Individualized Tracking of Neurocognitive-State-Dependent Eye-Movement Features Using Mobile Devices

Individualized Tracking of Neurocognitive-State-Dependent Eye-Movement Features Using Mobile Devices

HSIN-YU LAI, Harvard University, USA
CHARLES G. SODINI, Massachusetts Institute of Technology, USA
VIVIENNE SZE, Massachusetts Institute of Technology, USA
THOMAS HELDT, Massachusetts Institute of Technology, USA

With current clinical techniques, it is difficult to assess a patient's neurodegenerative disease (e.g., Alzheimer's) state accurately and frequently. The most widely used tests are qualitative or only performed intermittently, motivating the need for quantitative, accurate, and unobtrusive metrics to track disease progression. Clinical studies have shown that saccade latency (an eye movement measure of reaction time) and error rate (the proportion of eye movements in the wrong direction) may be significantly affected by neurocognitive diseases. Nevertheless, how these features change over time as a disease progresses is underdeveloped due to the constrained recording setup.

In this work, our goal is to first understand how these features change over time in healthy individuals. To do so, we used a mobile app to frequently and accurately measure these features outside of the clinical environment from 80 healthy participants. We analyzed their longitudinal characteristics and designed an individualized longitudinal model using a Gaussian process. With a system that can measure eye-movement features on a much finer timescale in a broader population, we acquired a better understanding of eye-movement features from healthy individuals and provided research directions in understanding whether eye-movement features can be used to track neurocognitive states.

CCS Concepts: • **Applied computing** → **Bioinformatics**.

Additional Key Words and Phrases: mobile health monitoring, longitudinal model, saccade latency, saccade directional error rate, neurodegenerative diseases, Gaussian Process

ACM Reference Format:
Hsin-Yu Lai, Charles G. Sodini, Vivienne Sze, and Thomas Heldt. 2023. Individualized Tracking of Neurocognitive-State-Dependent Eye-Movement Features Using Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 19 (March 2023), 23 pages. https://doi.org/10.1145/3580843

1 INTRODUCTION

The ability to objectively, accurately, and frequently track neurocognitive state is important. For example, drowsy driving contributes to 9.5% of all crashes [37]. An objective assessment of neurocognitive state may help reduce the rate of accidents. Neurocognitive states also degrade over the progression of neurodegenerative diseases. Since it is believed that treatments are more effective in early disease stages [16, 41], with the increase in life

Authors' addresses: Hsin-Yu Lai, hsinyul@fas.harvard.edu, Harvard University, Cambridge, Massachusetts, USA, 02138; Charles G. Sodini, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02139, sodini@mtl.mit.edu; Vivienne Sze, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02139, sze@mit.edu; Thomas Heldt, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02139, thomas@mit.edu.

2474-9567/2023/3-ART19 $15.00
https://doi.org/10.1145/3580843

expectancy and the lack of disease-modifying medications for many diseases (e.g., Alzheimer's disease), tracking neurocognitive states becomes a pressing need.

However, current clinical assessment of neurocognitive states requires trained specialists, is mostly qualitative, and is commonly done only intermittently [18, 31]. Therefore, these assessments are affected by an individual physician's acumen and by confounding factors such as a patient's level of attention [34]. Quantitative, objective, and more frequent measurements are needed to mitigate the influence of these factors. On the other hand, blood, cerebrospinal fluid sampling, and brain imaging techniques are obtrusive; we need more accessible methods to detect early disease stages.

A promising candidate for a quantitative and accessible neurocognitive-state monitor is pro/anti-saccadic eye movement [26], a fast eye movement moving towards or away from a stimulus. Saccadic features are observed to be significantly different between healthy individuals and patients with neurodegenerative diseases [1, 47]. In addition, it is studied that several brain regions that are crucial in generating pro/anti-saccades (such as frontal cortex and basal ganglia) may be affected by neurodegenerative diseases [19, 26, 33].

However, since these pro/anti-saccade features are commonly measured clinically with high-speed, IR-illuminated cameras and under controlled conditions (chinrest, lighting), the accessibility is limited. With only one or two data-collection sessions, the results are often pooled across participants. While we see on average there are differences in eye-movement features between normal elderly participants and neurodegenerative-disease patients, there are often significant overlaps between the ranges of these measurements, which highlights the importance of individualized tracking of eye-movement features. Nevertheless, the constraints on the recording environments also limit the number of longitudinal studies. While some studies showed promising results that saccadic features may change over disease progression [2, 42], the measurements of these studies were usually too sparse (less than or equal to twice per year) to detect neurodegenerative disease onset or efficiently evaluate treatment effects.

An alternative to this approach could be afforded by performing eye-movement tracking and analysis at the convenience of the patient on mobile devices with user-facing cameras. In fact, the use of such "digital biomarkers" has recently attracted significant attention in neurology [4, 11, 12, 24, 34] (where biomarkers refer to biological signs for a disease or a condition). In a recent work [22, 23], an instructive and easy-to-use iOS app was developed to allow a user to record themselves in their own homes and offices while they are performing a pro-saccade or an anti-saccade task with an iPad. In addition, a robust and automated measurement pipeline was proposed to measure two of the most widely measured saccadic features in the clinical literature from the recordings: saccade latency (time difference between a stimulus presentation and the initiation of the corresponding eye movement) and error rate (the proportion of eye movements towards the wrong direction) [9, 14, 32, 46]. Using the app and measurement pipeline in [22, 23], we have collected over 6,800 videos and over 235,000 individual eye movements from 80 self-reported healthy participants across the adult age spectrum.

Our goal in this paper is to study how pro/anti-saccade latency/error rate change over time in healthy individuals. In particular, we want to answer the following questions: a) How large are the day-to-day variations? b) How do these features change over time and how are these features correlated? c) Is there any trend in these features? By knowing the characteristics of eye-movement features in healthy individuals, we can put into context how neurocognitive impairment may affect these eye-movement features. In addition, if we can characterize these features in a statistical model, we may be able to extend the model for disease progression modeling. Different from previous works on eye-movement features where analyses were often based on data pooled across participants from one or two recording sessions, we study individualized longitudinal characteristics. The contributions are as follows: a) We observe that there is significant inter-subject variability in the day-to-day variations in these eye-movement features. b) When considering the median pro/anti-saccade latency and error rate per day, we observe that some participants present significant linear correlations across these eye-movement features that may be related to their task-performing strategies. c) We show that when we have more than 25

days of recordings, our individualized longitudinal model outperforms the baseline where we assume the features in healthy individuals are fixed over time. By characterizing these longitudinal eye-movement features from healthy individuals, our study provides future directions on evaluating the possibility to use these features to track neurocognitive states objectively, accurately, and frequently.

2 RELATED WORK

Multiple directions of work are related to our research. First, several studies were conducted to understand how clinical biomarkers may be affected by neurodegenerative-disease progression. However, since these biomarkers rely on cognitive tests, neuroimaging techniques, and cerebrospinal fluid analysis, the assessments may not be sufficiently quantitative, objective, or frequent to identify early or even prodromal stages. Digital biomarkers such as gaits, finger tapping, and saccadic movements are promising unobtrusive measurements to help detect disease onsets. While mobile-device-based monitors have been proposed to measure gaits and finger tapping, mobile-device-based saccade measurements were still underdeveloped. Since currently most saccade measurements rely on clinical environments, longitudinal studies on saccade measurements were also underdeveloped. Therefore, to the best of our knowledge, our work is the first to enable saccade measurements using mobile devices and is the first to have characterized how these measurements change over days in healthy individuals. (While we focus on ocular biomarkers, eventually it may be best to include multiple modalities for monitoring disease progression as explained in [38].) Finally, since our ultimate goal is to potentially track the neurocognitive states using eye-movement features, we gave a concise review on existing disease-progression models (including models developed for other diseases).

2.1 Clinical Biomarkers of Neurodegenerative Diseases

One of the most studied open dataset for the clinical biomarkers of Alzheimer's disease (AD) is the Alzheimer's Disease Neuroimaging Initiative (ADNI) [50]. The study has started in 2004 and has collected a relatively comprehensive set of biomarkers. The Coalition Against Major Diseases (CAMD) provided another online dataset for AD [35]. Large-scale cohort studies are also being developed for other diseases including Parkinson's diseases (PD) [25] and Huntington's diseases (HD) [39, 48]. Most of these studies focused on measurements instrumented by physicians. Therefore, the measurements were usually sparse in time.

2.2 Digital Biomarkers of Neurodegenerative Diseases

To enable frequent and quantitative measurements of neurodegenerative diseases, several mobile-device monitors have been proposed to measure digital biomarkers, including mobile-device monitors for AD [20, 40], gait-based monitors for PD[10, 54], apps for self-reported PD symptoms [27], or apps that incorporated multiple modes such as tapping, voice, and walking [6]. The main focus of most of these studies is to detect a disease rather than to track the progression of cognitive states.

2.3 Eye-Movement Features in Clinical Literature

Eye-movement features are another promising digital biomarker. Saccadic eye movements, for example, require attention to the environment as well as appropriate decision-making and execution of oculomotor responses once a stimulus is registered. This stimulus-response paradigm probes cognitive and oculomotor function, both of which can be impaired in neurocognitive diseases [1]. In addition, in clinical studies, several saccadic features are observed to be significantly different between healthy individuals and patients [28].

Eye-movement features were also studied in healthy individuals to evaluate the reliability of these features both in the context of test/retest reliability and their correlations with cognitive load. Recently, a large-scale study on young healthy adults showed that eye-movement features can be a robust signature [3], which suggests

that monitoring these features over time may help detect changes in the health of an individual. Evidence of eye-movement features as potential cognitive-state markers can also be found in studies that tested the effect of fatigue [29] and cognitive demands [30].

While all of these studies suggest that eye-movement features are promising features for monitoring disease progression, there are two main differences between their results and our work. First, these studies were mostly conducted in one or two recording sessions. In addition, the results were often pooled across participants. As discussed in [3], eye-movement features are promising signatures for each individual. It is thus important to understand how these features change over time in each individual. By understanding the longitudinal characteristics in healthy individuals, when we start collecting longitudinal eye-movement data from patients, we may better discover patterns associated with disease progression.

2.4 Disease Progression Modeling

There are several approaches to disease progression modeling – a graphical model [52], a Gaussian process (GP) model [8, 13, 44, 45], and an recurrent neural network [51]. We use GP to develop individualized longitudinal models for the eye-movement features we collected from healthy participants for three reasons. First, it can capture the correlation over time and the correlation across the features. Second, GP is a nonparametric model and its complexity can be adapted to the complexity of the training data. That is, compared to a linear model which can only characterize a linear function, GP can characterize an infinite dimensional function. Thus, it is more flexible than any model consisting of a finite number of basis functions. Third, because any finite samples from GP form a Gaussian distribution, the computation for learning and inference is relatively simple (when compared to other nonparametric models). Therefore, a GP model provides interpretability, flexibility, and computability. An in-depth overview of GP models can be found in [7]. Our models are special cases of a multi-task GP model [5], which is known as linear models of coregionalization (LMC) in the geostatistics literature [15]. With the amount of data we have collected, we develop a model similar to that in [36]. However, we carefully designed the hyperparameters and whether they should be individualized or shared across the participants based on the characteristics of the eye-movement features. This design allowed us to enable individualized tracking of saccade latency and error rates from healthy individuals.

3 MATERIALS

In this section, we summarize the recruitment efforts, the recording setup, the task design, the measurement pipeline, and the data collection summary. The procedures follow [22] where details can be found.

3.1 Recruitment

A person can participate in one or multiple recording sessions. Each recording session consists of three pro-saccade tasks and three anti-saccade tasks. (Six recordings in total.) If a participant chooses one recording session, each task will consists of 40 stimuli. If a participant chooses to take part in multiple recordings sessions, the participant will be asked to take the recordings everyday for at least two weeks while they can choose 20 stimuli per task. Fig. 1 shows the number of participants in one or multiple recording sessions. In total, there are 80 self-reported healthy adult participants, ranging in age from 20 to 92 years, in this study. Video recording of volunteers was approved by MIT's Committee on the Use of Humans as Experimental Subjects (protocol # 1711147147), and informed consent was obtained from each participant before recording. Most participants were students, professors, staffs from MIT, and their family members.

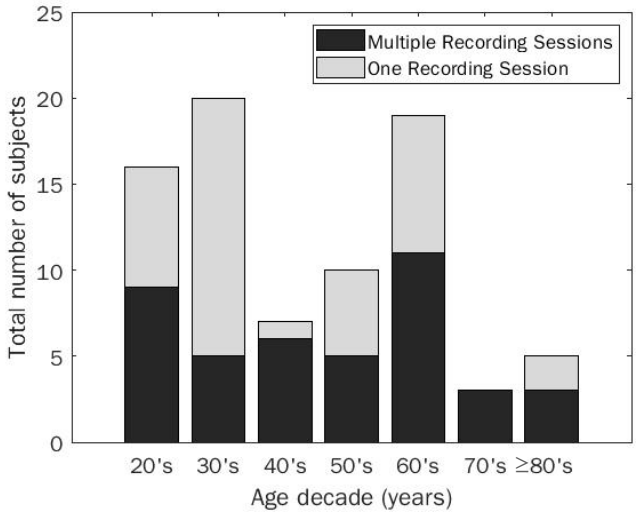

Fig. 1. Age distribution of participants with single or multiple recording sessions.

3.2 Recording Setup

In [22], an iOS app is designed to guide a participant to use an iPad (Generation 2 and 3) to record themselves with sufficient lighting and with appropriate distance to the camera in the comfort of their homes or offices. The app first asks the participant to enter their Participant ID. It then asks the participant to choose whether they would perform a pro- or anti-saccade task with 20 or 40 stimuli. As shown in Fig. 2, the app then displays the participant and a rectangle on the screen and instructs the participant to align their face with the rectangle. This is to ensure that the participant is within a desirable distance (30-50 cm) to the camera. If the iPad can measure distance, the rectangle will be green if the participant is within the proper range of distance and red if otherwise. In addition, to ensure proper lighting, the participant will be asked to move to a brighter location if the automatically detected ISO is greater than 1000. When the participant is ready, they can start the recording. The participant will then be recorded by the frontal camera while performing the task shown on the screen. After the task, the recording and a meta file will be saved for measurements of saccade latency and error rates. Details can be found in [22].

3.3 Task Design

We implemented a gap-pro-saccade and a gap-anti-saccade task as in [9, 14, 43]. As shown in Fig. 3, both tasks start with a fixation period. During the fixation period (1 s), a participant is instructed to look at the fixation point (a green square at the top center of the screen). Followed by the fixation period is a 200-ms gap period, where the screen is black. After the gap period is a 1.2-s stimulus period. A stimulus (white square) would appear on either left or right side of the screen. A participant is instructed to look towards/away from the stimulus as quickly and accurately as possible in a pro/anti-saccade task. After that is another 200-ms gap period. This sequence of "fixation-gap-stimulus-gap" will repeat for 20 or 40 times, with half of these stimuli on the right and half on the left in randomized order.

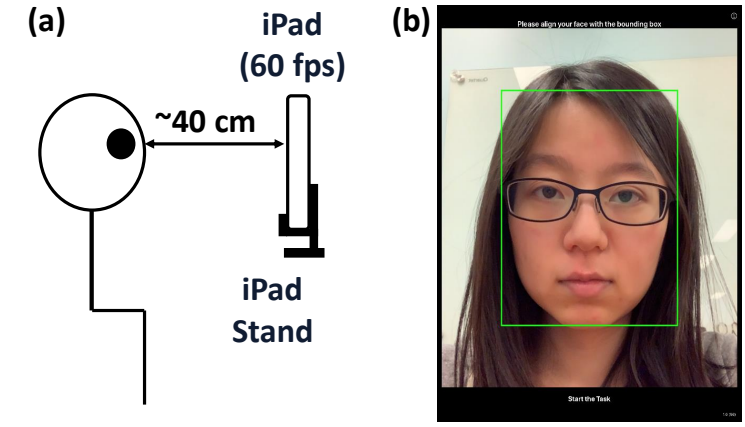

Fig. 2. (a) Recording setup; (b) before showing the task, the app displays the face of the participant with a bounding box. If the distance measurement from the camera to the participant's face is accessible (i.e. between 30 and 50 cm), the box will turn green. If the automatically detected ISO is greater than 1000, a warning will be shown to guide the participant to move to a better-illuminated place.

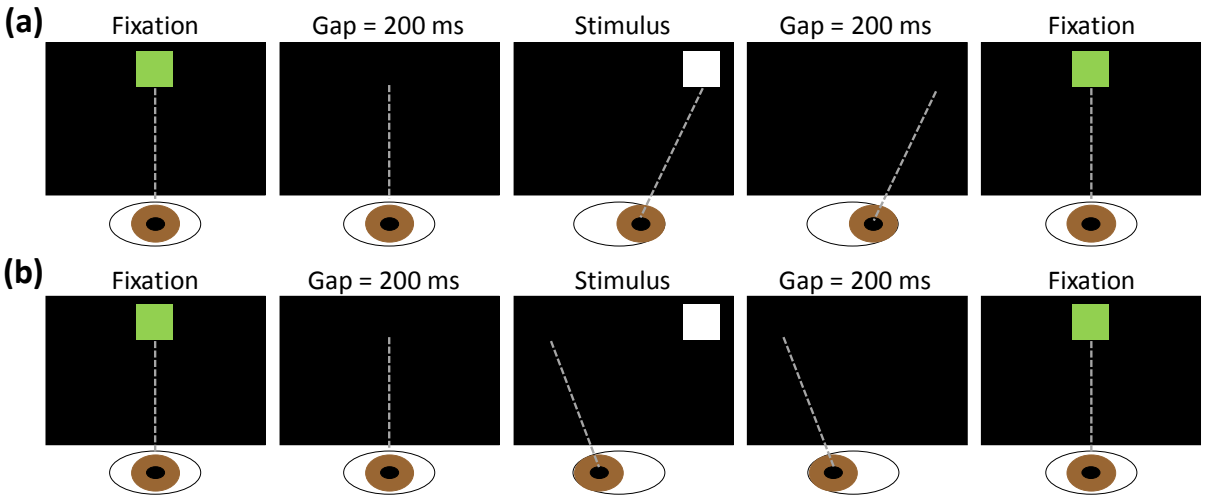

Fig. 3. (a) Pro-saccade task: Look toward the stimulus. (b) Anti-saccade task: Look away from the stimulus.

3.4 Measurement Pipeline

Our measurement pipeline is shown in Fig. 4. The pipeline first uses an eye-tracking algorithm [21, 23] to estimate where a participant is looking at on the screen from 200 ms before to 800 ms after each stimulus presentation and generate a saccade trace per stimulus. As discussed in [22], then the pipeline automatically characterizes each trace into "Declared a Low Signal (dLS)", "Declared an Error (dE)", and "Declared a Correct Saccade (dC)" using the algorithm described in [22]. The dLS traces typically consist of eye movements that are visually hard to tell whether they are correct saccades or directional errors, for example, due to eyelid droops. Therefore, as explained

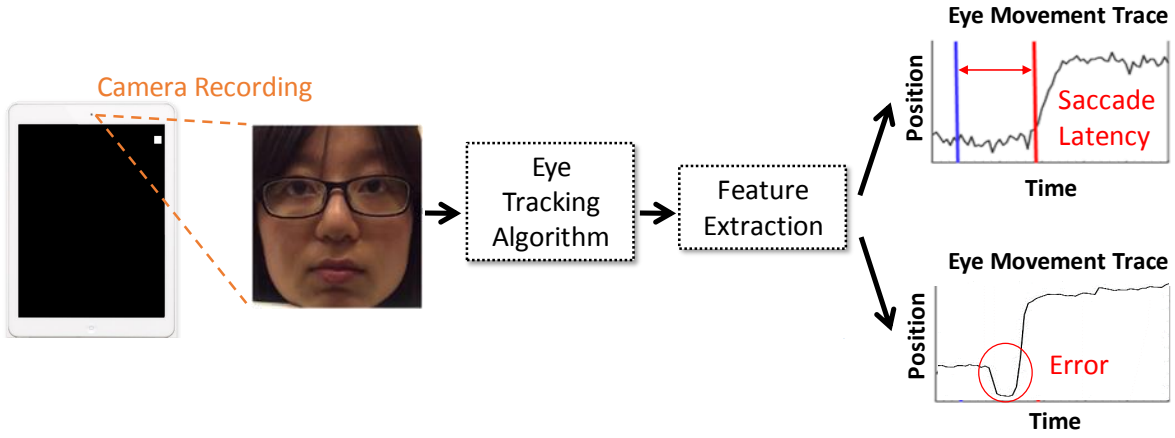

Fig. 4. The measurement pipeline includes the tablet-based video recording, an eye tracking algorithm, a saccade-latency measurement algorithm, and an error detection algorithm.

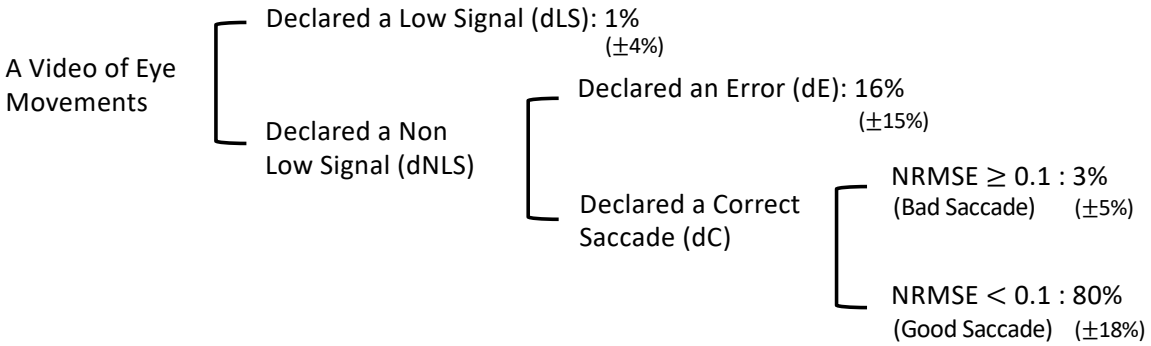

Fig. 5. Breakdown of saccades into declared low signal (dLS) and declared non low signal (dNLS). Breakdown of dNLS into declared error saccades (dE) and declared correct saccades (dC). Breakdown of dC into good saccades and bad saccades.

in [22], we then estimate the directional error rate as #dE/(#traces-#dLS). The pipeline further automatically characterizes each dC into a good or a bad saccade. A bad saccade refers to an eye movement where the algorithm considers it as a correct saccade but cannot accurately measure its saccade latency, e.g., due to the existence of a head movement. Therefore, bad saccades are excluded from the analysis of saccade latency.

3.5 Data Collection Summary

With a flexible system, we have collected 6,823 videos and 236,900 eye movements from 80 participants across the adult age spectrum. We observe that in videos with a substantial number of dLSs, participants' eyes were often partially occluded due to eyelid droop. Videos with a large number of bad saccades tend to contain more head movements. As a result, the number of dLSs and bad saccades indicates whether a participant recorded themselves properly. We therefore discard a video if more than half of the saccades are dLSs or bad saccades.

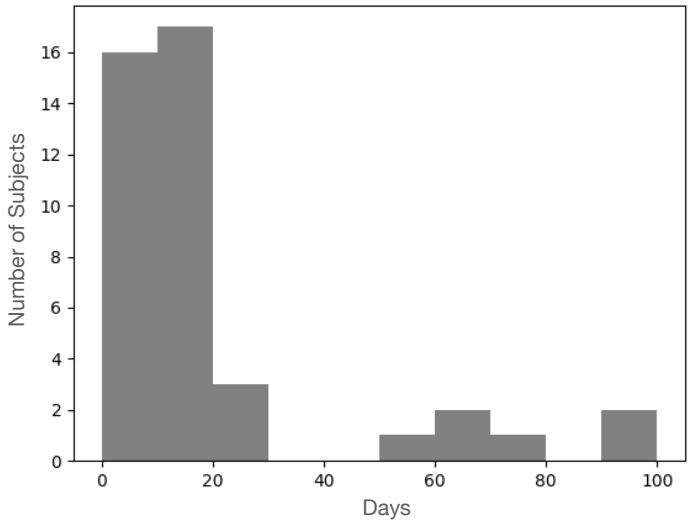

Fig. 6. Distribution of the number of days of recordings from participants with multiple recording sessions.

After discarding the videos with too many dLSs and bad saccades, we retained 6,787 videos and 235,520 eye movements from 80 participants. As shown in Fig. 1, there are 42 participants with multiple recording sessions. In Fig. 6, we show the distribution of the number of days of recordings per participant with multiple recording sessions.

4 EYE-MOVEMENT CHARACTERISTICS

The motivation of our work is to track individual eye-movement features over time and analyze how these features correlate with the neurocognitive states. To achieve this goal, we need to first understand how eye-movement features change over time in healthy individuals. In particular, in this section, we analyze the inter-subject variability in the day-to-day variations of the features. Then we examine the characteristics of individual day-to-day variations.

To analyze how eye-movement features vary over time, we group the measurements by days. For each day of measurements, we calculate four eye-movement features – median pro-saccade latency, pro-saccade error rate, median anti-saccade latency, and anti-saccade error rate. We use the median rather than the mean to reduce the impact of outliers. The size of the day-to-day variations can be calculated by the standard deviation of these daily eye-movement features. Fig. 7 shows the distribution of the standard deviations from participants with more than five days of recordings. We notice significant inter-subject variability in the day-to-day variations, which highlights the importance of individualized tracking of eye-movement features.

Since we assume that these day-to-day variations are not caused by disease progression, they may be introduced by measurement errors, changes in the task-performing strategies, and fatigue effects. As discussed in [53], measurement errors can be classified into random errors and systematic errors. Since random errors affect the measurements of each saccade randomly, they only contribute to the variations within a day. On the contrary, systematic errors bias all measurements in a recording session. Thus, these errors contribute to the day-to-day variations. These systematic errors can, for example, be caused by the differences in the recording setup.

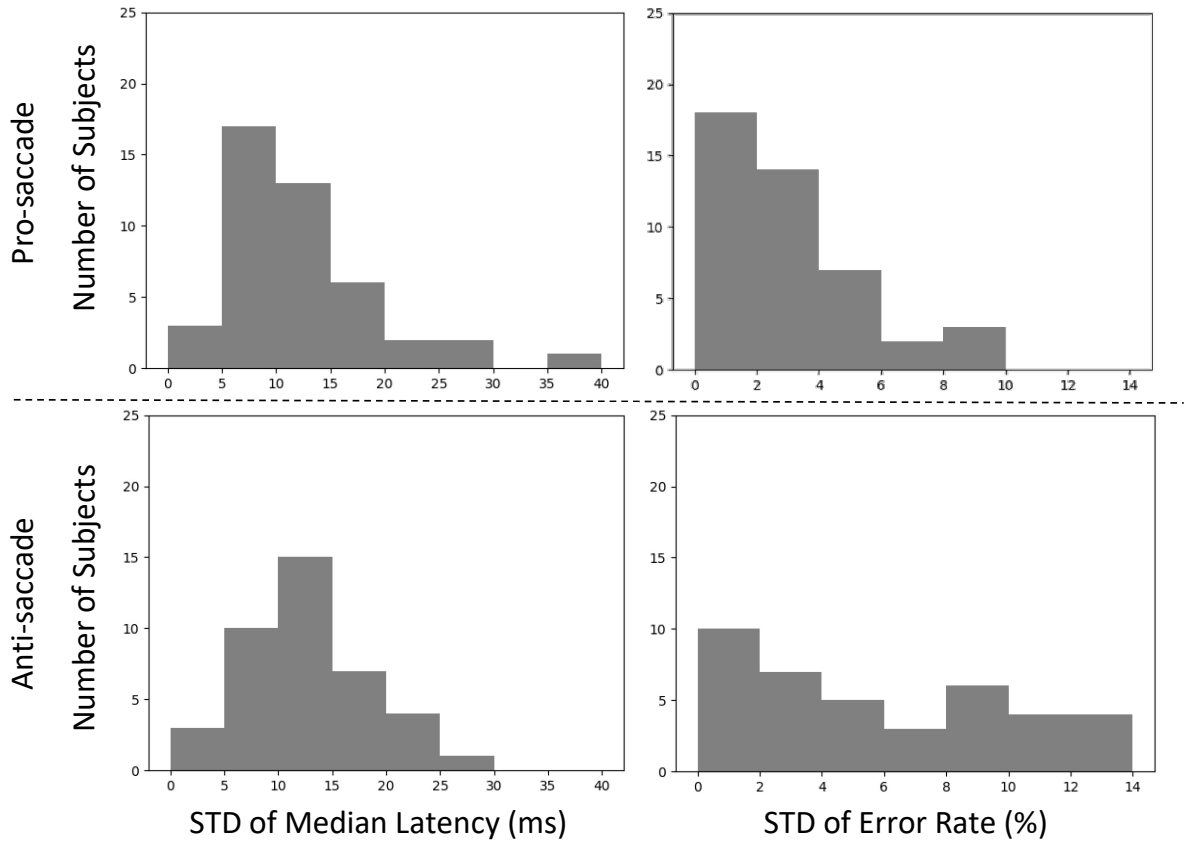

Fig. 7. The histogram of the standard deviation of four daily eye-movement features – pro/anti-saccade latency/error rate from participants with more than five days of recordings.

To analyze the size of the random errors, we use bootstrapping to estimate the variations within a day. Fig. 8 shows the four eye-movement features over days from two example participants with the 95% confidence interval estimated by bootstrapping. We see that the variations across days are larger than the variations within a day. Therefore, we know that random errors cannot fully explain the day-to-day variations. On the other hand, we explained in [22] how we minimize the variations in a recording setup (e.g., the distance to the camera and the lighting condition) by providing guidance in the app. This design should reduce systematic errors.

Besides systematic errors, the day-to-day variations can also be caused by a participant's task-performing strategy. This effect can be illustrated by Participant 4 in Fig. 8. We observe that the trajectories of pro/anti-saccade latency are similar. The trajectories of pro/anti-saccade error rate are also similar. However, the changes in the trajectories of latency and error rate are opposite to each other. For example, we notice that latencies measured around Day 35 are larger whereas error rates measured around Day 35 are smaller. We hypothesize that the participant was trading off between accuracy and speed when performing the tasks. That is, by moving their eyes faster, a participant can attain a lower latency and a higher error rate, and vice versa.

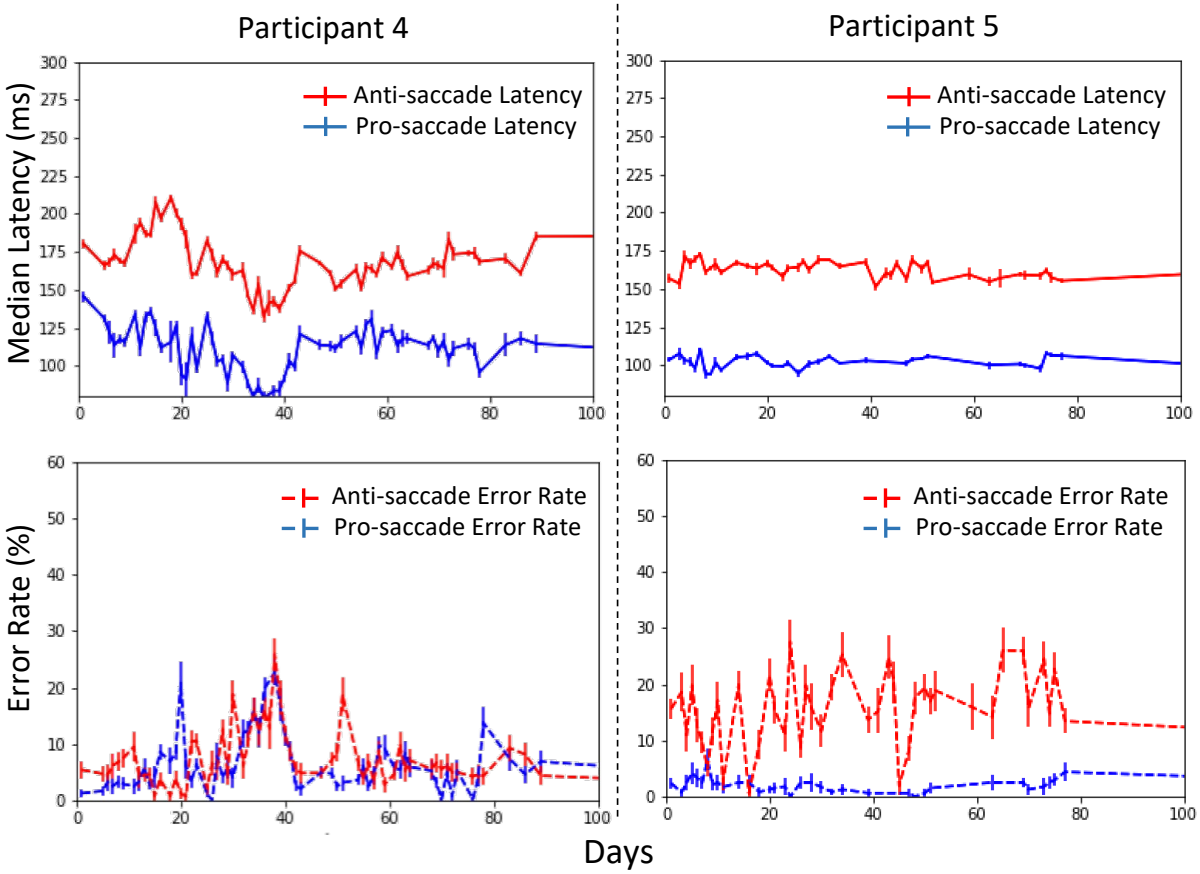

Fig. 8. Median saccade latency and error rate over days from two participants. The error bars indicate 95% confidence intervals. Here the index numbers for the participants follow the experiment result shown in Figure 11 where Participant 4 and 5 are the participants with the fourth and fifth most data in the experiment.

However, not every participant has a clear strategy. As shown in Fig. 8, Participant 5's strategy is not as clear as Participant 4's. A strategy naturally introduces correlation across features. Fig. 9 shows the Pearson correlation coefficients across the eye-movement features from five example participants. Here, Participant 1 and 4 present a trade-off between latency and error rate. The strategies in Participant 2, 3 are slightly different from the latency and error rate trade-off. It is not clear what Participant 5's strategy is. To design an individualized longitudinal model, we need to design individualized parameters to learn the correlations across eye-movement features to account for these differences.

Besides the differences in each participant's correlation across features, how these features correlate over time is also different among individuals. In Fig. 8, we observe that Participant 4's eye-movement features clearly correlate over time. Task-performing strategies and tiredness may be the cause of this correlation. By contrast, the anti-saccade error rate in Participant 5 in Fig. 8 changes over time more abruptly. The individualized model we develop should be able to learn these various characteristics.

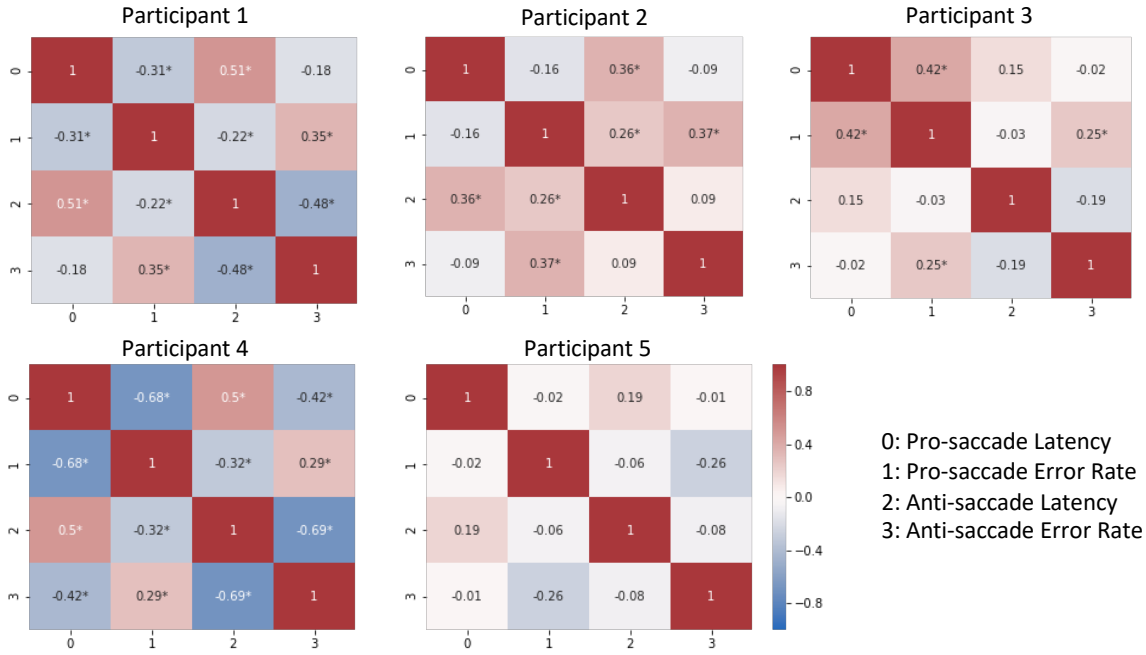

Fig. 9. The Pearson correlation across the four eye-movement features from five example participants. Stars mark the significance.

5 LONGITUDINAL MODELS

In this section, we propose three candidate GP models to characterize longitudinal eye-movement features from healthy individuals and test their performances on the data we collected. Since there is no neurocognitive state for us to predict, the performance metric is the ability to "characterize" normal eye-movement features so that once data are collected from patients, one may identify the effect of disease progression.

5.1 Data Preprocessing and Notations

As in Section 4, we group the measurements by day. We calculate the median pro/anti-saccacde latency and pro/anti-saccade error rate per day. Since the day-to-day variations vary across participants, we normalize each participant's data by the mean and variance before fitting the data to the model. There are several implications from this preprocessing step. To begin with, with the variations normalized, the model is designed to learn the shape of the longitudinal data instead of the size of the day-to-day variations. Moreover, since the size of the day-to-day variations is the same across participants after the normalization, we can share some hyperparameters of the model across the participants to avoid over-fitting. However, if the size of the day-to-day variations are indications of difference neurocognitive states, we may need to modify the model.

Before introducing the candidate models, we first define the notations. We consider eye movement features $y_p = \{y_{pi}\}_{i=1}^{4}$ where $y_{pi} = \{y_{pin}\}_{n=1}^{N_p}$, p denotes the p-th participant, i denotes the i-th feature, n denotes the n-th day of measurements, and N_p denotes the number of days of measurements from the p-th participant. We denote the corresponding day of measurements as $t_p = \{t_{pn}\}_{n=1}^{N_p}$. Notice that we can still use GP to do inference if there is any missing measurements. Such condition may happen when 1) a participant decides to only take pro-saccade

tasks or anti-saccade tasks in a day 2) the recordings are discarded because more than half of the saccades are dLSs.

5.2 Model Setup

With the notations, we can present our candidate models. In addition, we provide some remarks about the strengths and the weaknesses of these models.

5.2.1 *Baseline Model.*

$$p(y_p|g_p) = \prod_{i=1}^{4}\prod_{n=1}^{N} N(y_{pin}; \mu_{pi}, \sigma_i^2), \tag{1}$$

where μ_{pi} is the mean of the i-th feature from the p-th participant.

Remark: The baseline model assumes the day-to-day variations can be modeled as random noise. Therefore, the correlation over time and the correlation across the eye-movement features are assumed to be zero.

5.2.2 *Multi-task Model.*

$$p(y_p|g_p) = \prod_{i=1}^{4}\prod_{n=1}^{N} N(y_{pin}; w_{pi}g_p(t_n), \sigma_i^2), \tag{2}$$

where

$$g_p \sim GP(0, K^g(t, t')), \tag{3}$$

and $K^g(t,t') = (a^g)^2 \exp\{-\frac{|t-t'|}{l^g}\}$.

Remark: This model is a simplification of the multi-task model [5]. There are two reasons why we choose a simplified form. The first reason is interpretability. Motivated by Participant 4 in Fig. 8, we assume that there is an underlying process $g_p(t_n)$ shared across the four eye-movement features, and that the scale w_{pi} of this underlying process on each feature is associated with each participant's task-performing strategy. For example, for Participant 4, the sign of w_{pi} for the pro-saccade latency will be the same as the sign for the anti-saccade latency but opposite to the sign for the error rates. The second reason is to avoid overfitting. With the number of data we have per participant, learning four individualized parameters per participant is a reasonable choice. For the day-to-day variations that cannot be explained by a shared process, this model assumes that they are caused by random noise (σ_i^2).

In contrast to w_{pi} that is learned per participant per feature, the hyperparameters σ_i^2, a^g and l^g are shared across participants. (Notice that the g here is used as a notation rather than a power.) We notice that if for participant p, the effect size of the shared process on feature i is smaller than another participant, i.e., $|w_{pi}| < |w_{pi'}|$, then since a^g is shared across participants, $|w_{pi}g_p(t)|$ will also be smaller. However, since the random noise σ_i is shared across participants, it cannot become larger to compensate for a smaller $|w_{pi}g_p(t)|$. Therefore, this model may suffer if the effect size of the shared process on the features is not uniform across participants.

5.2.3 *Feature-specific Model.*

$$p(y_p|h_p) = \prod_{i=1}^{4}\prod_{n=1}^{N} N(y_{pin}; h_{pi}(t_n), \sigma_i^2), \tag{4}$$

where

$$h_{pi} \sim GP(0, K_i^h(t, t')), \tag{5}$$

and $K_i^h(t,t') = (a_i^h)^2 \exp\{-\frac{|t-t'|}{l_i^h}\}$.

Remark: This model assumes that all features are independent. This assumption contradicts with the observation in Fig. 8. While one may still use this model to predict the values of missing eye-movement features, this model cannot learn the correlation across the features and thus cannot learn individualized strategies.

5.2.4 Mixed Model. Motivated by the limitations in the presented multi-task model and the feature-specific model, we designed a mixed model as follows:

$$p(y_p|g_p) = \prod_{i=1}^{4}\prod_{n=1}^{N} N(y_{pin}; w_{pi}g_p(t_n) + h_{pi}(t_n), \sigma_i^2), \tag{6}$$

where

$$\begin{aligned} g_p &\sim GP(0, K^g(t,t')), \\ h_{pi} &\sim GP(0, K^h_{pi}(t,t')), \\ K^g(t,t') &= \exp\{-\frac{|t-t'|}{l^g}\}, \\ K^h_{pi}(t,t') &= a_i^2(1-\tilde{w}_{pi}^2)\exp\{-\frac{|t-t'|}{l_i^h}\}, \end{aligned} \tag{7}$$

and

$$w_{pi} = a_i\tilde{w}_{pi}, \tilde{w}_{pi} \in (-1,1). \tag{8}$$

Remark: As noted in the remark in the multi-task model, the multi-task model assumes that except the shared process, the day-to-day variations are caused by random noise. In addition, the multi-task model assumes that the effect size of the shared process is uniform across participants. These assumptions contradict with the intuition shown in Fig. 8, where the data from Participant 4 can be explained by a shared process but not the data from Participant 5. Thus, in this model, we include a feature-specific GP. We notice that if a participant's $|\tilde{w}_{pi}|$ is small, then the term $|w_{pi}g_p(t)|$ will be small. However, with a smaller $|\tilde{w}_{pi}|$, the covariance function of the $h_{pi}(t)$ will be larger. As a result, $|h_{pi}(t)|$ can be larger. That is, $|\tilde{w}_{pi}|$ not only controls how the four features are correlated, it also controls the effect size of the shared process.

5.3 Model Learning

The hyperparameters include $\{w_{pi}, \tilde{w}_{pi}, a^g, a_i^h, a_i, l^g, l_i^h, \sigma_i\}$. As in [7], these hyperparameters are learned by maximizing the likelihood functions. The maximization is performed using gradient descent with momentum (learning rate= 0.001 and momentum= 0.9). In particular, the signal variances and noise variances are both initialized to be one. The weights are initialized by the main principal component of the estimated linear correlation matrix using the training data. The length scales are initialized as twenty. All the hyperparameters are then re-parameterized to range from minus infinity to infinity before we perform the gradient descent. The length scales are the only hyperparameters where we set a lower bound in the learning process, and we set it to be two. We do so to ensure some correlation across the time and avoid overfitting.

While all the other initial values make sense, the initial values of the length scales can seem arbitrary. Our simulation result shows that most length scales converge to around five days. The only exceptions are the length scales for the pro-saccade latency error rate in the feature-specific model and the mixed model and the length scale for the shared process in the mixed model; these length scales converge to the lower bound. Since we do not have sufficient participants, we do not draw special attention to this result. However, we imagine that it is worth looking into in the future.

5.4 Model Evaluations

To evaluate the candidate models, we use two performance metrics – normalized L2 error and normalized log-likelihood. Let (t_*, y_*) be the testing data and say the algorithm predicts the values at t_* to be distributed as $N(\mu_*, \Sigma_*)$. The normalized L2 error is defined as $\frac{\|y_* - \mu_*\|_2}{\|y_*\|_2}$. Notice that since we remove the mean before fitting the data, for the baseline model, we have $\mu_* = 0$. Therefore, the normalized L2 error for the baseline model is one. However, the normalized L2 error does not quantify the uncertainty estimate Σ_*. To incorporate the uncertainty estimate, we can define the normalized log-likelihood as the log-likelihood normalized by the number of non-missing entry in y_*. A model performs well when the normalized L2 is small and the normalized log-likelihood is large.

Since we are assuming that there is no neurodegenerative-disease progression, our goal is to characterize the eye-movement features rather than predict the disease states. In other words, we are trying to test whether the characteristics we see in Section 4 are learned by the models as well. To do so, we first evaluate the performance of these models over different numbers of days of recordings. After we understand how many days of recordings is sufficient to characterize a participant's eye-movement features, we next evaluate how well the candidate models characterize the correlation across the features from participants with sufficiently many data. Finally, we evaluate whether a linear trend should be included in the model to account for learning effects.

5.4.1 Number of Days of Recordings. In order to understand how many days of recordings is needed, we analyze participants with more than 60 days of recordings. As shown in Fig. 6, there are six participants with more than 60 days of recordings. However, we only analyze data from five participants and remove one participant because the participant's pro-saccade latency is larger than the anti-saccade latency; we are uncertain whether the participant understood the task. To test the performance of the models with $N = 15, 25, 35, 45, 60$ days of recordings, we keep the first N days of recordings and perform 3-fold cross validation with days of recordings missing at random. For each fold, we average over the participants and acquire one normalized L2 value and one normalized log-likelihood value. In Fig. 10, we average over the three folds, and the error bars mark the maximum and the minimum values from the three folds.

Several observations can be made. First, we notice that GP-based models outperform the baseline when there are more than 25 days of recordings regarding both normalized L2 and log-likelihood. Since the baseline model does not assume correlation over time, this observation suggests that it is beneficial to consider correlation over time. That is, we can characterize eye-movement features from healthy individuals better than assuming that healthy individuals have fixed eye-movement features over time. In addition, with more than 25 days of recordings, a mixed model performs the best, followed in order by the feature-specific model, the multi-task model, and the baseline. Without assuming the correlation across the features, the feature-specific model can still predict a missing data point using its neighboring data. However, the mixed model still outperforms the feature-specific model since the correlation across the features can help the prediction and reduce the uncertainty in the prediction. As explained in the remark in Section 5.2, the multi-task model assumes that besides the shared process across all four features, all the other day-to-day variations are caused by noise. As a result, it cannot utilize the correlation over time as flexibly as a feature-specific model. When there is no clear shared process across all four features, it can only perform as well as the baseline model. Therefore, it generally performs worse than the mixed model and the feature-specific model.

5.4.2 Correlation across Features. We notice that a GP model can use the correlation over time to predict the missing data from the neighboring data. To evaluate how well a model characterizes the correlation across the features, we remove a continuous segment of a feature instead of randomly removing data as in the previous experiment. In this case, a GP model cannot use the neighboring data to predict the missing data but to use the other features. More precisely, for each participant, we cut the data into three segments and remove the middle

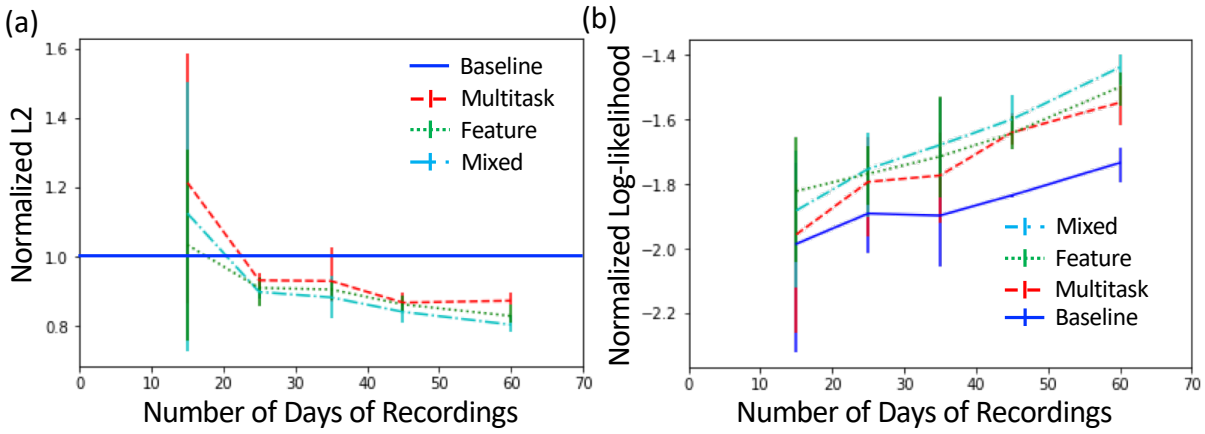

Fig. 10. The performance of the baseline and the three GP models with different number of days of recordings regarding (a) normalized L2 and (b) normalized log-likelihood. The experiments were performed using 3-fold cross validation. The error bars show the maximum and minimum values from the three folds.

segment of each of the four features, one at a time. Thus, there are two thirds of the recordings with four features intact and one thirds of the recordings missing one feature. We test it on participants with more than 45 days of recordings since in there would be 30 (>25) days of recordings with the four features for the models to learn the correlation. We then average over the four features and show the performance of the models in Fig. 11.

We notice that regarding normalized L2, the performance of the feature-specific model is comparable to the baseline in all five participants. This is to be expected since the model assumes all the features are independent. As shown in Fig. 9, almost all the features from Participant 1 and Participant 4 are significantly correlated. As a result, we see in Fig. 11 that the multi-task model and the mixed model perform better than the baseline in Participant 1 and Participant 4. To observe it more closely, we show in Fig. 12 how the missing pro-saccade values from Participant 4 are predicted by the three GP models. As shown in Fig. 8, it is clear that the missing data can be predicted from the anti-saccade latency. Since the feature-specific model fits each model independently, as shown in Fig. 12(b), it can only assume the pro-saccade latency increases gradually from Day 25 to Day 60. On the contrary, with the assumption of a shared process, both the multi-task model and the mixed model can predict the trend of the missing data using the anti-saccade latency. In addition, since the mixed model is more flexible than the multi-task model regarding the effect size of the shared process, we see that the mixed model performs better than the multi-task model.

As for normalized log-likelihood, we see that the mixed model generally performs the best. However, in Participant 5, the baseline model performs the best. If we look at the correlations across Participant 5's features in Fig. 9, we notice that the features are not significantly correlated. As a result, a baseline model may be least prone to over-fitting.

We further compare the learned correlation from the mixed model with the estimated correlation from the data. (The learned correlation refers to the linear correlation implied by the mixed model.) As shown in Fig. 13, the model can learn the signs of the correlation correctly if the correlation is significant. However, we also notice that the learned correlations tend to be smaller in general when compared to the estimated correlations from the data. It may be due to the fact that we simplify the modeling of the correlation across the features as a rank-one matrix plus noise.

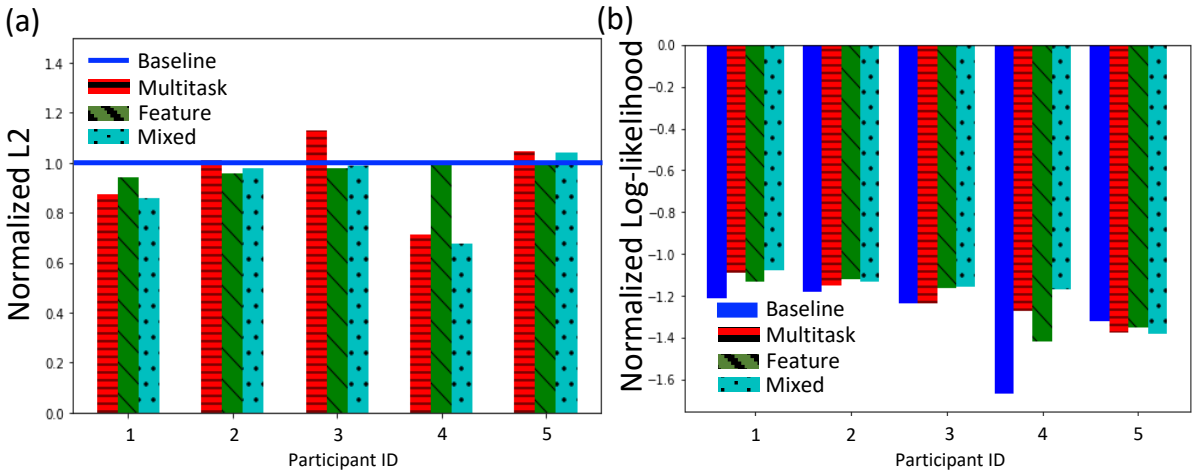

Fig. 11. The performance of the baseline and the three GP models on participants with more than 45 days of data regarding (a) normalized L2 and (b) normalized log-likelihood. Participants are ordered by their number of recordings in decreasing order.

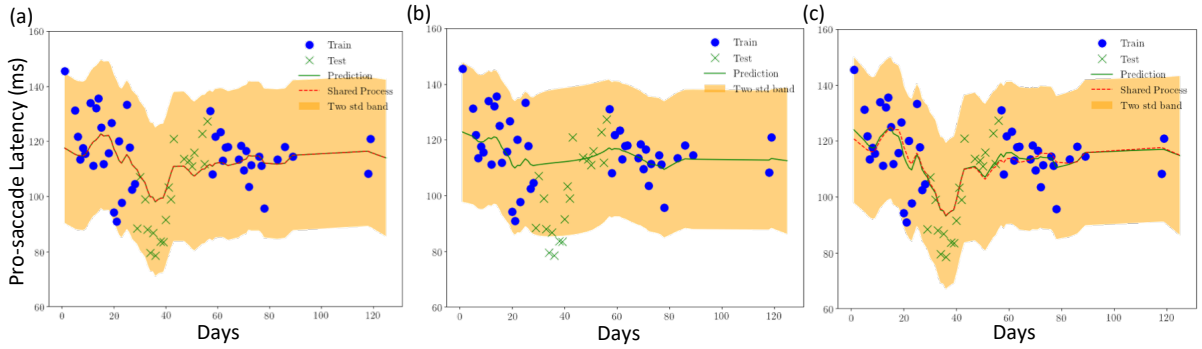

Fig. 12. The performance of the three GP models on Participant 4 in Fig. 8 with missing pro-saccade latency values – (a) the multi-task model, (b) the feature-specific model, and (c) the mixed model. The training data, the testing data, the predictions, the learned shared processes, and the two-standard-deviation bounds are shown. In the multi-task model, the prediction is the same as the learned shared process.

5.4.3 Linear Trend. In the mixed GP model, we assume the mean functions of $h_i(t)$ to be zero. In this section, we test whether the model performs better if we instead assume the mean functions to be linear. That is, whether there is a significant linear trend in the data (which can model learning effects). To do so, we modify the mixed model as follows:

$$p(y_p|g_p) = \prod_{i=1}^{4} \prod_{n=1}^{N} N(y_{pin}; w_{pi} g_p(t_n) + h_{pi}(t_n), \sigma_i^2), \tag{9}$$

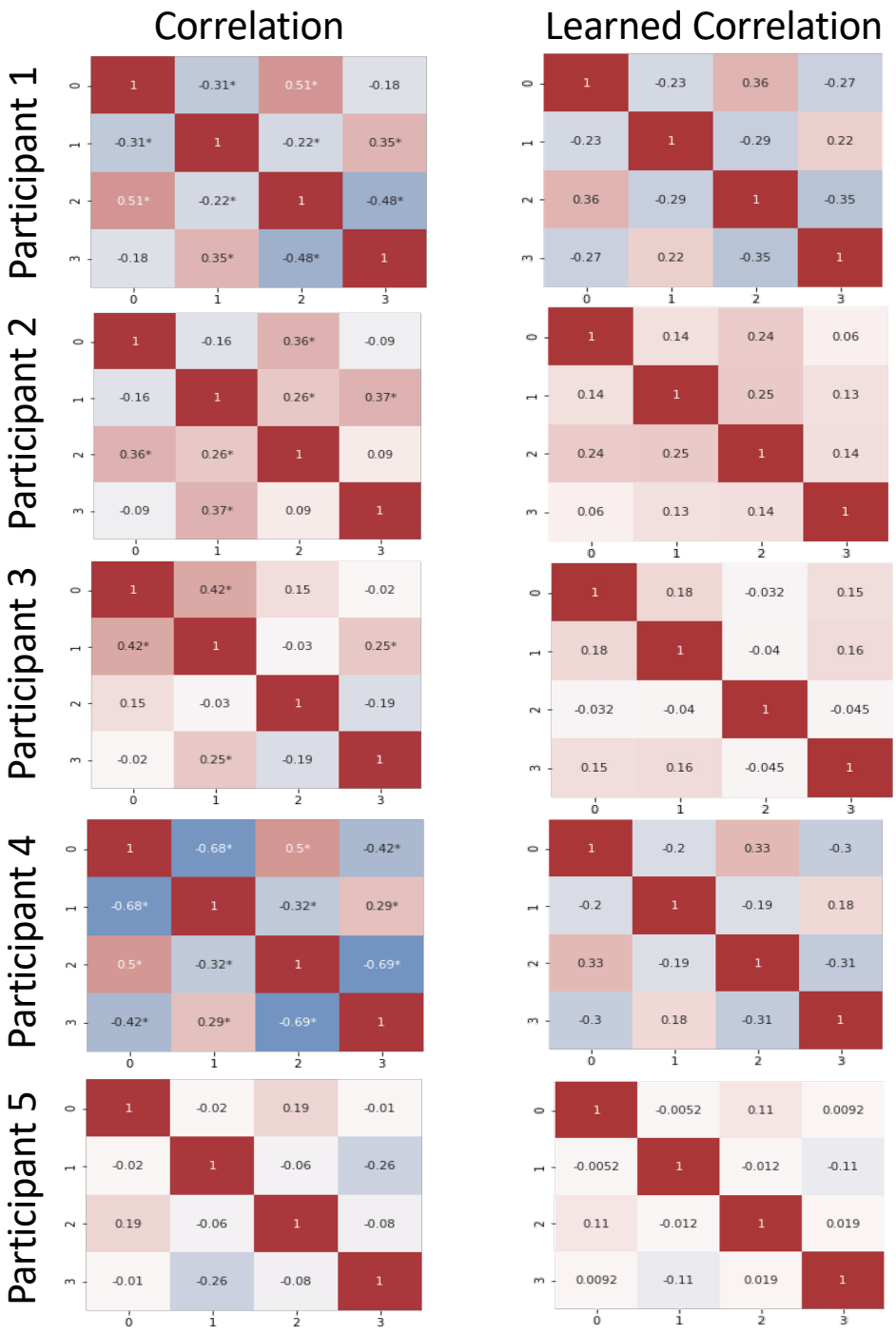

Fig. 13. The correlation estimated from the data versus the correlation learned by the mixed model.

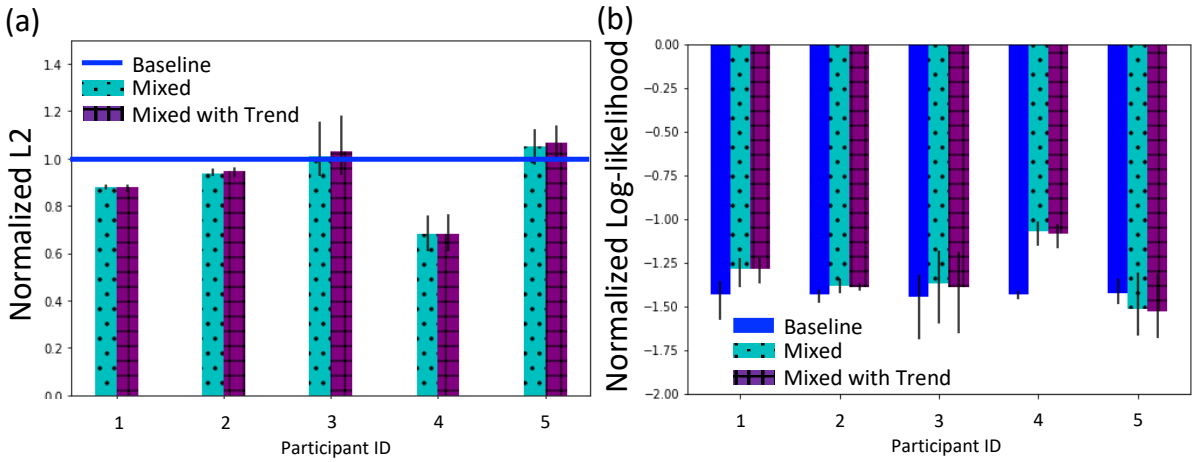

Fig. 14. The performance of the baseline, the mixed model, and the mixed model with a linear trend on participants with more than 45 days of data regarding (a) normalized L2 and (b) normalized log-likelihood. Participants are ordered by their number of recordings in decreasing order.

where

$$
\begin{aligned}
g_p &\sim GP(0, K^g(t,t')), \\
h_{pi} &\sim GP(\Phi_{ind}^{(i)}(t)^T b_{pi}, K_{pi}^h(t,t'), \\
K^g(t,t') &= \exp\{-\frac{|t-t'|}{l^g}\}, \\
K_{pi}^h(t,t') &= a_i^2(1-\tilde{w}_{pi}^2)\exp\{-\frac{|t-t'|}{l_i^h}\}, \\
w_{pi} &= a_i\tilde{w}_{pi}, \tilde{w}_{pi} \in (-1,1),
\end{aligned}
\tag{10}
$$

and

$$
b_{pi} \sim N(0, B^{(i)}). \tag{11}
$$

Here, $\Phi_{ind}(i)(t)$ are the two bases (the slope and the intersection) for the linear functions and b_{pi} are the corresponding coefficients. We assume that not all participants may present a linear trend. As a result, these coefficients are learned for each individual p and are drawn from a population distribution $N(0, B^{(i)})$. We simplify the setup by assuming that $B^{(i)}$ is diagonal. Therefore, there are two hyperparameters to learn for each i. In total, this model has eight more hyperparameters to learn than the mixed GP model.

To test whether a model with a linear trend helps, we test a 3-fold cross validation on randomly missing days of recordings from participants with more than 45 days of data. As shown in Fig. 14, we observe that the performances with and without a linear trend are almost identical. It's likely that the performance with a linear trend model is slightly worse the performance without a linear trend due to overfitting.

6 DISCUSSION

The motivation of this work is to evaluate whether eye-movement features can be used to track the progression of neurocognitive states. Currently, there are few studies that track the longitudinal changes in saccade latency among patients [2, 42], especially within the same cohort. Because the data in these studies were collected in

clinical environments and the analyses usually involved manual removal of outliers, longitudinal measurements are sparse (typically with an interval greater than six months). Therefore, current methods cannot assess disease progression sufficiently frequently to detect disease onset or efficiently evaluate treatment effects.

With the system and the methods discussed in Section 3, we are able to collect significantly more saccades and more sessions per participant than previously possible – 6787 recordings and 235520 eye movements from 80 participants, 45 of whom with multiple recording sessions. These sizable data allow us to study the day-to-day variations in the eye-movement features and the correlation across the eye-movement features. Here, we show a preliminary analysis of the characteristics of the eye-movement features from healthy participants. By doing so, we can put into better context the changes seen in patients with a neurodegenerative disease and potentially use these features to track the disease progression.

6.1 Day-to-day Variations

Fig. 7 shows that there is significant inter-subject variability in the day-to-day variations, which highlights the importance of individualized tracking of eye-movement features. We further analyze the variations within a day using bootstrapping and show that the variations within a day is smaller than the variations across the days. This observation suggests that the source of the day-to-day variations cannot be solely explained by random measurement noise.

One source of day-to-day variations is the change of a participant's task-performing strategy. We observe that participants may be testing different strategies throughout the course of the recordings. As shown in Fig. 8, Participant 4 seems to trade off between speed and accuracy. Therefore, when the latency values decrease, the error rates rise, and vice versa. This task-performing strategy introduces the correlation over time and the correlation across eye-movement features. As shown in Fig. 9, Participant 1 and Participant 4 present significant correlations across the eye-movement features whereas the correlations across the eye-movement features in Participant 5 are insignificant. This observation suggests that not all participants have similar strategies. Therefore, when we design an individualized longitudinal model, we need to model individualized correlations over time and across features.

6.2 Longitudinal Model

With a better understanding of how eye-movement features change over time in healthy individuals, we can design individualized longitudinal models that can characterize the features in the hope that the models can be extended for monitoring disease progression. GP models have been commonly used in disease progression modeling [8, 13, 45]. In particular, we evaluate the performances of three GP models. While all these models are special cases of a multi-task GP model, the mixed model particularly is designed based on the intuition we learn about the individual task-performing strategies. We ensure that the mixed model can model the impact of the strategy flexibly.

We compare the three GP models with a baseline model where we assume that the features in healthy individuals are fixed and that the day-to-day variations are caused by random noise. We notice that when we have collected more than 25 days of recordings, all three models outperform the baseline. It suggests that the eye-movement features are correlated over time and that we can characterize the eye-movement features better than assuming that they are fixed over time in healthy individuals. In addition, we evaluate the capabilities of the three GP models in characterizing the correlation across the eye-movement features. We see that the mixed model performs the best when the correlations across the eye-movement features are significant. Moreover, when we further inspect the linear correlations learned by the mixed model, we notice that the signs of the linear correlations can be learned correctly when they are significant. This result suggests that the mixed model learn individual task-performing strategies.

Last but not least, we test whether the performance can be improved by adding a linear trend in the mixed model. We notice that the performance hardly changes after we assume a linear trend. We hypothesize that it is because 1) the learning effect only lasts for a short period of time and may not be noticeable after 25 days of recordings, and 2) the eye-movement features were not affected by disease progression.

6.3 Limitations and Future Directions

While we have taken 100x more recordings than most previous literature and analyzed longitudinal characteristics in pro/anti-saccade latency and error rate, there are several limitations in our work. To begin with, the system currently only measures saccade latency and error rate. Eye-movement features such as gaze amplitude and velocity [1] may also be affected by disease progression. They are currently not measured because it is challenging to measure saccade amplitude accurately using mobile devices, although with a recent work [49], these amplitude-related features may be included. In addition, we need a diverse group of participants with more than 25 days of data to fully understand the strengths and weaknesses of our longitudinal models. (Notice that with current measurements of neurodegenerative diseases being relatively obtrusive, we may need to rely on the "self-reported healthy" criterion to recruit participants before we start to take measurements from clinics where we may be able to collect data from clinically-reported healthy individuals. While this may be another limitation, we are hoping that by analyzing data from more individuals, the effect of a potential missed diagnosis can be minimized.)

Nevertheless, we hope that our analyses provide some insights about future research directions. For example, the distributions of day-to-day variations in neurodegenerative-disease patients have not been reported in the literature before. It will be interesting to see if the day-to-day variations measured from people with neurodegenerative diseases are generally larger than normal elderly participants. We can ask similar questions about the correlations among the eye-movement features. For example, will the correlations between the daily median pro- and anti-saccade latency be different between healthy participants and patients? (Notice that different from the literature where the correlation is often calculated over the population, here we get correlation estimates for each individual.)

Moreover, while the mixed model can be a good candidate model for healthy individuals, it can be improved to characterize more participants and characterize data collected over a much longer period. For example, we may include the effect of age or extend our model to a disease progression model using the ideas in [13, 45], or we may reduce the computation complexity using the stochastic variational inference for GP [17, 36]. Notice that unlike the work in [45] where the ground-truth measurement of the disease may be relatively clear (e.g., PFVC for scleroderma), our ground-truth measurement is not clear. One way to evaluate how eye-movement features are correlated with different measurements of disease states (such as cognitive scores and cerebrospinal fluid (CSF) measurements) is to include current clinical measurements as output features in a GP model. It will then be interesting to know how to design this multi-output GP when the linear correlation across the eye-movement features is taken into account. So far, all these future directions focus on "characterizing" eye movement features. Only after we understand these eye-movement features better can we start to consider a better definition of neurocognitive states, include it in our model, and "predict" the neurocognitive state of an individual.

Last but not least, several improvements are required on the measurement system before we can analyze data from patients. We need to re-adjust the app and task design to ensure that the measurement is user-friendly for patients. Additionally, we need to keep the app engaging. As mentioned in [48], the disease stage affects an individual's willingness to participate in a study. We may need to think about how to motivate patients to take recordings on a regular basis without making it burdensome. To achieve these goals, we should interact with patients and iterate the app design.

7 CONCLUSION

A recent work has enabled frequent measurements of pro/anti-saccade latency and directional error rate using an app and measurement pipeline. In this work, we studied longitudinal characteristics of pro/anti-saccade latencies/error rates from healthy participants collected using the framework, which was barely studied in the literature due to the constrained environment setup. We noticed substantial inter-subject variability in day-to-day variations and recognized potential task-performing strategies in multiple participants. From there, we proposed a flexible GP model that can characterize individuals with more than 25 days of recordings and learn the correlation across the features introduced by the task-performing strategies. Together with the app and measurement pipeline, we demonstrated the potential to track eye-movement features from healthy individuals and open up the possibility to use eye-movement features to track neurocognitive states more frequently and widely than previously possible.

ACKNOWLEDGMENTS

This work was supported in part by the MIT-IBM Watson AI Lab, the MIT's Aging Brain Initiative, and by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] T.J. Anderson and M.R. MacAskill. 2013. Eye movements in patients with neurodegenerative disorders. *Nature Reviews Neurology* 9, 2 (2013), 74–85.

[2] C.A. Antoniades, Z. Xu, S.L. Mason, R.H.S. Carpenter, and R.A. Barker. 2010. Huntington's disease: Changes in saccades and hand-tapping over 3 years. *Journal of Neurology* 257, 11 (2010), 1890–1898.

[3] G. Bargary, J. M. Bosten, P. T. Goodbourn, A. J. Lawrance-Owen, R. E. Hogg, and J.D. Mollon. 2017. Individual differences in human eye movements: An oculomotor signature? *Vision Research* 141 (2017), 157–169. https://doi.org/10.1016/j.visres.2017.03.001

[4] L. Barrios, P. Oldrati, M. Hilty, D. Lindlbauer, C. Holz, and A. Lutterotti. 2021. Smartphone-Based Tapping Frequency as a Surrogate for Perceived Fatigue: An in-the-Wild Feasibility Study in Multiple Sclerosis Patients. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 89 (2021), 30 pages. https://doi.org/10.1145/3478098

[5] E.V. Bonilla, K. Chai, and C. Williams. 2008. Multi-task Gaussian Process Prediction. In *Advances in Neural Information Processing Systems*, Vol. 20.

[6] B.M. Bot, C. Suver, E.C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E.R. Dorsey, S.H. Friend, and A.D. Trister. 2016. The mPower study, Parkinson's disease mobile data collected using ResearchKit. *Scientific data* 3, 3 (2016), 160011.

[7] C.E. Rasmussen and C.K.I Williams. 2005. *Gaussian Processes for Machine Learning*. MIT Press.

[8] L.-F. Cheng, G. Darnell, B. Dumitrascu, C. Chivers, M.E. Draugelis, K. Li, and B.E. Engelhardt. 2018. Sparse Multi-Output Gaussian Processes for Medical Time Series Prediction. arXiv:1703.09112 [stat.ML]

[9] T.J. Crawford, S. Higham, T. Renvoize, J. Patel, M. Dale, A. Suriya, and S. Tetley. 2005. Inhibitory control of saccadic eye movements and cognitive impairment in Alzheimer's disease. *Biological Psychiatry* 57, 9 (2005), 1052–1060.

[10] A.L. Silva de Lima, L. Evers, T. Hahn, L. Bataille, J.L. Hamilton, M.A. Little, Y. Okuma, B.R. Bloem, and M.J. Faber. 2017. Freezing of gait and fall detection in Parkinson's disease using wearable sensors: a systematic review. Journal of neurology. *Journal of neurology* 264, 8 (2017), 1642–1654.

[11] E.R. Dorsey, A.M. Glidden, M.R. Holloway, G.L. Birbeck, and L.H. Schwamm. 2018. Teleneurology and mobile technologies: The future of neurological care. *Nature Reviews Neurology* 14, 5 (2018), 285–297.

[12] E.R. Dorsey, S. Papapetropoulos, M. Xiong, and K. Kieburtz. 2017. The First Frontier: Digital Biomarkers for Neurodegenerative Diseases. *Digital Biomarkers* 1 (2017), 6–13.

[13] J. Futoma, M. Sendak, B. Cameron, and K. Heller. 2016. Predicting Disease Progression with a Model for Multivariate Longitudinal Clinical Data. In *Machine Learning for Healthcare*. PMLR, 42–54.

[14] S. Garbutt, A. Matlin, J. Hellmuth, A.K. Schenk, J.K. Johnson, H. Rosen, D. Dean, J. Kramer, J. Neuhaus, B.L. Miller, S.G. Lisberger, and A.L. Boxer. 2008. Oculomotor function in frontotemporal lobar degeneration, related disorders and Alzheimer's disease. *Brain* 131, 5 (2008), 1268–1281.

[15] H. Wackernagel. 1998. *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, Berlin, 2nd edition.

[16] P.D. Harvey, S. Cosentino, R. Curiel, T.E. Goldberg, J. Kaye, D. Lowenstein, D. Marson, D. Salmon, K. Wesnes, and H. Posner. 2017. Performance-based and observational assessments in clinical trials across the Alzheimer's disease spectrum. *Innovations in Clinical Neuroscience* 14, 1-2 (2017), 30–39.

[17] J. Hensman, N. Fusi, and N.D. Lawrence. 2013. Gaussian Processes for Big Data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* (Bellevue, WA) *(UAI'13)*. AUAI Press, Arlington, Virginia, USA, 282–290.

[18] 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.

[19] J.E. McDowell JE, K.A. Dyckman, B.P. Austin, and B.A. Clementz. 2008. Neurophysiology and neuroanatomy of reflexive and volitional saccades: evidence from studies of humans. *Brain and Cognition* 68, 3 (2008), 255–270.

[20] L.C. Kourtis, O.B. Regele, J.M. Wright, and G.B. Jones. 2019. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *npj Digital Medicine* 2, 9 (2019).

[21] H.-Y. Lai, G. Saavedra-Peña, C.G. Sodini, T. Heldt, and V. Sze. 2018. Enabling saccade latency measurements with consumer-grade cameras. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 3169–3173.

[22] H.-Y. Lai, G. Saavedra-Peña, C. G. Sodini, T. Heldt, and V. Sze. 2022. App-based saccade latency and directional error determination across the adult age spectrum. *IEEE Transactions on Biomedical Engineering* 69, 2 (2022), 1029–1039. https://doi.org/10.1109/TBME.2021.3112007

[23] H.-Y. Lai, G. Saavedra-Peña, C. G. Sodini, V. Sze, and T. Heldt. 2020. Measuring Saccade Latency Using Smartphone Cameras. *IEEE Journal of Biomedical and Health Informatics* 24, 3 (2020), 885–897. https://doi.org/10.1109/JBHI.2019.2913846

[24] R. Langevin, M.R. Ali, T. Sen, C. Snyder, T. Myers, E.R. Dorsey, and M.E. Hoque. 2019. The PARK Framework for Automated Analysis of Parkinson's Disease Characteristics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 54 (2019), 22 pages. https://doi.org/10.1145/3328925

[25] J.C. Latourelle, M.T. Beste, T.C. Hadzi, R.E. Miller, J.N. Oppenheim, M.P. Valko, D.M. Wuest, B.W. Church, I.G. Khalil, B. Hayete, and C.S. Venuto. 2017. Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *The Lancet Neurology* 16, 11 (2017), 908–916. https://doi.org/10.1109/10.1016/S1474-4422(17)30328-9

[26] R.J. Leigh and D.S. Zee. 2015. The Saccadic System. In *The Neurology of Eye Movements*. Oxford University Press, Oxford, Chapter 4, 169–288.

[27] M. Little, P. Wicks, T. Vaughan, and A. Pentland. 2013. Quantifying Short-Term Dynamics of Parkinson's Disease Using Self-Reported Symptom Data From an Internet Social Network. *Journal of Medical Internet Research* 15, 1 (2013), e20.

[28] R.Z. Marandi and R. Gazerani. 2019. Aging and eye tracking: in the quest for objective biomarkers. *Future Neurology* 14, 4 (2019), FNL33. https://doi.org/10.2217/fnl-2019-0012

[29] R. Z. Marandi, R. Madeleine, Ø. Omland, N. Vuillerme, and A. Samani. 2018. Eye movement characteristics reflected fatigue development in both young and elderly individuals. *Scientific Reports* 8, 13148 (2018). https://doi.org/10.1038/s41598-018-31577-1

[30] R. Z. Marandi, R. Madeleine, Ø. Omland, N. Vuillerme, and A. Samani. 2018. Reliability of Oculometrics During a Mentally Demanding Task in Young and Old Adults. *IEEE Access* 6 (2018), 17500–17517. https://doi.org/10.1109/ACCESS.2018.2819211

[31] A.J. Mitchell. 2009. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research* 43, 4 (2009), 411–431.

[32] U.P. Mosimann, R.M. Müri, D.J. Burn, J. Felblinger, J.T. O'Brien, and I.G. McKeith. 2005. Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies. *Brain* 128, 6 (2005), 1267–1276.

[33] D.P. Munoz and S. Everling. 2004. Look away: the anti-saccade task and the voluntary control of eye movement. *Nature Review Neuroscience* 5, 3 (2004), 218–228. https://doi.org/10.1038/nrn1345

[34] National Academies of Sciences, Engineering, and Medicine. 2018. *Harnessing mobile devices for nervous system disorders: Proceedings of a Workshop*. The National Academies Press, Washington, DC. https://doi.org/10.17266/25274

[35] J. Neville, S. Kopko, S. Broadbent, E. Avilés, R. Stafford, C.M. Solinsky, L.J. Bain, M. Cisneroz, K. Romero, and D. Stephenson. 2015. Development of a unified clinical trial database for Alzheimer's disease. *Alzheimer's and dementia : the journal of the Alzheimer's Association* 11, 10 (2015), 1212–1221.

[36] V.T. Nguyen and E.V. Bonilla. 2014. Collaborative Multi-Output Gaussian Processes. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI'14)*. 643–652.

[37] J.M. Owens, T.A. Dingus, F. Guo, Y. Fang, M. Perez, J. McClafferty, and B.C. Tefft. 2018. Prevalence of Drowsy Driving Crashes: Estimates from a Large-Scale Naturalistic Driving Study (Research Brief). *AAA Foundation for Traffic Safety* (2018).

[38] N.P. Oxtoby, A.L. Young, D.M. Cash, T.L.S. Benzinger, A.M. Fagan, J.C. Morris, R.J. Bateman, N.C. Fox, J.M. Schott, and D.C. Alexander. 2018. Data-Driven Models of Dominantly-Inherited Alzheimer's Disease Progression. *Brain* 141, 5 (2018), 1529–1544.

[39] J. Paulsen, J. Long, H. Johnson, E. Aylward, C. Ross, J. Williams, M. Nance, C. Erwin, H. Westervelt, D. Harrington, H. Bockholt, Y. Zhang, E. McCusker, E. Chiu, and P. Panegyres. 2014. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Frontiers in Aging Neuroscience* 6 (2014), 78.

[40] A. Piau, K. Wild, N. Mattek, and J. Kaye. 2019. Current State of Digital Biomarker Technologies for Real-Life, Home-Based Monitoring of Cognitive Function for Mild Cognitive Impairment to Mild Alzheimer Disease and Implications for Clinical Care: Systematic Review. *Journal of Medical Internet Research* 21, 8 (2019), e12785.

[41] H. Posner, R. Curiel, C. Edgar, S. Hendrix, E. Liu, D.A. Loewenstein, L. Morrison, G. Shinobu, K. Wesnes, and P.D. Harvey. 2017. Outcomes assessment in clinical trials of Alzheimer's disease and its precursors: Readying for short-term and long-term clinical trial needs. *Innovations in Clinical Neuroscience* 14, 1-2 (2017), 22–29.

[42] M. Proudfoot, R.A. Menke, R. Sharma, C.M. Berna, S.L. Hicks, C. Kennard, K. Talbot, and M.R. Turner. 2015. Eye-tracking in amyotrophic lateral sclerosis: A longitudinal study of saccadic and cognitive tasks. *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 17, 1-2 (2015), 101–111.

[43] S. Rivaud-Péchoux, M. Vidailhet, J. Brandel, and B. Gaymard. 2006. Mixing pro- and antisaccades in patients with Parkinsonian syndromes. *Brain* 130, 1 (2006), 256–264.

[44] O. Rudovic, Y. Utsumi, R. Guerrero, K. Peterson, D. Rueckert, and R.W. Picard. 2019. Meta-Weighted Gaussian Process Experts for Personalized Forecasting of AD Cognitive Changes. arXiv:1904.09370 [cs.LG]

[45] P. Schulam and S. Saria. 2015. A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure. In *Neural Information Processing Systems (NIPS)*. 748–756.

[46] R. Shafiq-Antonacci, P. Maruff, C. Masters, and J. Currie. 2003. Spectrum of Saccade System Function in Alzheimer's disease. *Archives of Neurology* 60, 9 (2003), 1275–1278.

[47] S.J. Tabrizi, D.R. Langbehn, B.R. Leavitt, R.A.C. Roos, A. Durr, D. Craufurd, C. Kennard, S.L. Hicks, N.C. Fox, R.I. Scahill, B. Borowsky, A.J. Tobin, H.D. Rosas, H. Johnson, R. Reilmann, B. Landwehrmeyer, and J.C. Stout. 2009. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: Cross-sectional analysis of baseline data. *The Lancet Neurology* 8, 9 (2009), 791–801.

[48] S.J. Tabrizi, R.I. Scahill, G. Owen, A. Durr, B.R. Leavitt, R.A. Roos, B. Borowsky, B. Landwehrmeyer, C. Frost, H. Johnson, D. Craufurd, R. Reilmann, J.C. Stout, and D.R. Langbehn. 2013. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *The Lancet Neurology* 12, 7 (2013), 637–649.

[49] N. Valliappan, N. Dai, E. Steinberg, J. He, K. Rogers, V. Ramachandran, P. Xu, M. Shojaeizadeh, L. Guo, K. Kohlhoff, and V. Navalpakkam. 2020. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications* 11, 1 (11 Sep 2020), 4553. https://doi.org/10.1038/s41467-020-18360-5

[50] D.P. Veitch, M.W. Weiner, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack, W. Jagust, J.C. Morris, R.C. Petersen, A.J. Saykin, L.M. Shaw, A.W. Toga, and J.Q. Trojanowski. 2019. Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's and Dementia* 15, 1 (2019), 106–152.

[51] T. Wang, R.G. Qiu, and M. Yu. 2018. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Scientific Reports* 8, 9161 (2018).

[52] X. Wang, D. Sontag, and F. Wang. 2014. Unsupervised Learning of Disease Progression Models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. Association for Computing Machinery, 85–94. https://doi.org/10.1145/2623330.2623754

[53] J.P. Weir. 2005. Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM. *Journal of Strength and Conditioning Research* 19, 1 (2005), 231–240.

[54] H. Zhang, C. Xu, H. Li, A.S. Rathore, C. Song, Z. Yan, D. Li, F. Lin, K. Wang, and W. Xu. 2019. PDMove: Towards Passive Medication Adherence Monitoring of Parkinson's Disease Using Smartphone-Based Gait Assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 123 (sep 2019), 23 pages. https://doi.org/10.1145/3351281