Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams

Richard Chen*
Apple Inc.

Filip Jankovic*
Evidation Health, Inc.

Nikki Marinsek*
Evidation Health, Inc.

Luca Foschini
Lampros Kourtis
Alessio Signorini
Evidation Health, Inc.

Melissa Pugh
Jie Shen
Roy Yaari
Vera Maljkovic
Marc Sunga
Eli Lilly and Company

Han Hee Song
Hyun Joon Jung
Belle Tseng
Andrew Trister
Apple Inc.

ABSTRACT

The ubiquity and remarkable technological progress of wearable consumer devices and mobile-computing platforms (smart phone, smart watch, tablet), along with the multitude of sensor modalities available, have enabled continuous monitoring of patients and their daily activities. Such rich, longitudinal information can be mined for physiological and behavioral signatures of cognitive impairment and provide new avenues for detecting MCI in a timely and cost-effective manner. In this work, we present a platform for remote and unobtrusive monitoring of symptoms related to cognitive impairment using several consumer-grade smart devices. We demonstrate how the platform has been used to collect a total of 16TB of data during the Lilly Exploratory Digital Assessment Study, a 12-week feasibility study which monitored 31 people with cognitive impairment and 82 without cognitive impairment in free living conditions. We describe how careful data unification, time-alignment, and imputation techniques can handle missing data rates inherent in real-world settings and ultimately show utility of these disparate data in differentiating symptomatics from healthy controls based on features computed purely from device data.

CCS CONCEPTS

• **Applied computing → Consumer health**.

KEYWORDS

Multimodal sensor data; cognitive impairment; real-world clinical studies; machine learning

ACM Reference Format:
Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, Marc Sunga, Han Hee Song, Hyun Joon Jung, Belle Tseng, and Andrew Trister. 2019. Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3292500.3330690

*Contributed equally.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3330690

1 INTRODUCTION

An estimated 5.7 million Americans and 46.8 million people worldwide live with dementia with a global cost of approximately $1 trillion [32]. Despite this prevalence, early diagnosis is a clinical challenge and is time consuming. Early symptoms are subtle, insidious, and easily dismissed as "normal aging" [5].

Common clinical screening tools for cognitive impairment, such as the Mini Mental State Examination [12] or the Montreal Cognitive Assessment [31], do not consistently detect the earliest stages of cognitive impairment [19]. More sensitive testing is limited by the need for highly specialized raters, lengthy duration of testing, rater bias, cultural and educational bias, and practice effects [17, 20]. Efforts to reduce these limitations have focused on computerization of assessments, such as the CogState CBB [29], however computerized tests are still limited (e.g. in their ability to discriminate the earliest forms of cognitive impairment) [3].

Other efforts have focused on porting the testing of specific cognitive domains from the clinical setting to apps through gamification [1, 24]. While leveraging ubiquitous computing devices may solve the issues of access to testing, such tests may introduce new limitations in the form of practice effects, limiting their clinical utility. Purely passive measurements would avoid practice effects, though often these measurements require complex installation of sensors within the home, limiting the scalability [22].

The near-continuous passive data collection of sensors in mobile devices and other consumer technologies can overcome these limitations and may have the potential to transform our ability to detect and track cognitive decline with minimal intrusion and burden [11]. Recognizing the potential utility for Real-World Evidence (RWE) for drug development, the US Food and Drug Administration (FDA) launched a framework to advance the use of RWE collection [14], including MyStudies App, a digital tool to help capture real world data from patients [13]. Informed by our experience with frequently measured app-derived data [34], a pragmatic approach is required

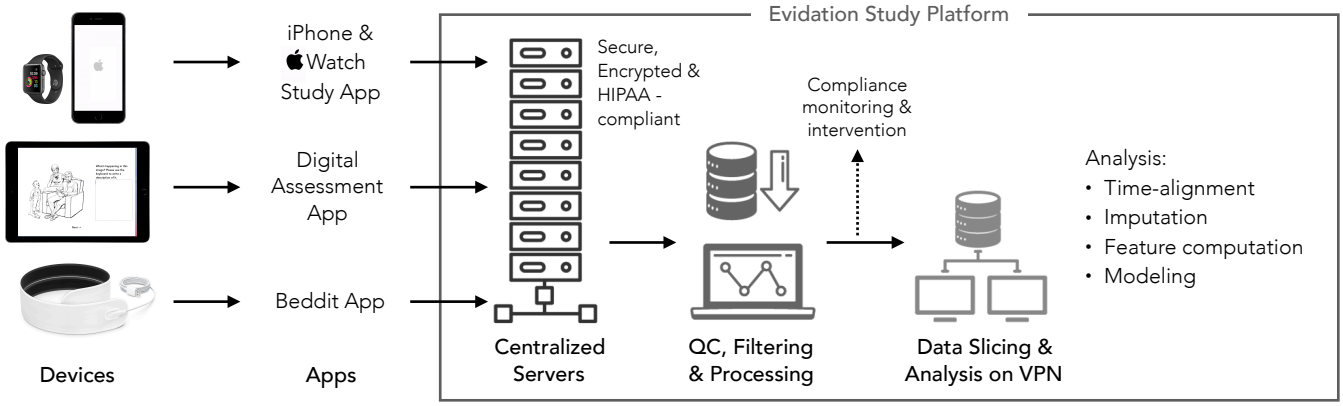


Figure 1: End-to-end data flow. Adapted from "Clinical Trials Transformation Initiative (CTTI) Recommendations: Advancing the Use of Mobile Technologies for Data Capture & Improved Clinical Trials" [9]. Released July 2018.

to temporally align data sampled at different rates and impute missing data. Particularly in the case of cognitive decline, missing data may, in itself, be a signal that should be captured.

Contributions. This study aimed to assess the feasibility of collecting data in individuals with and without cognitive impairment from multiple smart devices and test whether these data can differentiate between them. Our contributions are summarized as follows:

(1) We present a unified platform for remote and unobtrusive monitoring of potential symptoms related to cognitive impairment. This secure and compliant platform (Section 3.1) collects and harmonizes multi-modal high frequency data streams from multiple sensors on multiple consumer devices. Through the platform, we collected over 1.5 GB of data per participant per day on average, for a total of 16 TB of data during the course of this real-world study.
(2) We describe methods for effectively processing the data into meaningful features, including handling missing data and aligning data collected at different sampling rates.
(3) We demonstrate the utility of these processed features in distinguishing participants with symptoms of cognitive impairment (*symptomatics*) from healthy controls. We also explore which individual features make the strongest contributions to model outputs, to drive hypothesis generation for further investigations.

2 STUDY DESIGN

The Lilly Exploratory Digital Assessment Study was an IRB-approved multi-site 12-week exploratory study conducted by Evidation Health, Inc. on behalf of Eli Lilly and Company and Apple Inc.. The study aimed to assess the feasibility of using smart devices to differentiate individuals with mild cognitive impairment (MCI) and early Alzheimer's disease (AD) dementia from healthy controls. MCI is the clinically symptomatic, pre-dementia stage of AD in which cognitive deficits do not yet impair the ability to function at work or in usual daily activities.

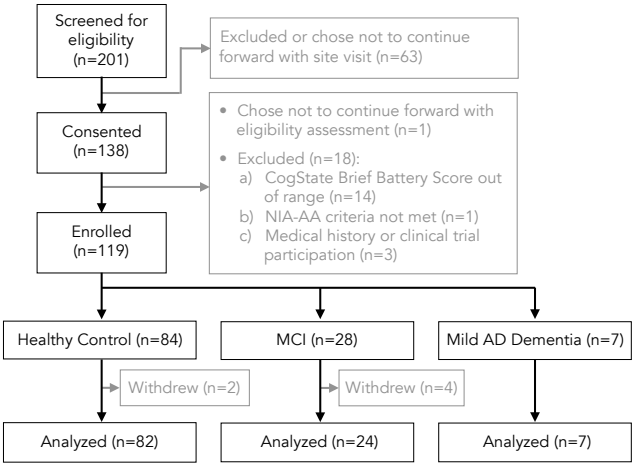


Figure 2: Flowchart of participants' enrollment.

2.1 Participant Screening and Enrollment

From December 2017 through August 2018, 201 potential participants initiated screening procedures and 138 of those individuals were consented and fully screened at 12 centers across the United States. 119 participants enrolled in the study. Key inclusion criteria were: (1) being 60-75 years old, (2) speaking English as their primary language, and (3) being familiar with digital devices, including currently having and actively using an iPhone and having an at-home WiFi network.

Participants with MCI and mild AD dementia had to meet the NIA-AA core clinical criteria for their respective AD disease states[21]. For symptomatic participants, a study partner was consented to monitor the compliance with study procedures.

Upon enrollment, each participant was provided an iPhone 7 plus (to be used as their primary phone), an Apple Watch Series 2, a 10.5" iPad pro with a smart keyboard, and a Beddit sleep monitoring device along with apps to collect all sensor and app-usage events during the 12 week study period.

Table 1: Sources of data collected in this study, along with their sampling rates and estimated sizes. Data size estimates are reported in MB collected per participant per day. *Data sources are outside the scope of this paper.

Domain	Hypotheses for Symptomatic cohort	Device	Datastream	Sampling frequency	Est. Size (MB)
Gross Motor Function	Poorer motor coordination, slower and more variable gait.	**Watch**	Accelerometer	100 Hz, while worn	> 200
			Gyroscope *	100 Hz, while worn	> 200
			Pedometer	2-5 seconds	~ 0.1
			Stairs climbed	Event-triggered	< 0.1
			Stand hours	Hourly	< 0.1
			Workout sessions	Event-triggered	< 0.1
		Phone	Accelerometer	100 Hz, continuous	> 400
			Gyroscope *	100 Hz, continuous	> 400
			Pedometer	2-5 seconds	~ 0.2
Autonomic	Impaired parasympathetic system activity.	**Watch**	Heart-rate	Seconds, dynamic	~ 0.2
Circadian Rhythm	Disruption in sleep cycle and daily routines.	**Beddit**	Sleep sensors	Multiple	~ 0.4
			Sleep summaries	Daily	< 0.1
		Phone	Energy survey	Daily	< 0.1
Behavioral, Social, and Cognitive	Increased withdrawal from social engagements, electronics usage, hobbies, etc. Over-reliance on helper apps due to difficulty with cognitive control and attention.	**Watch**	App usage	Event-triggered	< 0.1
			Phone unlocks	Event-triggered	< 0.1
			Message meta-data	Event-triggered	< 0.1
			Phone call meta-data	Event-triggered	< 0.1
			Breathe sessions	Event-triggered	< 0.1
		Phone	Distance	Event-triggered	< 0.1
			Mood survey	Daily	< 0.1
Fine Motor Control	Slower and more variable typing and tracing.	**Assessment App**	Tapping task	Bi-weekly	—
			Dragging task	Bi-weekly	—
Language	Impairments in language content and quality (pauses, grammatical errors, etc.).	**Assessment App**	Typed Narrative task	Bi-weekly	—
			Verbal Narrative task	Bi-weekly	—
			Video *	Bi-weekly	—
			Audio *	Bi-weekly	—

In all, 84 healthy controls and 35 symptomatic participants were enrolled (Figure 2). Participants were asked not to change any therapies for dementia or other medications that could affect the central nervous system over the course of the study, though this was not a requirement for participation.

2.2 Study Procedures

Over the course of the 12 weeks of data collection, participants were instructed to use their iPhone and Apple Watch as normal, and to keep them charged. Data from sensors in these devices and device usage, including phone lock/unlock, calls, messages, and app history, were passively collected by a bespoke study app and transmitted nightly to the study servers (Table 1). Central review of incoming data allowed for outreach when no data were received from devices. Participants with gaps in device data were contacted via email or phone to remind them to use their devices and to troubleshoot any problems.

Participants were also asked to answer two one-question surveys daily (one about mood, one about energy) as well as perform simple activities every two weeks on the Digital Assessment App. The app consisted of several low-burden active language and psychomotor tasks, including a dragging task in which participants dragged one shape onto another, a tapping task in which participants tapped a circle as fast as possible and then as regularly as possible, a reading task in which participants read easy or difficult passages, and a typed narrative task in which participants typed a description of a picture. These activities were selected because they have the potential to be monitored passively in the future. Study procedures included recording and transmitting video and audio of the participants while completing tasks on the Digital Assessment App. Though the analysis of audio/video data is outside the scope of this paper, we describe challenges on reliable collection of high-quality video for clinical purposes in the next section. At the conclusion of the 12 weeks of data collection, the devices were returned to the study center.

3 DATA PROCESSING

The collection and processing of high volume sensitive health-related data mandates high security standards and requires following strict protocols to comply with regulatory requirements and protect the privacy of the individuals involved. The complexity of this task increases with the number of sources, formats, and different sampling rates at which the data is collected (Table 1).

In this study, we used the Study Platform, developed by Evidation Health, Inc., to aggregate and analyze the data collected from the iPhone, Watch, and Beddit devices, as well as from the active tests performed on the iPad over the 12-week study period.

3.1 Study Platform

The Study Platform is a high-security environment designed to manage clinical studies, ingest and process device data, and provide

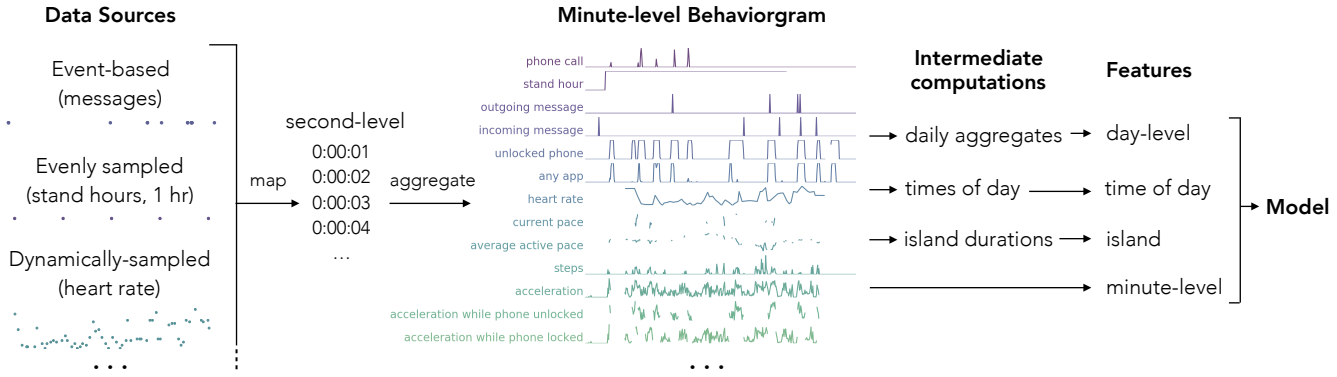


Figure 3: Preprocessing of Raw Data using Time Alignment and Data Imputation. Raw data sources (on the left) are imputed, time aligned, and combined into data channels of a behaviorgram.

a secure platform for analysis (Figure 1). The study platform stores participants' consents and data generated from eligibility screening to study completion, while monitoring compliance.

All data collected, including from sensors and smart phones, surveys, active tasks, and audio/video, were encrypted in transit and then stored on the platform data lake. Evidation's Study Platform uses a chain of custody for data that is compliant with the Health Insurance Portability and Accountability Act (HIPAA) and the FDA Code of Federal Regulations (CFR) Title 21 Part 11.

In total, the types of data collected from the daily use of these devices pose a potential concern to privacy for the participants. In addition to explicit consent to this risk during the enrollment, the data collected from each device was limited to reduce the risks of sensitive information being transmitted. For example, information about messages, calls, and social media usage was restricted to a binary indication of whether the application was in use. No private messages, voice calls, or other data packets were collected passively. There was an active recording of an interview that was used to monitor language with the express consent of the participant.

Data ingested by the platform was time-stamped (for compliance, by a third-party), checked for consistency, normalized to a standard schema (to facilitate data analysis) and saved using an optimized format in a distributed and replicated data store.

Researchers acessed the analytics portion of the data platform through a secure Virtual Private Network (VPN). Due to the large volume of data, the Study Platform used a distributed data processing system based on Apache Spark [37]. Data could not be accessed directly except by an internal interface that allowed researchers to request an encrypted, check-pointed slice of data. Depending on their role and permissions, each researcher was allowed to see subsets of the sources and data types available. The decryption key and the location of the data were only communicated to researchers.

3.2 Data Preparation

The collection of data from multiple sources posed several engineering and analytical challenges. Some input sources were sampled at a constant frequency (e.g. sleep quality data), while others were sampled only when relevant events happened (e.g. the time when a specific app was opened) or the sample frequency adapted to the context (e.g. sampling rates of pedometer and heart-rate measurements increased during high-activity and workout periods). Among the evenly-sampled data sources, sampling time ranged from day (e.g. surveys) to minute (e.g. aggregate physical activity) to sub-second (e.g. raw accelerometer sampled at 100Hz) intervals.

Behaviorgrams. As a first step to the analysis, we proceeded to map all event streams and time-series *data sources* (raw) into a common representation that we call a behaviorgram (Figure 3). A behaviorgram is comprised of time-aligned *data channels* (processed) with values at a 1-minute resolution. The behaviorgram succinctly represents the behavior of a participant in the study over time. In our case, the behaviorgram of a participant consisted of 65 data channels and 100,800 timepoints, corresponding to each minute in the 10 week period following enrollment [1].

Transforming the input source into the behaviorgram representation required time-alignment between channels, resampling of sources at different time scales, channel-aware aggregations, and careful handling of missing values. First, input sources timestamps were aligned in a timezone-aware fashion. Values from event-based sources were assigned to the second in which they occurred and either summed (for steps, stairs, missed calls, and messages) or averaged (for pace, stride, heart rate, and survey responses) to produce the minute-level-resolution sampling. Input sources representing intervals (e.g. for workout sessions, breathe sessions, stand hours, exercise, phone calls, phone unlocks, and app usage) were converted into minutes by encoding the fraction of the minute covered by the interval. We chose a minute-level resolution as the base resolution for the behaviorgram, following our experience on behavioral patterns associated with several health conditions manifesting at that timescale [34]. For domains that required sub-minute (or sub-second) precision (e.g. fine-motor functions) we first computed statistics at higher time-resolution before aggregating them to a minute-level resolution. For example, accelerometer measurements at 100Hz were aggregated into minute-level values by averaging the L^2 (Euclidean) norm of the X, Y, and Z accelerations taken at

[1] Only 10 total weeks of data were available during the writing of this manuscript, due to the week 11 & 12 still undergoing quality control

Table 2: Summary of aggregations applied to minute-level data during feature computation. Features for the active psychomotor tasks are not reported here. (Abbreviations: TOD, time of day; IQR, inter-quartile range, pctl: percentile)

	Feature Type			
Channel Type	**Minute-level**	**Time of day (TOD)**	**Day-level**	**Island**
Average Values *accelerometer, pace, stride, heart rate, sleep cycle, distance from home*	• 5, 10, 25, 50, 75, 90, 95th pctl, IQR, Skew, Fraction null	TODs of first, middle, last occurrences and peak: • Median, IQR	Daily 5, 50, 95th pctls, Fraction null: • 5, 50, 95th pctl, IQR, Skew	—
Counts *steps, stairs climbed, messages*	• Sum	TODs of first, middle, last occurrences: • Median, IQR	Daily sums: • 5, 50, 95th pctl, IQR, Skew	—
Fractions of a minute *workout sessions, breathe sessions, standing hours, exercise minutes, phone calls, apps, sleep stages*	• Sum	—	Daily sums: • 50, 95th pctl, IQR, Fraction non-zero	• Island durations • 5, 50, 95th pctl, IQR • Count
Impulses *missed calls, new apps, new contacts, top 3 contact*	• Sum	—	Daily sums: • 95th pctl, IQR, Fraction non-zero	—
Surveys *energy survey, mood survey*	—	TOD of survey: • Median, IQR	Daily response: • 5, 25, 50, 75, 95th pctl, IQR, Fraction null	—

each 100th of a second, after applying a low-pass filter to reduce the effects of gravity.

The behaviorgram has proven to be a helpful tool to explore patterns of associations between different channels. First, as a tool for data quality diagnostics, behaviorgrams allow inspecting missing data and outliers in one channel within the context of others. Second, as a data representation format, a behaviorgram makes it easy to capture interactions between different input data sources and may provide a means to conceptually replicate dual-task experiments that are administered in the lab or clinic. For example, previous studies have shown that individuals with dementia show greater impairment when they attempt to do two tasks at the same time (e.g. walking and having a conversation) than when they do a single task (e.g. only walking) [30]. With the behaviorgram representation, it's easy to add a channel that represents "walking while talking" at the minute level resolution by merging information from data channels that represent phone calls and average walking pace.

Missing Data. As the data in this study was collected in free living conditions over an extended period of time, there existed periods of no data collection due to participants not using or wearing the devices, being outside the vicinity of a sensor, or not participating in the active tests. We adopted a conservative approach to handling missing data. For on-event data streams in which we received data when an event was triggered (when an app was opened or a message was received, for example), we filled in minutes with no values with zero, which represented the absence of a triggering event in that minute. We also linearly interpolated heart rate within gaps shorter than 15 minutes, since heart rate was sparsely and dynamically sampled (heart rate was sampled more during high-activity periods, so aggregate values would be skewed toward higher values without imputation). We kept all remaining missing data as non-imputed. The choice to treat missing data as a signal was driven by the hypothesis that a person may demonstrate gaps in behaviors when they have cognitive impairment. As a result, we did not want to lose the potential signal that gaps in data might represent by imputing across missing values, a type of informative missingness [6].

Feature Computation. The features used in the Analysis (Section 4) were computed as time-aggregates over the behaviorgram data channels. We tailored the features we computed to the different types of data channels in order to create a set of interpretable, hypothesis-driven variables. We grouped data channels into five different channel types - average values, counts, intervals, impulses, and surveys - and computed four general types of features, consisting of aggregates of 1) all minutes, 2) the times of day of different events, 3) daily aggregates, and 4) the durations of continuous "islands" of activity (Table 2).

We also computed a set of 98 features from the data gathered from the psychomotor tasks in the Digital Assessment App. The features captured different psychomotor components, such as tapping speed, tapping regularity, typing speed, sentence complexity, drag path efficiency, and reading times. In total, we created 996 features, including 98 psychomotor features. Additional information about data processing is provided in Section 7 on reproducibility.

4 ANALYSIS

In this section, we demonstrate utility of the collected data in discriminating between individuals with and without cognitive impairment. We cast our problem as a machine learning regression task and report the performance of our models to differentiate participants as symptomatics vs. healthy controls.

4.1 Methods

We chose modeling techniques that provide direct interpretability of the results in feature space. Even if methods based on representation learning (e.g. deep learning) that directly model outcomes from the raw time series [26] are becoming increasingly popular in

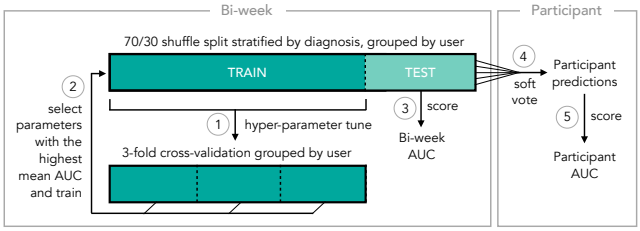


Figure 4: Diagram of model structure.

the medical machine learning community, interpretability of findings, model diagnostics, and overall complexity of model developed remain largely unsolved issues [16].

We trained models with the Extreme Gradient Boosting (XGBoost) algorithm, which allowed us to rapidly construct an ensemble of decision trees in a stage-wise fashion [8]. Advantages of XGBoost are that the boosting algorithm handles missing data and it can achieve low generalization error even when the number of features highly exceeds the number of samples. To leverage the additional information on severity of symptomatic participants (MCI vs. mild AD dementia), we trained the model using XGBoost Regression with a pairwise ranking objective function and the following labels: healthy control = 0, MCI = 1, and mild AD dementia = 2.

4.2 Data Augmentation

The small number of examples relative to the number of features poses a challenge in performing any machine learning task. To overcome this limitation we use a data augmentation technique, popular for data-inefficient applications such as CNN for image classification [27]. Our approach consists of computing features on non-overlapping subsets of 2-week periods for a total of 5 bi-weeks per participant: $BW_{i,1} \ldots BWi$, 5 for each user i with each bi-week $BW_{i,j}$ assigned the same label (healthy control or symptomatic) assigned to user i. This technique is sometimes referred to as Window Slicing in the Time Series Classification literature [24]. The scores returned by the model on $BW_{i,j}$ are then averaged ("soft-voting") into a final score for the user i.

We chose a two-week window because it provides a substantial boost in data size (increasing examples by a factor of 5x), while at the same time still capturing daily and weekly patterns within an individual [33]. A two-week window was also a natural choice for the features computed on the psychomotor tasks, which were administered every two weeks.

For all tasks we used a 100-repeat holdout procedure to evaluate out-of-sample generalization performance on classifying each bi-week as belonging to a healthy control or symptomatic participant. In each of 100 iterations, we split the dataset into train and test sets using a 70/30 shuffle split that was stratified by diagnosis (symptomatic vs. healthy control) and grouped by participant (bi-weeks from the same participant must all end up in the same set to prevent the model from memorizing a specific participant's pattern). We performed hyper-parameter tuning on the training set using grouped 3-fold cross validation. We used Hyperopt [4] to select the following parameters: number of estimators, learning rate, maximum tree depth, and gamma. For each combination of parameters, up to 30 combinations, we evaluated the performance of the model. The model hyperparameters that yielded the highest average Area Under the ROC Curve (AUROC) across the three folds were selected to train on the full training set in the outer split. We computed the bi-week model performance metrics on the held-out test set in the outer split. Then, in order to make predictions at the participant-level, we aggregated bi-week scores for a participant via soft-voting to rank each participant in the test set. The participant model performance metrics were computed on these scores. Finally, this procedure was repeated for 100 iterations to estimate average performance metrics and their associated errors.

A schematic of the modeling steps is illustrated in Figure 4. Additional details on model selection, parameter tuning, and alternative models used to reproduce results are discussed in the optional Reproducibility Section 7.

5 RESULTS

We measure performance using Area Under the Receiver Operating Characteristic curve (AUROC), averaged across splits. AUROC, which is optimized by ranking positive examples ahead of negative examples, is an appropriate metric of success for the intended application of targeting interventions. We also report Area Under the Precision-Recall Curve (AUPRC, computed as average precision over all possible recall thresholds), which is a more informative metric in our case where the emphasis is on accurate identification of the positives with a majority of negative samples [35].

At the participant level and on the full cohort, demographics alone are very discriminative between conditions, attaining AUROC of 0.757 (Table 3). The device-derived features alone obtained an AUROC of 0.771. Device-derived features alone were more precise on average than demographics alone (AUPRC=0.628 vs 0.546) in identifying symptomatic participants. The AUROC of the model increased to 0.804 (AUPRC = 0.701) when demographics were added to the feature set. When comparing AUROC and AUPRC scores between the demographics-only models and the models that included device-derived features, all scores were significantly different ($p<0.0001$), except for the demographics vs. device-derived features trained on the full cohort ($p=0.2$). Reported p-values for testing significance between differences of mean model scores were computed using a permutation test. We also repeated the training/test procedure on a dataset with randomly shuffled labels, and found that AUROC scores of biweek- and user-level models were not significantly different from a randomly performing model (AUROC 0.5).

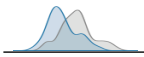
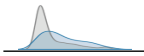
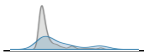
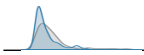
Age-Matched Cohort. Participant recruitment in this study was not age- and gender-matched, but the distributions of age and gender were monitored throughout enrollment and preferred not to exceed a 60/40 ratio for gender (in either direction) or an average difference of 2 years for age. Even so, due to difficulties with recruiting symptomatic participants, the symptomatic cohort was an average of 3 years older than the healthy control cohort.

In order to verify that the device-derived features were not detecting differences in behavior due to normal aging, we selected the nearest age-matched control within the healthy control cohort for every participant in the symptomatic cohort. Doing so produced

Table 3: Summary of modeling results.

		Healthy Control vs. Symptomatic				Healthy Control vs. Mild AD	
		Bi-weeks		Participants		Participants	
Cohort	Feature Set	AUC (±95 CI)	AUPrC (±95 CI)	AUC (±95 CI)	AUPrC (±95 CI)	AUC (±95 CI)	AUPrC (±95 CI)
Full	Demographics (Demo)	—	—	0.757 (±0.016)	0.546 (±0.020)	0.803 (±0.030)	0.327 (±0.027)
	Device features	0.739 (±0.014)	0.556 (±0.020)	0.771 (±0.016)	0.628 (±0.023)	0.933 (±0.016)	0.742 (±0.047)
	Device features + Demo	0.782 (±0.014)	0.650 (±0.020)	0.804 (±0.015)	0.701 (±0.021)	0.916 (±0.025)	0.804 (±0.050)
Age-matched	Demographics (Demo)	—	—	0.519 (±0.018)	0.536 (±0.012)	0.608 (±0.031)	0.294 (±0.020)
	Device features	0.704 (±0.018)	0.709 (±0.017)	0.726 (±0.021)	0.758 (±0.018)	0.897 (±0.027)	0.816 (±0.043)
	Device features + Demo	0.701 (±0.018)	0.705 (±0.018)	0.725 (±0.022)	0.754 (±0.020)	0.887 (±0.028)	0.799 (±0.045)

Table 4: Top 5 feature descriptions and cohort means for Healthy Controls (gray) and Symptomatics (blue).

	Data stream	Feature Description	Healthy Controls Mean (±95 CI)	Symptomatics Mean (±95 CI)
	Assessment App Typing Task	Typing speed with no pauses (keystrokes per minute).	115 (± 11)	87 (± 10)
	Pedometer (phone)	Median time of day of first active pace from phone.	7:42 am (± 0:19)	9:08 am (± 0:42)
	Energy Survey	Fraction of days with no energy survey response.	0.17 (± 0.07)	0.33 (± 0.08)
	Energy Survey	Median time of day of energy survey response.	10:38 am (± 0:35)	12:43 pm (± 0:52)
	Messages	Total number of messages received.	179 (± 68)	110 (± 43)

two age-matched cohorts of 31 symptomatics and 31 healthy controls with an average age difference of less than six months[2]. We re-ran the full analysis on these age-matched cohorts and report the results in Table 3. After controlling for age via matching, there was a large drop in performance for the demographics only models (AUROC=0.519, AUPRC=0.536). However, device-derived features still show moderate performance, with AUROC decreasing from 0.771 to 0.726 on the full cohort, and AUPRC increasing from 0.628 to 0.758. The boost in precision is mainly to be attributed to the change in class balance (31:31 in the age-matched case, vs 31:82 in the full cohort), which decreases the ratio between true positives and predicted positives. AUROC and AUPRC scores were significantly higher ($p<0.0001$, permutation test) for the models which included the device-derived features than the demographics-only models. Finally, model performance no longer improved when demographic features were added, indicating that device-derived features capture differences between healthy controls and symptomatic individuals that go above and beyond normal aging.

Mild AD dementia Cases. We additionally report the performance of the model when classifying healthy controls vs. individuals with mild AD dementia. Although results have been reproduced with different models (see Section 7 for details) we caution against optimistic interpretation due to the very small number of individuals with mild AD dementia in the symptomatic cohort (n=7). The rationale for making these comparisons is that detecting differences in cognitive impairment should be easier as impairment increases.

The model performed better when classifying individuals with mild AD dementia. Device-derived features and device features plus demographics achieved AUROC in the low 90s [3], (AUPRC = 0.804 for device-derived features and demographics). Using only device-derived features on the age-matched cohorts gets AUROC = 0.897.

5.1 Feature Importance

To understand the predictions made by XGBoost, we used a recent approach called SHapley Additive exPlanations (SHAP), which combines game theory with local explanations to explain machine learning models [28]. SHAP values are reported for an XGBRegressor model with a pairwise objective function (and default parameters otherwise) that was trained on the device-derived features for the age-matched cohorts. The SHAP values for the top 20 features are illustrated in Figure 5 and the top 5 features are described in Table 4. Overall, a few trends emerged that were important in identifying symptomatic individuals for our model:

[2]Due to age distributions of the two sub-cohort (see Section 7, attaining a higher ratio of healthy control vs. symptomatics was not possible, despite availability of more healthy controls. Similarly, matching on gender in addition to age also resulted in a drastic reduction of sample size.

[3]Due to the small number of mild AD dementia, 95% CIs of device-derived features and device-derived features + demographics are quite large, the inversion of AUROC when adding demographics to device-derived features is non-significant

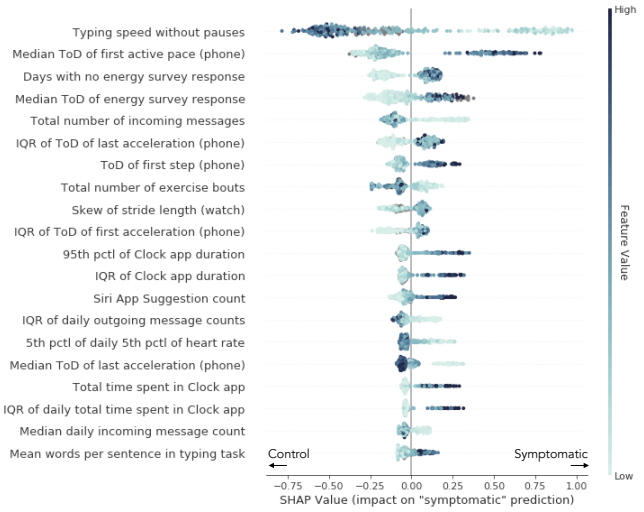


Figure 5: SHAP values of top 20 features of hand-tuned XGBoost model trained on the age-matched cohorts.

- **Slower typing:** Symptomatic participants tended to have slower typing than healthy controls. These results are in line with previous work [36], and may be the result of impaired fine motor control, difficulties with language, or both.
- **Less regularity and later first steps:** In general, symptomatic individuals exhibited less routine behavior compared to healthy controls, as measured by the larger interquartile range of the times of the first and last phone acceleration each day, which likely correspond to picking up the phone for the first time and setting down the phone for the last time each day, respectively. We also found that symptomatic participants tended to take their first step (as measured by the phone's pedometer) later in the day. Similar patterns have been observed in previous work as associated with MCI [25].
- **Fewer text messages:** Symptomatic participants received fewer text messages in total (and per day) and had a lower interquartile range of daily outgoing messages.
- **Greater reliance on helper apps:** Symptomatic individuals spent more total time in the Clock app than healthy controls and were more likely to view or access Siri's app suggestions.
- **Poorer survey compliance:** Symptomatic individuals answered the daily one-question surveys less often than healthy controls and, when they did respond, tended to respond later in the day.

6 DISCUSSION

The goal of this study was to assess the feasibility of collecting data in cognitively impaired individuals and healthy controls from multiple smart devices and to test whether the data can differentiate between these groups. We addressed the engineering and analytic challenges that accompany collecting large amounts of data from different devices and we adopted an approach that can appropriately handle data quality issues (including missing data) that are inherent in real-world settings. We also demonstrated the utility of using device-derived features to detect cognitive impairment in the small cohort of 31 symptomatics and 82 healthy controls included in the analysis, presenting a model achieving AUROC=0.80 using device-derived features and demographic data. To put our results into context, diagnostic AUROCs of computerized cognitive tests in analogous groups range from 0.67 to 0.97 [2]. In contrast, the authors of TATC [25] reported AUROC=0.62 in detecting MCI from actigraphy data only. Other digital assessments to discriminate between AD and healthy controls have been tested, including typing speed, speech and language, eye movements, and pupillary reflex [23]. Although individual sensors and domains show promise, no other study has yet created a digital signal to assess cognitive status from multiple sensors. Two ongoing projects are using passive data in Alzheimer's disease. The PRISM study employs an app to assess and characterize social withdrawal from passive data in specific diseases including AD dementia [10]. The RADAR-AD study measures disability progression associated with AD using smart phones, wearables, and home-based sensors (https://www.radar-ad.org/).

Since this is a feasibility study, we prioritized obtaining interpretable results that can be used in designing future studies. We also explored using TICC [18], which was recently adopted on another study on AD dementia using actigraphy data [25], but found that it was too sensitive to missing data to be applied to the current data set. These results are a starting point, and more accurate classification may be possible with longer longitudinal data, larger cohort sizes, and other advances in passive data collection. Among the next steps in the analysis of this dataset specifically, are more in-depth explorations of accelerometer, audio, and video data.

In the future, smart devices may be harnessed to monitor the symptoms of patients who have already been diagnosed with MCI or AD, detect individuals who may be vulnerable to developing MCI, test the effectiveness of current symptomatic therapies, accelerate the development of new therapies, or be used in conjunction with traditional diagnostic tools (such as medical history, imaging, cognitive tests, or self-reports) to improve the accuracy of dementia diagnosis. However, additional research and validation are needed before these applications become a reality. Privacy is of particular importance in any clinical application. Regulations such as the General Data Protection Regulation (GDPR) require applications dealing with longitudinal data to implement the "right to be forgotten." To comply, any implementation of these algorithms would require limiting the data collected centrally and providing users more on-device control.

Our approach is not without limitations. First, some of the patterns we found are associated with behaviors that are *modifiable*. Shifts in behaviors not associated with the progression of the underlying disease must be properly accounted for in future work. Further, there is the potential that a passive measure of cognitive performance could be self-reinforcing; without the knowledge of actions to take to mitigate any potential decline, the knowledge of the decline might cause decline itself.

Finally, we recognize potential risks in the creation of automated decision making tools trained on data whose distribution may not be representative of the target population, or may shift over time [15], and of the complex tradeoffs between fairness and accuracy of predictive modeling in the context of applications where human well-being is at stake, such as healthcare and criminal justice [7]. We believe a promising direction to address these challenges is to

minimize the cross-sectional nature of the model by considering applications to N-of-1 longitudinal settings, in which the system is set to detect changes of an individual's behavior relative to the behavior of the same individual in the past.

ACKNOWLEDGMENTS

The authors would like to thank Emily Fox at Apple, Cora Sexton at Lilly, and Jessie Juusola at Evidation for their feedback. The authors would also like to thank the study participants, investigators and the anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] Joaquin A Anguera, Jacqueline Boccanfuso, James L Rintoul, Omar Al-Hashimi, Farhoud Faraji, Jacqueline Janowich, Eric Kong, Yudy Larraburo, Christine Rolle, and Eric Johnston. 2013. Video game training enhances cognitive control in older adults. *Nature* 501, 7465 (2013), 97–101.

[2] Rabeea'h W Aslam, Vickie Bates, Yenal Dundar, Juliet Hounsome, Marty Richardson, Ashma Krishan, Rumona Dickson, Angela Boland, Joanne Fisher, Louise Robinson, et al. 2018. A systematic review of the diagnostic accuracy of automated tests for cognitive impairment. *International journal of geriatric psychiatry* 33, 4 (2018), 561–575.

[3] Russell M Bauer, Grant L Iverson, Alison N Cernich, Laurence M Binder, Ronald M Ruff, and Richard I Naugle. 2012. Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist* 26, 2 (2012), 177–196.

[4] James Bergstra, Dan Yamins, and David D Cox. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*. Citeseer, 13–20.

[5] Andrea Bradford, Mark E Kunik, Paul Schulz, Susan P Williams, and Hardeep Singh. 2009. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer disease and associated disorders* 23, 4 (2009), 306.

[6] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 6085.

[7] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? *arXiv preprint arXiv:1805.12002* (2018).

[8] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[9] Clinical Trials Transformation Initiative (CTTI). 2018. CTTI Recommendations: Advancing the Use of Mobile Technologies for Data Capture & Improved Clinical Trials - Data Flow Diagram. www.ctti-clinicaltrials.org/sites/www.ctti-clinicaltrials.org/files/data-flow-diagram.pdf

[10] Bruce N Cuthbert. 2019. The PRISM project: Social withdrawal from an RDoC perspective. *Neuroscience & Biobehavioral Reviews* (2019), 34–37.

[11] E Ray Dorsey, Michael V McConnell, Stanley Y Shaw, Andrew D Trister, Stephen H Friend, et al. 2017. The use of smartphones for health research. *Academic Medicine* 92, 2 (2017), 157–160.

[12] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. 1975. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12, 3 (1975), 189 – 198.

[13] US Food and Drug Administration (FDA). 2018. FDA launches new digital tool to help capture real world data from patients to help inform regulatory decision-making. www.fda.gov/NewsEvents/Newsroom/FDAInBrief/ucm625228.htm

[14] US Food and Drug Administration (FDA). 2018. Framework for FDA's Real-World Evidence Program. www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf

[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010* (2018).

[16] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Rajesh Ranganath. 2018. Opportunities in Machine Learning for Healthcare. *arXiv preprint arXiv:1806.00388* (2018).

[17] Terry E Goldberg, Philip D Harvey, Keith A Wesnes, Peter J Snyder, and Lon S Schneider. 2015. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1, 1 (2015), 103–111.

[18] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 215–223.

[19] S. Hoops, S. Nazem, A. D. Siderowf, J. E. Duda, S. X. Xie, M. B. Stern, and D. Weintraub. 2009. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology* 73, 21 (2009), 1738–1745.

[20] Diane Howieson. 2019. Current limitations of neuropsychological tests and assessment procedures. *The Clinical Neuropsychologist* 0, 0 (2019), 1–9.

[21] Clifford R. Jack, Marilyn S. Albert, David S. Knopman, Guy M. McKhann, Reisa A. Sperling, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* 7, 3 (may 2011), 257–262.

[22] Jeffrey A Kaye, Shoshana A Maxwell, Nora Mattek, Tamara L Hayes, Hiroko Dodge, Misha Pavel, Holly B Jimison, Katherine Wild, Linda Boise, and Tracy A Zitzelberger. 2011. Intelligent systems for assessing aging changes: home-based, unobtrusive, and continuous assessment of aging. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 66, suppl_1 (2011), i180–i190.

[23] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham Jones. 2019. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ Digital Medicine* (2019).

[24] C. Leurent, E. Pickering, J. Goodman, S. Duvvuri, P. He, E. Martucci, S. Kellogg, D. Purcell, J. Barakos, G. Klein, JW. Kupiec, and R. Alexander. 2016. A Randomized, Double-Blind, Placeo Controlled Trial to Study Difference in Cognitive Learning Associated with Repeated Self-administration of Remote Computer Tablet-based Application Assessing Dual Task Performance Based on Amyloid Status in Healthy Elderly Volunteers. 4 (2016), 280–281.

[25] Jia Li, Yu Rong, Helen Meng, Zhihui Lu, Timothy Kwok, and Hong Cheng. 2018. TATC: Predicting Alzheimer's Disease with Actigraphy Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 509–518.

[26] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

[27] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. 2017. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442* (2017).

[28] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[29] P. Maruff, E. Thomas, L. Cysique, B. Brew, A. Collie, P. Snyder, and R. H. Pietrzak. 2009. Validity of the CogState Brief Battery: Relationship to Standardized Tests and Sensitivity to Cognitive Impairment in Mild Traumatic Brain Injury, Schizophrenia, and AIDS Dementia Complex. *Archives of Clinical Neuropsychology* 24, 2 (mar 2009), 165–178.

[30] Manuel M Montero-Odasso, Yanina Sarquis-Adamson, Mark Speechley, Michael J Borrie, Vladimir C Hachinski, Jennie Wells, Patricia M Riccio, Marcelo Schapira, Ervin Sejdic, Richard M Camicioli, et al. 2017. Association of dual-task gait with incident dementia in mild cognitive impairment: results from the gait and brain study. *JAMA neurology* 74, 7 (2017), 857–865.

[31] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society* 53, 4 (2005), 695–699.

[32] World Health Organization et al. 2017. Global action plan on the public health response to dementia 2017–2025. (2017).

[33] Emma Pierson, Tim Althoff, and Jure Leskovec. 2018. Modeling Individual Cyclic Variation in Human Behavior. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 107–116.

[34] Tom Quisel, Luca Foschini, Alessio Signorini, and David C Kale. 2017. Collecting and analyzing millions of mhealth data streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1971–1980.

[35] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, 3 (2015), e0118432.

[36] G Stringer, S Couth, LJE Brown, D Montaldi, A Gledson, et al. 2018. Can you detect early dementia from an email? A proof of principle study of daily computer use to detect cognitive and functional decline. *International Journal of Geriatric Psychiatry* 33 (2018), 867–874.

[37] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. 2016. Apache spark: a unified engine for big data processing. *Commun. ACM* 59, 11 (2016), 56–65.

7 OPTIONAL SUPPLEMENT - REPRODUCIBILITY SECTION

To aid reproducibility, we borrow the datasheet for datasets [4] and model card [5] formats to report details about our dataset, data processing pipeline, and models.

7.1 Dataset

- **Why was the dataset created?** This dataset was created as part of the Lilly Exploratory Digital Assessment Study, which aimed to 1) assess the feasibility of collecting data via smart devices in a population with cognitive impairment and 2) explore the utility of the data in measuring cognitive impairment.
- **Who funded the creation of the dataset?** Eli Lilly and Company funded the creation of this dataset.
- **What are the instances? How many instances of each type are there?** Instances consist of participants who completed the study: 82 healthy participants, 24 with MCI, and 7 with mild AD.
- **How was the data collected?** The data was collected from provided iPhones, Apple Watches, and Beddit devices via bespoke study apps. Active tasks were performed within the Digital Assessment App on an iPad every two weeks at the participants' homes.
- **Over what time-frame was the data collected?** Data was collected over a 12 week period for each participant. In all, data was collected from December 2017 to November 2018.
- **Does the dataset contain all possible instances?** No, the data only contains a subset of the population, which is all older adults.
- **Is there information missing from the dataset and why?** Yes, due to technical issues, 35 participants do not have any distance data, 5 do not have any data from health-kit (heart rate, stand hours, stairs climbed), 10 do not have accelerometer data from the phone, and 9 do not have any Beddit data. Data coverage for other data sources may be sporadic due to device usage, proximity to sensors, traveling, etc.
- **Other comments about data collection?** The location data was processed prior to ingestion to maintain anonymity. GPS data was converted to the distance away from home in meters. No distance data was received if the home location was not set. For participants who did not manually set a home location, a home location was inferred after 30 days.
- **How is the dataset distributed? Who is supporting/ hosting/ maintaining the dataset?** The dataset is proprietary of Eli Lilly and Company and Apple Inc. and cannot be shared or distributed. Evidation Health Inc. is maintaining the dataset on its secure Study Platform.
- **If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?** Yes, all participants were explicitly told what data would be collected and consented to participate in the study. Participants were told that their data would remain confidential and an alphanumeric code would be used to identify their data. The study was approved by the Western IRB and Boston University IRB.

[4] Gebru, Timnit, et al. 2018. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010*

[5] Mitchell, Margaret, et al 2018. Model Cards for Model Reporting. *arXiv preprint arXiv:1810.03993*

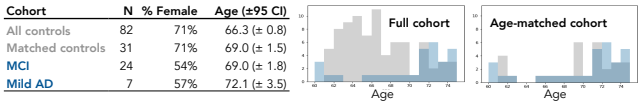


Figure 6: Demographics of the full and age-matched cohorts. Healthy controls in gray, Symptomatics in blue.

- **Does the dataset contain information that might be considered sensitive or confidential?** The full dataset includes information about demographics and medical history. We purposefully limited the type of data that was collected to reduce the collection of sensitive data. For example, distance data was processed prior to ingestion to remove information about location and we did not collect information about the content of phone calls or messages, just their timestamps.

7.2 Data Pre-processing: Behaviorgrams

This section describes the steps taken to convert the event stream and time-series data sources into behaviorgrams.

- To link participants' behaviors to times of day, we converted all timestamps from Coordinated Universal Time (UTC) to participants' local time. Conflicts that arose from timezone switches (due to travel or daylight saving time) were adjudicated by deleting earlier time points within the overlapping time periods.
- In order to map all participants' data onto a single time index, we converted all timestamps to the time elapsed since the midnight of participants' enrollment date.
- Timepoint data were mapped to second-level resolution evenly spaced time-series and the values falling within each minute were either summed (for steps, stairs, missed calls, and messages) or averaged (for pace, stride, heart rate, and survey responses) to produce the minute-level-resolution sampling.
- Data associated with time intervals were similarly converted to a minute-level resolution time series that represented the fraction of the minute spent doing an activity (for workout sessions, breathe sessions, stand hours, exercise, phone calls, phone unlocks, and app usage) or the average value during the minute (for distance away from home).
- We converted the 100 Hz raw accelerometer data to minute-level aggregates by taking the L^2 (Euclidean) norm of the X, Y, and Z accelerations at each timepoint, applying a low-pass filter to reduce the effects of gravity, and averaging the resulting values within each minute.

The following preprocessing steps were applied to the behaviorgrams prior to feature computation:

- Heart rate was linearly interpolated only within gaps of 15 minutes or less. This was done because heart rate was sparsely and dynamically sampled (more measurements were collected during high-activity bouts), so some features would be skewed towards higher values without imputation.
- Stride length was normalized by dividing by participants' height in meters.
- Upon inspection of the behaviorgrams, it was discovered that the phone accelerometer showed a spike of activity every two hours (but at different times across participants). To remove these

spikes, minutes of the day with a 5th percentile acceleration > 0.05 across the study days were replaced with NaNs and linearly interpolated.
- Sleep stage data were associated with start timestamps but no end timestamps, making the sleep stage durations unreliable. The sleep cycle data was used to clean the sleep stage data by setting the sleep stage channels to NaN for minutes with no sleep cycle data.
- Data was transmitted continuously for some of the evenly-sampled data channels. To aid the time-of-day computations and to avoid creating features that were biased by device usage levels, we set sleep cycle values equivalent to 0 and phone accelerometer values < 0.005 to NaN.

7.3 Model card

7.3.1 Model Details.

- The model was trained to differentiate individuals with and without MCI/mild AD, given features computed from two-weeks of passively collected data and performance on a set of psychomotor tasks. Biweekly predictions were averaged to predict diagnosis at the individual-level.
- The data set was divided into training (70%) and test (30%) sets and 3-fold cross validation was performed within the training set only to tune hyper-parameters. The hyper-parameters that produced the highest mean AUROC across the 3 folds were used to train the full training set and model performance was evaluated on the test set. The outer 70/30 split was performed 100 times, shuffling between iterations. Both the outer and inner splits were grouped by participant and the outer split was stratified by diagnosis.
- Developed by researchers at Apple, Lilly, and Evidation Health in 2018-2019.
- XGBRegressor with labels healthy control=0, MCI=1, mild AD=2.

7.3.2 Intended Use.

- Intended to be used by researchers to determine the feasibility of detecting cognitive impairment with passive data and to identify promising data sources.
- Not intended to be used for diagnosis or treatment decisions.

7.3.3 Factors.

- Potential factors that may influence the performance of the model include the severity of individuals' MCI or AD diagnoses, individuals' baseline cognitive abilities and behaviors, additional demographic information such as education or employment status, device usage, study compliance, and the time period on which passive features are computed (e.g., daily vs. weekly vs. monthly).

7.3.4 Metrics.

- The Area Under the Receiver Operating Characteristic (AUROC) curve and the Area Under the Precision-Recall curve (AUPRC) were used to measure model performance. We report the mean and 95% confidence intervals of the AUROCs and AUPRCs across all 100 outer shuffle splits.

7.3.5 Training Data.

- Ten weeks of data were split into five biweeks and features were computed on each biweek. An additional set of biweekly features were computed from a battery of digital psychomotor tasks.

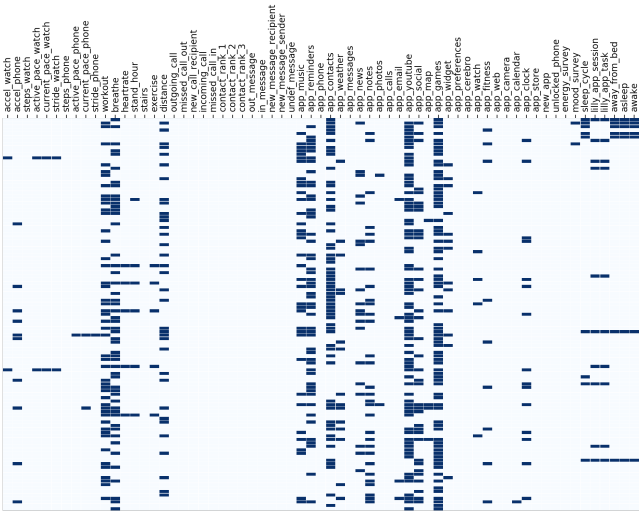


Figure 7: Data coverage for channels in the behaviorgram (columns) for all participants (rows). Channels with all NaN values are in dark blue. Channels with the most all-null values include workout and breathe sessions (which are user-initiated) and various apps (which may not be installed).

- 70% of the dataset was used to train the tuned model in each outer shuffle split. Scikit-learn's GroupShuffleSplit was used to split the data.

7.3.6 Evaluation Data.

- 30% of the dataset was used to evaluate the tuned model in each outer shuffle split.

7.4 Hyper-parameter search space

We used Hyperopt to select the best hyper-parameters for the XGBRegressor model within the given search space:

- Number of estimators: 100 to 400 with a step size of 100
- Learning rate: 0.05 to 0.20 with a step size of 0.05
- Max depth: 2 to 10 with a step size of 1
- Gamma: 1 to 10 with a step size of 1

7.5 Software and Hardware specifications

The following package versions were used (Python 3.5): XGBoost: 0.81; Scikit-learn: 0.19.1; Shap: 0.27.0; Pandas: 0.23.4; Numpy: 1.15.1; Hyperopt: 0.2. The model iterations were run in parallel on a server with 40 cores and 160 GB of RAM.

7.6 Independent Replication

The healthy control vs. symptomatic modeling results were independently reproduced using XGBoost's XGBClassifier with hand-tuned parameters (learning_rate=0.01, n_estimators=300, max_depth=4, gamma=2.0, reg_alpha=1.0, min_child_weight=5, subsample=0.85, colsample_bytree=0.8) and 60/40 outer grouped and stratified shuffle splits. All average AUROCs were reproduced within AUROC=±0.031.