical Dementia Rating (CDR) score of 0.5 collected during the annual in-home clinical evaluation.

2.2 | Home-based digital biomarkers

The passive infrared (PIR) motion sensors were used to estimate participant walking speed. For both studies, daily walking speed was

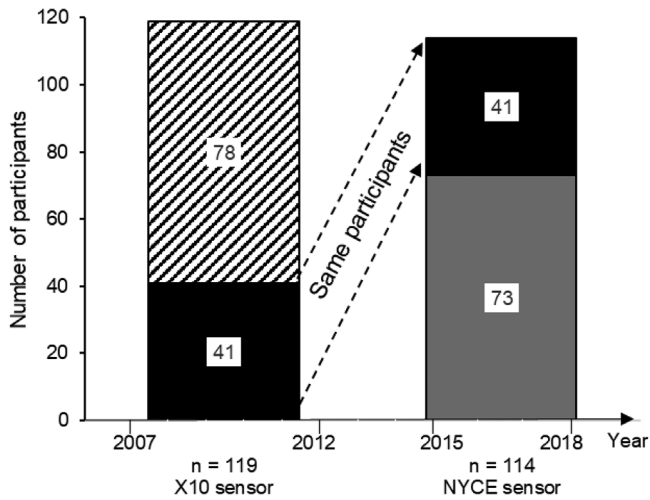


FIGURE 1 Comparison of the original and new studies

estimated using the time it would take a participant to walk beneath four motion sensors aligned and evenly spaced on the ceiling. The validity of using PIR motion sensors in measuring gait speed was tested in a previous study.^{17,18} While in the original study X10 (<https://www.x10.com/ms16a.html>) motion sensors were used, NYCE (www.nycesensors.com/product/motion-sensor) motion sensors were used in the new study. The X10 motion sensors report the presence on a continuous basis, and the NYCE motions sensors report the beginning time and ending time of detected presence. A more in-depth description of the algorithm used to extract walking speed from the PIR motion sensors has been discussed previously.¹⁸ Daily home computer usage was collected using commercial software (WorkTime) installed on participants' desktops.⁸ All the sensor data were collected using a wireless hub computer (original study: Globalscale DreamPlug, new study: Raspberry Pi 3 Model B) placed in each home. Sensor data were automatically uploaded from each hub computer to secure ORCATECH servers. Data collected after MCI incidence were excluded. All data were unobtrusively and continually collected with an average duration of 900 days (range: 97–1123 days) per participant in the current study.

2.3 | Annual neuropsychological tests

Seven neuropsychological tests were collected from annual home visits, including: Category Fluency (animal + vegetable) (language-based executive function),¹⁹ Trail Making Test A,²⁰ Trail Making Test B (executive function),²⁰ Wechsler Adult Intelligence Scale–Digit Symbol (attention, processing speed, working memory, visuospatial processing),²¹ Logical Memory Immediate Recall (learning),²² Logical Memory Delayed Recall (memory),²² and Boston Naming Test (language).²³

2.4 | Statistical analysis

First, longitudinal linear mixed effects models were fit to examine the slope difference between cognitively intact and incident MCI groups

on continuous variables (walking speed, walking speed variability, computer usage, all neuropsychological tests) without using subject-specific thresholds. Daily digital biomarker data (walking speed, computer usage) were processed as weekly mean and variability data, while neuropsychological test data were processed as yearly data. The models included random intercepts with group (cognitively normal vs. MCI converters) and time (days from baseline) being fixed-effect variables, and an unstructured covariance structure. Including non-linear terms and/or random slopes did not improve the model fit based on the Bayesian information criteria (BIC); therefore, we only included random intercepts without non-linear terms in the models. Using the missing at random (MAR) assumption, where missingness depends on observed variables (a reasonable assumption for our data), the modelling approach we used (linear mixed effects models) provided valid estimates.

The second modelling approach was to generate individual-specific thresholds using subject-specific baseline distribution on walking speed, walking speed variability, and computer usage. We calculated each participant's distributions of weekly mean walking speed, weekly walking speed variability, and weekly computer usage using the data observed during the first 90 days. This step generated individual-specific distributions of each activity and several measures of their variability (such as the subject-specific lowest 10th, 20th, 30th percentile, etc.) at baseline. We then defined weekly walking speed, walking speed variability, and computer usage data as either below or above pre-defined thresholds (YES/NO; 0/1) for each week using the baseline thresholds defined above. We used generalized linear mixed effects models with outcomes being the likelihood of experiencing values below the "subject-specific" lowest 10th, 20th, 30th, 40th, and 50th percentile thresholds (for walking speed and computer usage) and the likelihood of experiencing values above the "subject-specific" highest 60th, 70th, 80th, and 90th percentile thresholds (for walking speed variability) using a logit link. Generalized linear mixed effects models included group (cognitive normal vs. MCI converters) and time (days from baseline) variables as fixed-effect variables and a random intercept because as with the mixed effects models, model fitness (BIC) did not improve by adding random slopes. This approach estimated the likelihood of having individual-specific low-performance thresholds over time. We expected that those who develop clinical MCI later were more likely to experience worse outcomes over time defined by subject-specific threshold (e.g., low threshold for computer usage, high threshold for walking speed variability). Analytical approaches used were described in detail in the original study.¹⁰

2.4.1 | Sample size calculation

Percentages of effect sizes (20%, 30%, 40%, and 50%) were used to estimate sample sizes needed for a hypothetical 4-year trial to achieve 80% power. Here, a 30% effect size indicated the hypothetical pre-clinical AD trial could reduce 70% of the discrepancy on the outcome between treatment and placebo groups. The percentage of effect sizes followed previous studies estimating the sample size needed for

TABLE 1 Baseline characteristics of the new data collected between 2015 and 2018 (the same data for the original study collected between 2007 and 2012 are included in the supporting information)

Characteristics [n (%)]	All		Cognitively normal		MCI converter		t-statistics/ χ^2 - statistics	P
	114	(100)	96	(84.2)	18	(15.8)		
Age [mean (SD)]	84.54	(7.95)	83.84	(7.99)	88.13	(6.92)	$t_{(109)} = -2.13$.04
Female [n (%)]	86	(75.4)	71	(74.0)	15	(83.3)	$\chi^2_{(1)} = 0.42$.52
Years of education [mean (SD)]	15.66	(2.56)	15.73	(2.62)	15.33	(2.28)	$t_{(109)} = 0.60$.55
Duration of follow-up in days	898.52	(246.85)	927.10	(224.60)	734.60	(308.70)	$t_{(99)} = 2.89$.005
Duration of follow-up in days before MCI incidence					641.40	(476.85)		
In-home continuously monitored data								
Mean walking speed (cm/s)	68.33	(21.47)	69.32	(21.21)	62.51	(22.81)	$t_{(101)} = 1.14$.26
Mean daily computer usage (min)	109.89	(124.58)	110.00	(116.50)	109.60	(161.70)	$t_{(95)} = 0.01$.99
Neuropsychological tests								
Category Fluency (animals + vegetables)	33.32	(10.00)	35.08	(9.48)	24.22	(7.51)	$t_{(109)} = 4.58$	<.001
Trail Making Test A	42.11	(16.13)	39.98	(14.69)	52.67	(19.08)	$t_{(105)} = -3.17$.002
Trail Making Test B	102.61	(48.05)	95.84	(41.91)	137.70	(62.59)	$t_{(97)} = -3.35$.001
Digit Symbol	41.81	(10.63)	42.84	(10.64)	35.38	(8.36)	$t_{(92)} = 2.41$.02
Logical Memory Immediate Recall	14.43	(4.19)	14.92	(3.81)	11.89	(5.18)	$t_{(109)} = 2.91$.004
Logical Memory Delayed Recall	13.70	(4.42)	14.40	(4.10)	10.11	(4.39)	$t_{(109)} = 4.01$	<.001
Boston Naming (30 items)	27.06	(2.85)	27.39	(2.66)	25.39	(3.24)	$t_{(108)} = 2.82$.006

Abbreviations: MCI, mild cognitive impairment; SD, standard deviation.

preclinical AD trials.²⁴ Using five parameters derived from the models mentioned above (intercept, group intercept, slope estimate, group effect on slope and variance, and covariance matrices of these variables) and a parameter of time (1456 days = 7 days x 52 weeks x 4 years), Monte Carlo simulation of linear mixed effect models and generalized linear mixed effects models were conducted. A rejection rate (i.e., null hypothesis was rejected at alpha = 0.05) at 80% over 1000 iterations of modelling simulation was used to determine the sample size estimates.

2.4.2 | Reproducibility indices

Reproducibility is “the ability to generate the same experiments or findings based on the same approach and data.”¹¹ The same dataset, analytical approaches, and code from the original study were used to examine whether identical findings could be found. *P*-values, slope estimates, and sample size estimates were compared between the original and reproduced results. Analyses were conducted by an independent (from the original study) analyst (C.-Y.W) using the same software package (SAS).

2.4.3 | Replicability indices

Replicability is “the ability to replicate study results using newly collected data.”¹¹ Newly collected data were used to examine whether the findings could be replicated. Four indices were used to examine the suc-

cess of replication between the original and new results: (1) the concordance of a significant alpha, *P*-value ($P < .05$) using Cohen's kappa, (2) the concordance of slope estimates using Spearman's correlation, (3) the percentage of estimates where original slope estimates lay within the 95% confidence interval (CI) of new estimates, and (4) the concordance of sample size estimations using Spearman's correlation.¹⁵ Because 41 participants were in both the original and new studies (Figure 1), we conducted sensitivity analyses by including a covariate indicating the overlapped participants and the interaction with time in the new study and examined whether the results changed. The main conclusions did not differ; therefore, we report the results without these covariates in the subsequent section.

3 | RESULTS

Table 1 shows the baseline characteristics of participants from the new data. The same data for the original study derived from the previous publication is included in supporting information. Among 114 older adults cognitively healthy at entry, 18 (15.7%) developed MCI during the follow-up period. The mean age of participants in the original and new studies were 84.42 (standard deviation [SD] = 5.07) and 84.54 (SD = 7.95), respectively. In the original study 15.1% of the participants were male, and in the new study 24.6% were male. The duration of follow-up was 3.8 years in the original study and 2.5 years in the new study. A total of 11,822 and 7,348 weeks of walking speed and computer usage data were used in the current analysis. The proportion of

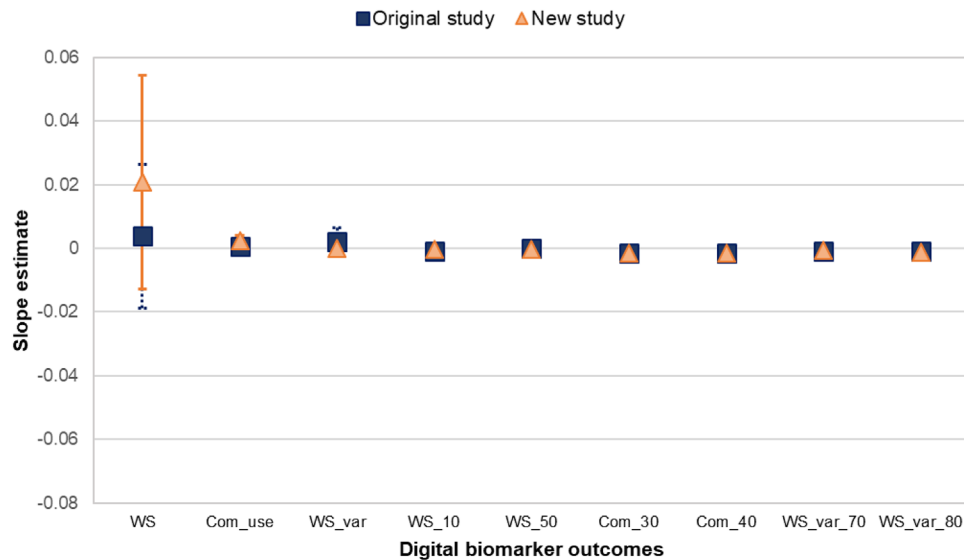


FIGURE 2 95% confidence interval (CI) of original and new slope estimates of digital biomarkers. WS, walking speed; Com_use, computer usage; WS_var, walking speed variability; WS_10, walking speed: likelihood of 10th percentile low; WS_50, walking speed: likelihood of 50th percentile low; Com_30, computer usage: likelihood of 30th percentile low; Com_40, computer usage: likelihood of 40th percentile low; WS_var_70, walking speed variability: likelihood of 70th percentile high; WS_var_80, walking speed variability: likelihood of 80th percentile high.

missing data for the weekly walking speed and computer usage data was 6.1% and 22.4%, respectively.

3.1 | Reproducibility results

The original study was reproducible. Identical *P*-values, slope estimates, and sample size estimates were found with both digital biomarkers and neuropsychological test measures.

3.2 | Replicability results

3.2.1 | Digital biomarkers

Table 2 shows the replicated results of digital biomarkers. A substantial agreement in the significance of *P*-values was found between original and new results in which eight out of nine *P*-values agreed in significance (Cohen's kappa = 0.78, *P* = .32; walking speed variability likelihood of having 70th percentile: *P*-value of 0.07 in new data, and *P*-value of 0.0009 in old data). A strong correlation was found between original and new slope estimates (interaction parameters; Spearman's *r* = 0.97, *P* < .001). All the original slope estimates were within the 95% CI of new slope estimates (Figure 2).

3.2.2 | Neuropsychological tests

Table 3 shows the replicated results of neuropsychological tests. A high agreement in the significance of *P*-values was found between the orig-

inal and new results. All seven *P*-values agreed in significance (Cohen's kappa = 1). A moderate correlation was found between the original and new slope estimates (interaction parameters; Spearman's *r* = 0.68, *P* = .09). All the original slope estimates were within the 95% CI of the new slope estimates (Figure 3).

3.3 | Sample size estimates

3.3.1 | Digital biomarkers

A strong correlation was found between the original and new sample sizes (Spearman's *r* = 0.87, *P* < .001). As expected, when the group differences were statistically significant (i.e., higher signal-to-noise ratio indicated by *P*-value listed the last column of Table 2), the estimated sample sizes were similar regardless of using the original or the new data sets. On the other hand, estimated sample sizes can differ largely for the outcomes for which we did not find significant group differences. For example, the likelihood of having low computer usage time, defined using subject-specific computer usage time at baseline, exhibited significant group differences in trajectories; hence, the sample sizes needed to achieve a 30% effect size with 80% statistical power in original and new studies were 26 and 38 subjects, respectively. Similarly, for the likelihood of having high walking speed variability, 86 and 52 subjects were needed in the original and new studies, respectively. On the other hand, when walking speed—which did not exhibit statistically significant group differences—was analyzed in linear mixed effects models (i.e., non-subject-specific models), then the estimated sample size differed largely: 41,156 versus 2,739.

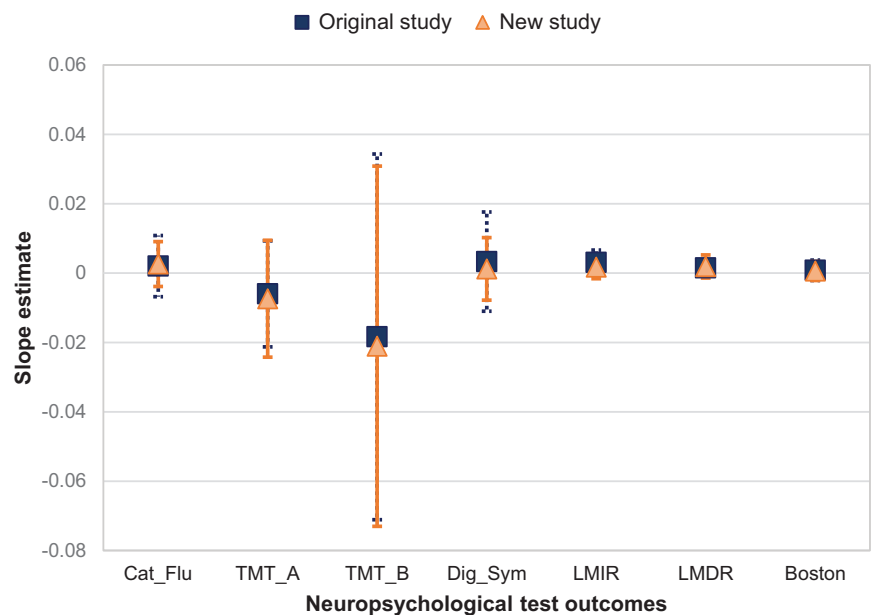
TABLE 2 Replicability results of digital biomarker outcomes

Model	Outcome	Results from new data						Results from original data						P-values of the interaction term*	
		Clinical trial sample size estimate (estimation based on 4 years of follow-up)						Clinical trial sample size estimate (estimation based on 4 years of follow-up)						New data	Original data
		Treatment effect size 20%	Treatment effect size 30%	Treatment effect size 40%	Treatment effect size 50%	Treatment effect size 20%	Treatment effect size 30%	Treatment effect size 40%	Treatment effect size 50%	Treatment effect size 20%	Treatment effect size 30%	Treatment effect size 40%	Treatment effect size 50%		
Linear mixed effects models	Walking speed	6162	2739	1541	986	92600	41156	23150	14816	.23	.74				
	Walking speed variability	2377	1057	595	381	7550	3358	1888	1208	.05	.34				
	Computer usage	1529	680	383	245	1100	490	276	176	.01	.01				
Generalized linear mixed effects models (with random intercept)	Walking speed: likelihood of 10 th percentile low	5206	2328	1304	802	588	262	148	94	.68	.10				
	Walking speed: likelihood of 50 th percentile low	1560	728	410	254	14550	6468	3638	2328	.38	.65				
	Computer usage: likelihood of 30 th percentile low	126	58	34	22	76	34	20	14	.04	<.0001				
	Computer usage: likelihood of 40 th percentile low	84	38	22	14	58	26	16	10	.02	<.0001				
	Walking speed variability: likelihood of 70 th percentile high	284	130	72	48	184	82	46	30	.07	.0009				
	Walking speed variability: likelihood of 80 th percentile high	116	52	28	20	192	86	48	32	.008	.0001				

* Significant interaction term indicates there is a group difference in longitudinal trajectories (those with normal cognition vs. those who developed mild cognitive impairment later on).

TABLE 3 Replicability results of neuropsychological test outcomes

Outcome	Results from new data				Results from original data				P-values of the interaction term	
	Clinical trial sample size estimate (estimation based on 4 years of follow-up)				Clinical trial sample size estimate (estimation based on 4 years of follow-up)				New data	Original data
	Treatment effect size 20%	Treatment effect size 30%	Treatment effect size 40%	Treatment effect size 50%	Treatment effect size 20%	Treatment effect size 30%	Treatment effect size 40%	Treatment effect size 50%		
Category Fluency (animal + vegetable)	25696	11421	6424	4112	8050	3578	2013	1288	.66	.43
Trail Making Test A	8337	3705	2085	1334	6800	3022	1700	1088	.45	.39
Trail Making Test B	9675	4300	2419	1548	7500	3334	1876	1200	.49	.43
Digit Symbol	13677	6079	3420	2189	75900	33734	18976	12144	.65	.80
Logical Memory Immediate Recall	1593	708	399	255	4900	2178	1226	784	.09	.32
Logical Memory Delayed Recall	2058	915	515	330	4300	1912	1076	688	.14	.28
Boston Naming (30 items)	19234	8549	4809	3078	26800	11912	6700	4288	.59	.66

FIGURE 3 95% confidence interval (CI) of original and new slope estimates of neuropsychological tests. Cat_Flu, Category Fluency; TMT_A, Trail Making Test A; TMT_B, Trail Making Test B; Dig_Sym, Digit Symbol; LMIR, Logical Memory Immediate Recall; LMDR, Logical Memory Delayed Recall; Boston, Boston Naming Test.

3.3.2 | Neuropsychological tests

A moderate correlation was found between original and new sample sizes (Spearman's $r = 0.75$, $P < .001$).

4 | DISCUSSION

To use digital biomarkers in clinical trials as key outcomes and for monitoring disease progression, replicating results using different sensors and different participants is important. The results of this study suggest that the sample sizes needed with outcomes generated using high-resolution digital biomarkers with subject-specific threshold models are reproducible and replicable. Clinically relevant activity collected

passively by an in-home sensor platform which can generate subject-specific thresholds within a short duration of time (e.g., 90 days) are viable complements to future AD and related trials to reduce cost and increase the efficiency of trials.

The extent of replication was different in two outcome measures. One digital biomarker outcome (walking speed variability: likelihood of 70th percentile high; $P < .001$) was significant in the original study while the replication finding ($P = .07$) was not significant at the pre-specified alpha level. This was the only replication criteria for which a digital biomarker performed worse than the cognitive tests. The National Academies of Sciences, Engineering and Medicine and others have suggested that the significance of P -value should not be overly interpreted when the P -value is close to .05 in the replicability studies.^{11,25} Instead, a combination of multiple replicability indices can better inform the

reliability of the study. Looking at the multiple replicability criteria, there was more variation in the replicability results of neuropsychological tests (moderate correlations in slope and sample size estimates) than the digital biomarkers (strong correlations in slope and sample size estimates). This variability may come from multiple sources, such as practice or learning effects because participants have been administered the cognitive tests over several years. Other factors, including the timing of the assessment (morning vs. afternoon) or the testing environment may affect consistency. This suggests a further value of the continuous everyday digital measures for longitudinal observational studies and clinical trials, in which the unobtrusive passive digital metrics provide objective responses that are not susceptible to classic learning effects.

Although from correlational statistics we showed replicability, we noticed relatively large discrepancies in estimated sample sizes between original and new study results in some neuropsychological tests and digital biomarkers. While Trail Making Tests A and B (TMT-A, TMT-B) demonstrated excellent replicability in sample sizes, the results of Digit Symbol and Category Fluency varied between original and new studies. Plausibly, across older adults with subtle cognitive decline, the impairment in attention or language executive (Digit Symbol; Category Fluency) are more variable while the impairment in the complex executive function (TMT-A, TMT-B) are more prevalent and progressive. As for the digital biomarkers, we found significant group differences in trajectories for computer usage and walking speed variabilities when subject-specific threshold models were applied, regardless of original or new data sets. For these outcomes, we also found high replication in estimated sample sizes. We recommend future real-world preclinical AD trials to adopt subject-specific threshold models¹⁰ as proof-of-concept studies.

The digital measures can also be considered to complement episodically captured cognitive tests by providing a measure of functional change that can augment interpretation of the conventional assessment. Thus, digital biomarkers offer advantages beyond reducing sample sizes. For early phase trials, the instant and delayed effects of drugs can be measured through high-resolution digital biomarkers, accompanied with medication adherence measures using an electronic pillbox.^{26,27} Assessment of function at home may be useful in interpreting cognitive tests on a particular testing day such as the effect of poor sleep preceding testing.⁵ Further, these digital biomarkers represent changes in health and function, which can be a surrogate measure of biological disease progression⁷ and uncover the unclear mechanisms of drug ingredients in delaying cognitive decline. Because these digital biomarkers reflect ecologically valid real-time conditions, this paradigm may also indicate or minimize the potential harm that trials bring to participants (e.g., overdose, side effects).

There are some limitations in these studies. We had a small sample size for both original and new data sets; therefore, it was not feasible to exclude 41 overlapping subjects for the statistical analyses. Notably, the 41 overlapping subjects had different sensors when contributing to the replication cohort and thus were contributing different data. Our sensitivity analysis also showed that using a model with an indicator of these overlapping subjects did not alter the results. Compared to the

original study, this new study has a shorter average duration of follow-up. This may explain the attenuated *P*-values in the new study compared to the original study. Oftentimes, a shorter follow-up duration may require a larger sample size to achieve sufficient statistical power. Yet, the replication results of digital biomarkers are similar to the original study in subject-specific threshold models, especially in computer usage. This suggests that digital biomarkers may differentiate cognitively intact versus MCI-emergent cases even earlier (2–3 years) than the original study (4 years) with a similar sample size. Another limitation is the method to collect computer usage time. Computer usage is estimated by the total time spent on desktop or laptop computers. As personal devices diversify and increase in popularity among the aging population, future studies will need to collect total time in using a potential array of computing devices. Because these measures are individualized, the difference in devices used across participants may not make a difference in intra-individual use metrics over time as long as the same device use profile is retained within a participant during a study. This issue will require empirical analysis with new data. Finally, determining consistency between two inferences can be approached in many ways, such as heterogeneity tests between original and new studies.

In summary, this study demonstrates that the sample sizes needed using in-home digital biomarkers with subject-specific thresholds as outcome measures are replicable. Using several in-home digital biomarkers to track neurological changes over time may not only improve the efficiency of trials, but at the same time offer additional value such as indicating adverse effects, providing ecologically valid functional performance data, and further facilitating the study of biological mechanisms.

ACKNOWLEDGMENTS

We would like to thank participants in the ORCATECH Life Lab (OLL) study. This work was supported by several grants, including the National Institute on Aging: P30AG024978, R01AG024059, P30-AG008017, P30AG066518 and U2C AG054397; Department of Veterans Affairs Health Services Research and Development: IIR 17-144; National Center for Advancing Translational Sciences: UL1 TR002369.

CONFLICTS OF INTEREST

J. Kaye in the past 36 months has been directly compensated for serving on a Data Safety Monitoring Committee for Eli Lilly, the Scientific Advisory Board of Sage Bionetworks, the Roche/Genentech Scientific Advisory Committee for Digital Health Solutions in Alzheimer's Disease, and as an external Advisory Committee member for two Alzheimer's Disease Research Centers. He has received research support awarded to his institution (Oregon Health & Science University) from the NIH, NSF, the Department of Veterans Affairs, USC Alzheimer's Therapeutic Research Institute, Merck, AbbVie, Eisai, Green Valley Pharmaceuticals, and Alektor. He holds stock in Life Analytics Inc., for which no payments have been made to him or his institution. He has received reimbursement through Medicare or commercial insurance plans for providing clinical assessment and care for patients. He has served on the editorial advisory board and as Associate Editor

of the journal *Alzheimer's & Dementia* and as Associate Editor for *Journal of Translational Engineering in Health and Medicine*. H. Dodge receives honorarium from the Alzheimer's Clinical Trials Consortium (ACTC) and is a consultant for Biogen, Inc. She serves as a Data Safety and Monitoring Board member for five clinical trials and an external Advisory Committee member for two Alzheimer's Disease Centers. She receives research funding from National Institute on Aging and serves as a senior editor for *Alzheimer's & Dementia: Translational Research and Clinical Interventions* and the Statistical Editor for the journal *Alzheimer's & Dementia*. Z. Beattie holds stock in Life Analytics Inc. for which no payments have been made to him or his institution. CY Wu, N Sharma, N Mattek have no conflicts of interest to declare.

ORCID

Chao-Yi Wu  <https://orcid.org/0000-0002-2187-6509>

References

- Sabbagh MN. Alzheimer's Disease Drug Development Pipeline 2020. *J Prev Alzheimer's Dis*. 2020;7(2):66-67.
- Schneider LS, Mangialasche F, Andreassen N, et al. Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014. *J Intern Med*. 2014;275(3):251-283.
- Kaye JA. Continuously acquired, home-based digital biomarkers of activity and function are related to Alzheimer's disease neuropathology. *J Prev Alzheimer's Dis*. 2019;6:S15.
- Kaye J, Mattek N, Dodge H, et al. One walk a year to 1000 within a year: continuous in-home unobtrusive gait assessment of older adults. *Gait Posture*. 2012;35(2):197-202.
- Seelye A, Hagler S, Mattek N, et al. Computer mouse movement patterns: a potential marker of mild cognitive impairment. *Alzheimer's Dement Diagnosis Assess Dis Monit*. 2015;1(4):472-480.
- Seelye A, Mattek N, Howieson D, Riley T, Wild K, Kaye J. The impact of sleep on neuropsychological performance in cognitively intact older adults using a novel in-home sensor-based sleep assessment approach. *Clin Neuropsychol*. 2015;29(1):53-66.
- Silbert LC, Dodge HH, Lahna D, et al. Less daily computer use is related to smaller hippocampal volumes in cognitively intact elderly. *J Alzheimer's Dis*. 2016;52(2):713-717.
- Kaye J, Mattek N, Dodge HH, et al. Unobtrusive measurement of daily computer use to detect mild cognitive impairment. *Alzheimer's Dement*. 2014;10(1):10-17.
- Dodge HH, Mattek NC, Austin D, Hayes TL, Kaye JA. In-home walking speeds and variability trajectories associated with mild cognitive impairment. *Neurology*. 2012;78(24):1946-1952. <https://doi.org/10.1212/WNL.0b013e318259e1de>.
- Dodge HH, Zhu J, Mattek NC, Austin D, Kornfeld J, Kaye JA. Use of high-frequency in-home monitoring data may reduce sample sizes needed in clinical trials. *PLoS One*. 2015;10(9):e0138095. <https://doi.org/10.1371/journal.pone.0138095>.
- National Academies of Sciences Engineering Medicine. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press; 2019. doi:10.17226/25303
- Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533(7604):452-454. <https://doi.org/10.1038/533452a>.
- Hua X, Hibar DP, Ching CRK, et al. Unbiased tensor-based morphometry: improved robustness and sample size estimates for Alzheimer's disease clinical trials. *Neuroimage*. 2013;66:648-661. <https://doi.org/10.1016/j.neuroimage.2012.10.086>.
- Etz A, Vandekerckhove J. A Bayesian perspective on the reproducibility project: psychology. *PLoS One*. 2016;11(2):e0149794. <https://doi.org/10.1371/journal.pone.0149794>.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science (80-)*. 2015;349(6251):aac4716. <https://doi.org/10.1126/science.aac4716>
- Kaye J, Reynolds C, Bowman M, et al. Methodology for establishing a community-wide life laboratory for capturing unobtrusive and continuous remote activity and health data. *J Vis Exp*. 2018;2018(137). <https://doi.org/10.3791/56942>.
- Hayes TL, Hagler S, Austin D, Kaye J, Pavel M. Unobtrusive assessment of walking speed in the home using inexpensive PIR sensors. In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE; 2009:7248-7251.
- Hagler S, Austin D, Hayes TL, Kaye J, Pavel M. Unobtrusive and ubiquitous in-home monitoring: a methodology for continuous assessment of gait velocity in elders. *IEEE Trans Biomed Eng*. 2009;57(4):813-820. <https://doi.org/10.1109/TBME.2009.2036732>.
- Lezak MD, Howieson DB, Loring DW, Fischer JS. *Neuropsychological Assessment*. USA: Oxford University Press; 2004.
- Reitan RM. Validity of the Trail Making Test as an indicator of organic brain damage. *Percept Mot Skills*. 1958;8(3):271-276. <https://doi.org/10.2466/pms.1958.8.3.271>.
- Wheeler D. *Wechsler Adult Intelligence Scale-Revised, Manual*. New York: Psychol Corp; 1981.
- Wechsler D. *Wechsler Memory Scale (WMS-III)*. San Antonio, TX: Psychological corporation; 1997.
- Kaplan E, Goodglass H, Weintraub S. *The Boston Naming Test*. Philadelphia, PA: Lea & Febiger. 1983.
- Donohue MC, Sperling RA, Salmon DP, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol*. 2014;71(8):961-970. <https://doi.org/10.1001/jamaneurol.2014.803>.
- Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. *Nat Methods*. 2015;12(3):179-185. <https://doi.org/10.1038/nmeth.3288>.
- Hayes TL, Larimer N, Adami A, Kaye JA. Medication adherence in healthy elders: small cognitive changes make a big difference. *J Aging Health*. 2009;21(4):567-580. <https://doi.org/10.1177/0898264309332836>.
- Austin J, Klein K, Mattek N, Kaye J. Variability in medication taking is associated with cognitive performance in nondemented older adults. *Alzheimer's Dement Diagnosis, Assess Dis Monit*. 2017;6:210-213. <https://doi.org/10.1016/j.dadm.2017.02.003>.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wu C-Y, Beattie Z, Mattek N, et al. Reproducibility and replicability of high-frequency, in-home digital biomarkers in reducing sample sizes for clinical trials. *Alzheimer's Dement*. 2021;7:e12220. <https://doi.org/10.1002/trc2.12220>