Classification and staging of Parkinson’s disease using video-based eye tracking

Donald C. Brien a,*, Heidi C. Riek a, Rachel Yep a, Jeff Huang a, Brian Coe a, Corson Areshenkoff a, David Grimes b, Mandar Jog d, Anthony Lang c, Connie Marras f, Mario Masellis g, Paula McLaughlin a,h,i, Alicia Peltsch j, Angela Roberts h, Brian Tan l, Derek Beaton m, Wendy Lou n, Richard Swartz o, the ONDRI Investigators p, Douglas P. Munoz a,b

a Centre for Neuroscience Studies, Queen’s University, Kingston, ON, Canada
b Department of Biomedical and Molecular Sciences, Queen’s University, Kingston, ON, Canada
c Ottawa Hospital, uOttawa Brain and Mind Research Institute, Ottawa, ON, Canada
d FF Centre of Excellence, London Movement Disorders Centre, Lawson Health Research Institute, London, ON, Canada
e Edmond J. Safra Program in Parkinson’s Disease, University Health Network and the Department of Medicine, Division of Neurology, University of Toronto, Toronto, ON, Canada
f Toronto Western Hospital Movement Disorders Centre and the Edmond J Safra Program in Parkinson’s Disease, University of Toronto, Toronto, ON, Canada
g Sunnybrook Health Sciences Centre, Medicine (Neurology), Toronto, ON, Canada
h Nova Scotia Health, Halifax, NS, Canada
i Department of Psychology and Neuroscience, Department of Medicine (Geriatrics), DalHousie University, Halifax, NS, Canada
j Faculty of Engineering and Applied Science, Queen’s University, Kingston, ON, Canada
k School of Communication Sciences and Disorders, Department of Computer Science, University of Western Ontario, London, ON, Canada
l Roman Research Institute, Baycrest Health Sciences, Toronto, ON, Canada
m Data Science & Advanced Analytics (DSAA), St. Michael’s Hospital, Unity Health Toronto, Toronto, ON, Canada
n Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
p Hurvitz Brain Sciences Program, Sunnybrook Health Sciences Centre; Faculty of Medicine, Department of Medicine (Neurology), University of Toronto, Toronto, ON, Canada

ABSTRACT

Introduction: 83% of those diagnosed with Parkinson’s Disease (PD) eventually progress to PD with mild cognitive impairment (PD-MCI) followed by dementia (PDD) - suggesting a complex spectrum of pathology concomitant with aging. Biomarkers sensitive and specific to this spectrum are required if useful diagnostics are to be developed that may supplement current clinical testing procedures. We used video-based eye tracking and machine learning to develop a simple, non-invasive test sensitive to PD and the stages of cognitive dysfunction.

Methods: From 121 PD (45 Cognitively Normal/45 MCI/20 Dementia/11 Other) and 106 healthy controls, we collected video-based eye tracking data on an interleaved pro/anti-saccade task. Features of saccade, pupil, and blink behavior were used to train a classifier to predict confidence scores for PD/PD-MCI/PDD diagnosis.

Results: The Receiver Operator Characteristic Area Under the Curve (ROC-AUC) of the classifier was 0.88, with the cognitive-dysfunction subgroups showing progressively increased AUC, and the AUC of PDD being 0.95. The classifier reached a sensitivity of 83% and a specificity of 78%. The confidence scores predicted PD motor and cognitive performance scores.

Conclusion: Biomarkers of saccade, pupil, and blink were extracted from video-based eye tracking to create a classifier with high sensitivity to the landscape of PD cognitive and motor dysfunction. A complex landscape of

* Corresponding author. Centre for Neuroscience Studies Rm 238 Botterell Hall. Queen’s University Kingston, Ontario, K7L 3N6, Canada.
E-mail addresses: briend@queensu.ca (D.C. Brien), heidi.riek@queensu.ca (H.C. Riek), 12ry@queensu.ca (R. Yep), 1jh19@queensu.ca (J. Huang), coe@queensu.ca (B. Coe), c.areshenkoff@queensu.ca (C. Areshenkoff), dagrimes@toh.ca (D. Grimes), Mandar.Jog@lhsc.on.ca (M. Jog), anthony.lang@uhnresearch.ca (A. Lang), connie.marras@uhnresearch.ca (C. Marras), Mario.Masellis@sunnybrook.ca (M. Masellis), paula.mclaughlin@inshealth.ca (P. McLaughlin), alicia.peltsch@queensu.ca (A. Peltsch), angela.roberts@northwestern.edu (A. Roberts), bstan@research.baycrest.org (B. Tan), derek.beaton@unityhealth.to (D. Beaton), wendy.lou@utoronto.ca (W. Lou), rick.swartz@sunnybrook.ca (R. Swartz), doug.munoz@queensu.ca (D.P. Munoz).

See Supplementary Material – ONDRI Investigators List.

https://doi.org/10.1016/j.parkreldis.2023.105316
Received 23 September 2022; Received in revised form 11 January 2023; Accepted 4 February 2023
Available online 8 February 2023
1353-8020/© 2023 Elsevier Ltd. All rights reserved.
1. Introduction

Parkinson’s Disease (PD) is a progressive and variable neurodegenerative disorder lacking a reliable biomarker. PD is currently diagnosed through the clinical evaluation of symptoms - tremor, rigidity, akinesia, and postural instability [1]. Furthermore, 83% of those diagnosed with PD will progress to PD with dementia (PDD), suggesting a complex mixture of pathology concomitant with aging [2]. Clinically evident neurodegeneration is preceded by prodromal pathology, years before the onset of first motor symptoms [3,4], suggesting an even more nuanced chronology. PD through PDD is clearly a complex, heterogeneous spectrum of neurodegeneration. Mitigation strategies start with early, specific, and sensitive diagnostics that can help proactively develop treatment regimens and predict outcomes. This work aimed to demonstrate the validity of eye-tracking tests as one possible diagnostic meeting these criteria of specificity and sensitivity over the early spectrum of PD.

The neural circuitry controlling eye movements is one of the most studied circuits in the brain, with pathways encompassing regions from frontal cortex to medulla [5]. The span of this network makes it an excellent non-invasive, low-cost model for studying neurobiology. The outputs of this system - saccade, pupil, and blink recordings - provide a high-fidelity proxy signal of the underlying intactness of brain circuits. Neurodegeneration disrupts a subset of these circuits, resulting in specific alterations of the output [6]. With PD, degeneration is characterized by either a body-first caudal to rostral pathological progression – possibly beginning in the enteric nervous system, advancing to medulla oblongata and then the substantia nigra, and eventually involving the entire neocortex [3] – or a brain-first progression that eventually descends to the peripheral autonomic nervous system [7]. An extensive history of resulting eye movement alterations has been reported in PD; in particular, increased saccade latency in both volitional and visually-guided saccades, increased errors in antisaccade tasks, hyper-reflexivity to known targets, saccade hypometria, altered pupil responses, and decreased blink rates [6,8,9]. The magnitude and progression of these alterations track with PD to PDD progression, with saccade amplitude abnormalities detectable earlier in the disease course than latency abnormalities [10-12].

While the body of research points to alterations of eye movements in PD patients, existing studies have concentrated on relatively small numbers of participants, at varying narrow stages of disease, and with varying tasks that can alter performance and confound the overall picture of PD eye movement dysfunction [13]. Recent work investigating the predictive power of eye-tracking measures has focused more and more on large-scale studies capturing the spectrum of PD from prodromal through PDD. Compared to PD patients, PDD patients are more impaired in measures of saccadic gain, latency, and execution and these saccade measures correlate with disease staging and the progression of motor and cognitive deficits [10,12,14]. Other recent work [9,15] has shown that pupil and blink are altered in prodromal PD where other saccade behaviors are not. Taken together, there is task-dependent variability in eye-movement markers, and more complex tasks are more sensitive to disease progression because they more fully sample the eye-movement network. The full range of eye-movement features are needed for best sensitivity to all stages of the disease and an aspect of voluntary control is key as the disease involves the frontostriatal pathways. This is confirmed in a recent meta-analysis [16], which showed that antisaccade reaction times and error rates were significantly increased in PD across a wide range of studies. They also confirmed that disease duration, staging, and motor scores were correlated to antisaccade latency. Thus, our focus is on using an interleaved pro/antisaccade gap task [5] to produce a rich set of features that fully capture the heterogeneity of the PD.

Many participants are required to cover the spectrum of PD and afford generalizable, longitudinally reproducible results. Through the Ontario Neurodegenerative Disease Research Initiative (ONDRI), such a sample was available to us [17,18]. The goal of ONDRI was to longitudinally study concomitant neurodegenerative pathologies across their spectrums to advance understanding of their common and divergent biotypes. Here we focus on the PD cohort to provide a large scale, longitudinally study of PD across the cognitive spectrum (normal cognition to dementia). We acquire the pro/antisaccade gap task on PD patients (as well as matching controls) to provide a unified and controlled view of the spectrum of PD. Advanced feature extraction methods and machine learning techniques are used to build a classification model that captures cognitive and motor alterations through these subgroups of PD. We predict a classifier that is both sensitive and specific to PD and PD progression. We hypothesize that oculomotor features are promising biomarkers that can be used to supplement the diagnosis of PD in the clinic.

2. Methods

2.1. Participants

One-hundred and forty PD participants were recruited through ONDRI, involving health care centers throughout Ontario, Canada [17,18], of which 121 were used in this study. Approval for experimental procedures was obtained from the respective Research Ethics Boards of these health centers. Written consent was obtained from each PD participant. At baseline, PD participants were 55–85 years of age, diagnosed with idiopathic PD based on the United Kingdom Parkinson’s Disease Society Brain Bank criteria [19] within the last 3–8 years and had a Hoehn & Yahr score of 1–3 [20]. PD participants had normal or corrected-to-normal vision in at least one eye. PD participants were on medication as required. PD participants were followed 3 years longitudinally on a series of assessment platforms; however, due to high attrition rates in the last year we only considered baseline, 1 year, and 2 years follow up timepoints.

One-hundred and six healthy control participants (see our previous work [21]) were recruited locally from the Kingston, Ontario, Canada community and experimental procedures were reviewed and approved by the Queen’s University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board. Control participants reported no history of neurodegeneration or other neurological dysfunction and had normal or corrected-to-normal vision. 96 control participants completed the Montreal Cognitive Assessment (MoCA) to screen for cognitive decline [22]. Written informed consent was obtained from every participant. There was no longitudinal data for controls.

PD participants and healthy controls were matched for age and education level, but not for sex. See Table 1 (and Supplementary methods) for a complete description of participants.

2.2. PD clinical and neuropsychological evaluation

Each PD participant was assessed by neurological examination including the MoCA, the Movement Disorder Society-Unified Parkinson’s Disease Rating Score (MDS-UPDRS) (Total Score of all 4 sections and Part III – Motor Examination), the Scales for Outcomes in Parkinson’s disease – Autonomic Dysfunction (SCOPA-AUT) [23], and an extensive neuropsychology battery (NPSY) [24]. As outlined in previous work on the ONDRI PD cohort, composite scores were calculated across
five cognitive domains using the NPSY protocol: Attention and Working Memory, Executive Function, Language, Memory, and Visuospatial Function [24–26]. This was done using the baseline visit, as well as the follow-up visits (year 1 and year 2). Composite scores are relative to the group in each time point and z-scored such that positive values indicate better performance than the group mean in each domain.

Performance on the NPSY battery at baseline was also used to classify participants as being cognitively normal (PD-CN), or meeting MDS Level 2 criteria [27] for PD-MCI or PDD. See Supplementary Methods for a description of these criteria.

### 2.3. Task and eye-movement recordings

The interleaved pro/anti saccade task (IPAST) (Fig. 1A) (See Coe, Huang, Brien, White, Yep, and Munoz (2022) [28] for details) with the gap condition was chosen to evaluate all participants. The randomly interleaved trials of this task required a participant to fixate a central gap and then, based on the color instruction of the central point, either override this automatic response and make a voluntary anti-saccade (red central point) to the opposite location of the target. This task allowed us to collect automatic and volitional saccades, pupil responses to the fixation, and blink rates throughout each trial. See Supplementary Methods for details on the task and eye-movement recordings.

### 2.4. Data processing and feature selection

Raw eye tracking data was first pre-processed in MATLAB (2019b, The MathWorks, Inc.) for quality assurance and the identification of saccades, blinks, and fixations. See Coe, Huang, Brien, White, Yep, and Munoz (2022) [28] and Supplementary Methods for detailed descriptions of the pre-processing pipeline.

The key to building a classification model of neurodegeneration is to fully describe the output of the oculomotor network in a set of numerical features. We describe 2 sets of features that we categorize by extraction technique (Supp. Table 3) that total 45 distinct features (Fig. 1B):

1. **Point Estimates (21 features).** These features were estimates of mean values and rates. For example, mean reaction times to the target for pro- and antisaccades, mean amplitudes, error rates, and anticipation (intrusive or premature saccades before the target appears) rates are included and have been shown in previous literature to be predictive of PD.

2. **Functional Estimates (24 features).** Output signals from video-based eye tracking are continuous functions and it is not always obvious how to fully capture their detail into summary features. For example, pupil or blink probabilities are continuous functions with complex structure throughout trials (see Supp. Fig. 1-4 for examples). Experimenter defined features such as slopes, or epochs may miss important, complex details. Instead, we used Functional Data Analysis (FDA) to analyze functional data. FDA allowed us to automatically extract functional summary features. One such automated method is functional principal components analysis (fPCA – a direct analog to PCA analysis), which we used to extract the 4 main components of each function. 4 was chosen subjectively as to capture an adequate amount of the variability. The key take-home from this analysis is that we can objectively extract 4 discrete values for each participant that summarizes the function in question. See Supplementary Methods for a complete description of this technique and the generated features.

### 2.5. Classifier model building

Our concentration was on building a classifier that generalized well (See Supplementary Methods and Supp. Fig. 6). We used a nested 10-fold cross-validation scheme to create repeated test samples from the data, while optimizing and training our classifier on the remaining data (Fig. 1C). We used a voting classifier that consisted of 3 sub-classifiers: a

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Participant characteristics.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Controls</td>
</tr>
<tr>
<td>Number of Participants&lt;sup&gt;a&lt;/sup&gt;</td>
<td>106&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Age (Years)</td>
<td>67.7 ± 8.2 (min 55, max 85)</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>69</td>
</tr>
<tr>
<td>Education Level (Years)</td>
<td>15.7 ± 2.7</td>
</tr>
<tr>
<td>MDS-UPDRS-Totals (4 sections)</td>
<td>46.3 ± 20.42</td>
</tr>
<tr>
<td>MDS-UPDRS-Part III</td>
<td>23.4 ± 12.3</td>
</tr>
<tr>
<td>MoCA</td>
<td>27.3 ± 1.9</td>
</tr>
<tr>
<td>SCOPA AUT-Total</td>
<td>11.4 ± 5.1</td>
</tr>
<tr>
<td>Hoehn &amp; Yahr&lt;sup&gt;d&lt;/sup&gt; (1/1.5/2.5/3)</td>
<td>5/1/89/15/11</td>
</tr>
<tr>
<td>Levodopa Equivalent Daily Dose (mg)</td>
<td>692.2 ± 365.1</td>
</tr>
<tr>
<td>On Cholinesterase Inhibitors</td>
<td>10</td>
</tr>
<tr>
<td>On Anticholinergics</td>
<td>16</td>
</tr>
</tbody>
</table>

All values are reported as mean ± standard deviation.

CN = Cognitively normal, MCI = Mild Cognitive Impairment, as defined by NPSY.

<sup>a</sup> See Data Preprocessing for quality control metrics and missing data. 11 PD participants are in the ‘Other’ category.

<sup>b</sup> Hoehn & Yahr is presented as counts per stages 1 through 3.

<sup>c</sup> p-values are displayed first for tests of Control vs PD ALL (when applicable), followed by tests of PD subtypes only. All tests are Wilcoxon (Control vs PD ALL)/Kruskal-Wallis (PD subtypes), except for categorical variables (Sex/Hoehn & Yahr/Cholinesterase Inhibitors/Anticholinergics) which are Fisher (binary) or cross-tabulation.
support vector machine, a logistic regression, and a random forest. Each of these classifiers produced a continuous score between 0 and 1 that was directly interpretable as Probability of PD. A soft vote (largest sum) of the probability for each sub-classifier was used as the final classification and the average probability was used for the final Probability of PD.

We evaluated the effectiveness of the classifier in several ways. First, we calculated the averaged accuracy of the classifier for each of the 10 held-out test data. These 10 scores provided bounds on the generalized performance of the classifier and were compared statistically to 10 scores from a dummy classifier that predicted classifications randomly. Second, within each fold we predicted the label of PD or control, as well as the Probability of PD, for each individual and used the aggregate labels to evaluate the sensitivity and specificity of the classifier. Finally, we summarized the overall quality of the classifier using Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) score from these. ROC curves summarize the quality of a classifier by plotting the true positive rate (correctly predicting PD for someone with PD) vs the false positive rate (incorrectly predicting PD for a control) at classification thresholds from 0 to 1 for the probability. The AUC of a classifier is 1 for a perfectly correct classifier, 0.5 for a random classifier, and 0 for a perfectly incorrect classifier. This allows for the direct comparison of classifiers, and we used the ROC curves and AUC scores to evaluate the overall efficacy of the classifier and to compare the classifications of subgroups of PD.

We used the individual probabilities to also assess correlations with the clinical measures: MDS-UPDRS-Total Score, MDS-UPDRS Part-III and MoCA.

2.6. Medication and sex effects

PD medications may impact eye-tracking measures. The use of levodopa, cholinesterase inhibitors, and anticholinergics were significantly increased in the PDD cognitive subtype (Table 1). We discuss the possible implications of these medications on our classifier in our Supplementary Methods. In addition, we discuss the implications of the

---

**Fig. 1.** The IPAST and classifier building process. The IPAST task (A) trial consisted of a blank inter-trial interval (ITI), followed by a colored fixation instruction, followed by a blank GAP period, and finally a gray target. Signals of saccade, pupil, and blink were reduced to a feature set (B) encompassing all domains of the oculomotor network. From this feature set, we built a classifier model (C) to predict PD diagnosis and Probability of PD using a nested cross-validation scheme and a voting classifier.
imbalance in sex.

2.7. Longitudinal statistical analysis

Linear mixed-effects models were built to investigate interactions between classifier Probability of PD and divergent, longitudinal clinical outcomes. Similar models have been used previously to study repeated measures in PD [12,14] and incorporate the individual subject repeated measures as random effects and can handle missing data. Longitudinal clinical scores (MDS-UPDRS-Total score, MDS-UPDRS-Part III, and MoCA) at baseline and years 1 and 2 were included. We also examined the 5 cognitive domain composite scores from the NPSY assessment. The fixed effects of each model were the Probability of PD, as determined by the classifier model, age, and interactions with year. As in previous studies [12], we subtracted the mean age of our participants (68) so that the intercept represented an average-aged participant. Sex, education, and their interactions with year did not generally improve the models and therefore were not included. The random effects were the intercepts and slopes of the subjects’ clinical scores through year. Analysis was performed in MATLAB using the `fitlme` function. Due to the exploratory nature and the large number of tests, traditional methods for correcting for multiple comparisons would be too restrictive at 5%. We present

Fig. 2. Classifier Results. Probability of PD (A) as assigned by the classifier for each subgroup. Square and bars indicate mean and standard error, respectively. ROC-AUC scores are shown for each PD-subgroup (B). The confusion matrix (inset) reveals a sensitivity of 83% and a specificity of 78%.
p-values uncorrected and discuss possible trends in the context of confidence intervals and skepticism regarding weak, but significant, predictions.

3. Results

3.1. Classifier Results

Fig. 2A compares the distribution of individual Probability of PD from the classifier for the 3 subgroups of PD and controls. There was a progression of increased confidence as PD participants progressed from PD-CN through PDD. We fit a linear mixed model of Probability of PD with group (control or PD subgroup) and age as fixed effects. Compared to PD-CN, the Probability of PD was significantly less for Controls (−30.3%, p < 0.000) and significantly greater for PDD (+11.6%, p < 0.005), while no significant difference was found for PD-MCI (+4.0%, p < 0.282). See Supp. Fig. 5 for an exploration of PD subgroups using dimensionality reduction techniques.

We plotted the ROC curve for each of the PD subgroups and PD overall (Fig. 2B). The overall AUC was 0.88, with the subgroups showing progressively increase AUC with progressing cognitive deficits, and the AUC of PDD alone being 0.95. We computed the confusion matrix for the binary task of control vs. PD (Fig. 2B, inset). The confusion matrix demonstrates the proportions of correctly and incorrectly classified data and the classifier was found to have a sensitivity of 83% and a specificity of 78%. Overall, based on the cross-validation scores, the classifier was significantly more accurate (median 82 ± 6.7%) when compared to a random dummy classifier (median 50 ± 9.3%) on a Mann–Whitney test (U = 100.0, p < 0.000).

Fig. 3. Probability of PD correlated with scores of overall disease burden (MDS-UPDRS-TOTAL – A) motor decline (MDS-UPDRS-Part III – B), global cognition (MoCA – C), and global cognition with PD only (MoCA – D) at baseline (year 0). TN = True negative controls. FP = False positive controls.
3.2. Clinical scores at baseline

We explored the ability of Probability of PD from the classifier to predict clinical measures of cognitive and motor impairment. We correlated the MDS-UPDRS-Total and MDS-UPDRS-Part III motor examination scores with Probability of PD (Fig. 3A and B, respectively). A significant correlation was observed, indicating that the classifier was sensitive to the severity of overall disease burden (Pearson’s, r(118) = 0.22, p < 0.016) and motor symptoms (Pearson’s, r(119) = 0.24, p < 0.001).

Because our controls also completed the MoCA at baseline, we plotted the correlation of MoCA with Probability of PD for all participants (Fig. 3C) and PD only (Fig. 3D) and found a moderate and significant correlation for all participants (Pearson’s, r(217) = −0.32, p < 0.000). False positive controls (defined by Probability of PD = 0.5) had significantly lower MoCA scores (median 26) than true negative controls (median 28) (Mann-Whitney U: U = 3915.5, n1 (TN) = 74, n2 (FP) = 22, p < 0.0004). There was no correlation with MoCA found for PD only (Pearson’s, r(121) = −0.089, p < 0.333).

3.3. Clinical scores longitudinally

For PD participants, there were clear trends for nearly all clinical measures, such that those individuals who were assigned a higher Probability of PD by the classifier tended to have significantly worse clinical scores at baseline, and those trends were stable longitudinally (see Supp. Fig. 11). To investigate these trends, we computed linear mixed-effects models for MDS-UPDRS-Total Score (overall disease burden), MDS-UPDRS-Part III (motor impairment), MoCA (overall cognitive decline), and cognitive domain composite scores. See Supplementary Tables 4-11 for the complete model outputs.

3.3.1. MDS-UPDRS-total

For MDS-UPDRS-Total, measuring overall PD disability, the model demonstrated a weak interaction with year (t = 2.04, p < 0.042). There was also a weak effect of Probability of PD on baseline scores (t = 2.11, p < 0.035). Following previous work [12], another view of these models is to list contributing effects in a linear equation:

MDS-UPDRS-TOTAL = 31.1 + 22.2 * PROBABILITY OF PD + 0.38 * AGE + 9.6 * PROBABILITY OF PD * YEAR + 0.30 * AGE * YEAR

This equation suggests that a baseline score of 31.1 is adjusted by Probability of PD and age both at baseline and longitudinally. For example, for a 68-year-old participant assigned a probability of PD of 0.9 at baseline, we would predict an MDS-UPDRS-TOTAL score of 68 at year 2, whereas we would predict a score of only 35 if they were assigned a probability of PD of 0.1.

3.3.2. MDS-UPDRS-part III (motor)

For MDS-UPDRS-Part III, which characterizes motor impairment, we found a weak effect of Probability of PD on baseline motor scores (t = 2.29, p < 0.023) and no interaction with year (t = 1.17, p < 0.242).

3.3.3. MoCA

For MoCA, while we did observe similar trends as with MDS-UPDRS scores, we found no significant effects of Probability of PD at baseline (t = −0.79, p < 0.430) or the interaction with year (t = −1.58, p < 0.116).

3.3.4. Cognitive domain composite scores

While MoCA was not predicted by Probability of PD, we found that individual domain scores of cognitive function were strongly predicted. At baseline, Probability of PD strongly predicted scores of attention/working memory (t = −3.52, p < 0.000) and executive function (t = −4.18, p < 0.000), moderately predicted visuospatial function (t = −2.40, p < 0.017), but not language (t = −0.56, p < 0.577) or memory (t = −1.62, p < 0.106). No significant interactions with year were found. As IPAST directly measures attention, executive function, and visuospatial function, but not language or memory, these results are consistent with the task design and demonstrate its sensitivity to the spectrum of cognitive decline present in the PD participants.

4. Discussion

Here we operationalized measures of saccade, pupil, and blink from the IPAST to achieve high sensitivity to the spectrum of PD. Functional feature extraction was used to provide a complete feature set that captured motor and cognitive impairment. Using machine learning, these features were used to tune a classifier which was sensitive to the landscape of PD across different classifications of cognitive health through PDD. Longitudinal predictions were less convincing across the first 2 years, but possible trends in overall disease burden were revealed. This work provides a framework for true clinical utility, where a simple eye tracking task could be used to screen for PD and predict outcomes.

A wide range of eye-tracking measures are altered in pro/anti-saccade tasks throughout the stages of PD progression, with pupil and blink disturbances more recently being implicated during prodromal stages [9,13,15,16]. Combining features of saccade, pupil, and blink in a complex task was expected to inform more fully the spectrum of PD. Indeed, controls and PD cognitive subgroups (Fig. 2A) formed a continuum of cognitive decline, where the classifier was more confident in PD diagnosis with increasing cognitive dysfunction. It should however be noted that cognitively burdened subgroups were concomitantly impaired in motor scores as well (Table 1). This supports the notion that eye-movement measures capture the complex interplay of cognitive and motor impairment in PD [12,14,16]. A machine learning pipeline was able to condense this multidimensional output (Supp. Fig. 5) into one continuous index predictive of PD subgroups and severity on the MDS-UPDRS scales. Consistent with previous results [12,14,16], we demonstrated that the output of our classifier was sensitive to motor and cognitive function.

To our knowledge, this machine learning model was the first built on a large, diverse cohort of PD and an equally large cohort of controls. In the same way that previous work has focused on specific eye-movement behaviors and how they are altered in specific stages of PD [12,16], most of the machine learning models built for PD classification have focused on specific eye-movement metrics in small cohorts of PD (See our previous work [29] and Supplementary Discussion for examples). While these studies achieved similar measures of sensitivity and specificity to ours, none have covered the diverse spectrum of PD to the extent achieved by ONDRI, which allowed us to explore the model’s sensitivity to staging and longitudinal outcomes. Future work on neurodegeneration will need to increasingly consider larger cohorts from diverse and converging disorders.

The goal of ONDRI was to study the subgroups and converging pathologies of a range of neurodegenerative disorders. A key question is what exact etiology eye-tracking measures, and the resulting classifier, are sensitive to? The proteinopathy of PD and other neurodegenerative disorders is complex and diverse, especially as they progress [4]. In PD, clinical evidence suggests that the prodromal buildup of α-synuclein in the brain enters either in the lower brainstem or the cortex before progressing to nigrostriatal pathways [3,5,7]. Comorbid accumulations of proteins are then observed as PD progresses to PDD. Diffuse amyloid-β (Aβ) plaques and tau neurofibrillary tangles (NFTs) are found in striatum in PDD and cortical α-synuclein, tau, and Aβ pathologies together may best predict PDD [30]. As eye tracking is a proxy for neurobiological intactness, it is likely measuring processes associated with the progression of α-synuclein and these converging and comorbid proteinopathies. As neurodegenerative disorders progress, it is likely that they converge in neurobiological dysfunction and that any measure, such as eye-tracking, that is more directly probing this dysfunction, will not necessarily follow clinical diagnoses. Nevertheless, these measures...
appear to predict subgroups accurately in PD. Future work will use these measures to potentially identify new biotypes from the larger ONDRI dataset of multiple and converging neurodegenerative disorders.

There are limitations to this work. Without longitudinal data for our controls, it is difficult to know which measures are specifically sensitive to PD and not age, especially cognitive scores. Longitudinal progressions of disease across the first 2 years appear to be hampered by the loss of more burdened participants in year 2. Because linear-mixed models assume that data is missing at random, this may impact the longitudinal results and more complex models may have more power in revealing longitudinal trends. IPAST was used to capture more complex saccadic behavior and while it is relatively easy to administer and understand for most participants, it can still be challenging, especially for those with cognitive impairment. The PDD cognitive subgroup was more burdened by motor and autonomic dysfunction, which was difficult to disentangle in our model. The impact of medications and sex differences are confounding factors. The degree to which the classifier is sensitive to these factors is discussed in our supplementary material, but a more thorough study is warranted. The ONDRI cohort did not include individuals in Hoehn & Yahr stages 4 and 5. The extent to which these models apply to the more advanced stages of PD is not known.

5. Conclusion

We have demonstrated the potential of eye tracking to classify PD in the clinic. With an even larger sample and optimization of the task and equipment, we believe this technology could be ready to screen potential PD and be used as a therapeutic marker. Our work provides a meaningful continuous index that lays the framework for larger clinical trials. A confidence value such as this could be used for neurodegeneration more broadly, much like continuous blood pressure continuous values are used to assess risk of heart disease.

Funding

This research was conducted with the support of the Ontario Brain Institute, an independent non-profit corporation, funded partially by the Ontario government, and a Foundation grant to DPM from the Canadian Institutes of Health Research grant (MOP-FDN-148418). DPM was supported by the Canada Research Chair Program.

Authors’ roles

All authors have contributed to and approved this final manuscript of this work. D.C. Brien is the primary author and responsible for the task implementation, eye-tracking programming/delivery/training at each site, analysis, writing, and editing. HCR and RY collected control data and edited. HCR calculated LED measures. J.H advised on methodology and edited. C.A advised on methodology, analysis, and edited. D.C. Crawford, J. Henderson, C. Kenndard, Ocular motor abnormalities in neurodegenerative disorders, Eye 29 (2) (2015) 200–207, https://doi.org/10.1038/eye.2014.276.

References
