npj Digital Medicine

https://doi.org/10.1038/s41746-026-02753-9

Article in Press

iPad Eye Tracking Reproduces Clinical Grade Oculomotor Differences in Parkinson's Disease

Received: 18 December 2025

Accepted: 4 May 2026

Cite this article as: Koerner, J., Zou, E., Karl, J.A. *et al.* iPad Eye Tracking Reproduces Clinical Grade Oculomotor Differences in Parkinson's Disease. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02753-9

Jamie Koerner, Erin Zou, Jessica A. Karl, Cynthia Poon, Roneil G. Malkani, Leo Verhagen Metman, Charles G. Sodini, Vivienne Sze, Thomas Heldt & Fabian J. David

iPad Eye Tracking Reproduces Clinical Grade Oculomotor Differences in Parkinson's Disease

Jamie Koerner[1], Erin Zou[2], Jessica A. Karl[3], Cynthia Poon[3], Roneil G. Malkani[3], Leo Verhagen Metman[3], Charles G. Sodini[1], Vivienne Sze[1], Thomas Heldt[1*†], Fabian J. David[4,5*†]

[1]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA.
[2]Chicago College of Osteopathic Medicine, Midwestern University, Downers Grove, IL, USA.
[3]Department of Neurology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA.
[4]Department of Physical Therapy and Human Movement Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA.
[5]US Field Medical Affairs Neuromodulation, LivaNova, Houston, TX, USA.

†Equally credited as senior authors.
*Corresponding author. E-mail: thomas@mit.edu; fabian.david@livanova.com;
Contributing authors: jkoerner@mit.edu; erin.zou@midwestern.edu; jessica.karl@nm.org; cynthia.poon@northwestern.edu; r-malkani@northwestern.edu; leonard.verhagenmetman@nm.org; sodini@mit.edu; sze@mit.edu;

Abstract

iPad-based eye tracking could support Parkinson's disease (PD) screening and longitudinal monitoring by enabling objective, low-cost, portable assessment of oculomotor function. We previously validated an iPad-based eye-tracking system against the EyeLink 1000 Plus for temporal and spatial saccade metrics. Here, in a convenience sample of 19 healthy controls and 12 patients with PD, we recorded eye movements simultaneously with both devices during pro-saccade, anti-saccade, memory-guided saccade, and self-generated saccade tasks. Across all pre-specified metrics, statistically significant PD–HC differences and null results were concordant between devices. In addition, saccade-level mixed-effects models showed small group × device interaction effects that remained below literature-based benchmarks for clinically meaningful PD–HC differences, indicating that iPad-based measurements preserved benchmark clinical-grade group-level effects. A compact three-metric iPad-based classifier comprising AS directional error rate, AS gain, and MGS gain supported strong subject-level PD–HC discrimination, with an area under the receiver operating characteristic curve of 0.98, sensitivity of 0.91, specificity of 1.00, and accuracy of 0.96. These findings support scalable tablet-based oculomotor assessment for PD-related screening and longitudinal monitoring.

Keywords: Parkinson's disease, eye tracking, iPad-based assessment, screening, longitudinal monitoring

Introduction

Parkinson's disease (PD) is a common neurodegenerative disorder with a rapidly growing incidence [1, 2] whose motor manifestations extend beyond limb and axial symptoms to include prominent abnormalities of eye movement control, or oculomotor deficits, reflecting dysfunction across cortico-basal ganglia and brainstem circuits [3–5]. Saccadic eye movements are especially informative because they are generated by a well-characterized network spanning cortical, basal ganglia, cerebellar, and brainstem regions, and distinct patterns of impairment can therefore reflect dysfunction in specific pathways [5–7]. In PD, volitional saccades such as anti-saccades (AS) and memory-guided saccades (MGS) are often hypometric and delayed, and patients with PD typically commit more directional errors on AS tasks than healthy controls (HC), whereas reflexive pro-saccades (PS) are comparatively spared [6, 7]. These features position saccadic eye movements as promising, noninvasive biomarkers for PD-related screening and disease monitoring.

Traditional eye movement assessment in research and specialized clinics relies on high-grade infrared video-oculography systems such as the EyeLink 1000 Plus (SR Research), which provide millisecond-level temporal resolution and sub-degree spatial accuracy. These systems remain benchmark tools for research and specialist assessment because of their precision [4, 8–12]. However, they are expensive, bulky, and dependent on dedicated infrastructure and trained operators, which limits deployment for large-scale screening, longitudinal monitoring, and routine use outside tertiary clinical centers [13]. In parallel, other proposed biomarker modalities, including neuroimaging, cerebrospinal fluid assays, and skin biopsies, can be invasive, costly, or logistically demanding [14, 15]. Together, these constraints motivate the search for scalable, lower-cost approaches that can capture PD-related oculomotor signatures outside specialized eye-tracking laboratories. If validated, a more portable and accessible eye-tracking approach could support quantitative oculomotor assessment in primary care and general neurology clinics, where a low-cost front-end screening or referral-support tool could help identify patients who may benefit from referral for specialist evaluation. In movement-disorders clinics, it could also support repeated measurements across routine follow-up visits to track oculomotor change over time. Longer term, such an approach could make home-based longitudinal monitoring between clinic visits more feasible.

Recent work has increasingly explored more accessible eye-tracking platforms for this purpose. Screen-based infrared systems such as the Tobii TX300 have been used in PD research [16] and in other clinical and behavioral settings [17, 18], offering a useful balance between cost and measurement precision. However, they still depend on fixed displays, controlled lighting, and structured environments, which limits portability and scalability outside laboratory or clinic-based settings [19]. At the same time, recent reviews have highlighted both the promise of eye-tracking markers in PD and the growing interest in translating them into practical clinical tools [5, 20, 21]. Consumer-grade approaches have also been explored, including head-mounted systems with integrated eye tracking, such as NeuroSync's virtual-reality-based platform for concussion diagnosis, although such systems remain relatively costly, bulky, and potentially uncomfortable for routine use [22].

Among scalable options, mobile devices such as smartphones and tablets are particularly attractive because they combine portability, relatively low cost, and wide availability. On smartphones, gaze estimation using standard front-facing cameras has advanced rapidly. Work by Valliappan et al. [23] showed that accurate smartphone-based eye tracking without specialized hardware is feasible, and later work extended these ideas to in-home and clinically oriented settings [24]. Portable smartphone-supported platforms such as BrainEye further illustrate the appeal of scalable video-based neurological assessment, although these efforts have largely targeted conditions such as concussion rather than PD [25]. Tablets have likewise been used for oculomotor assessment across the adult lifespan [26], in psychiatric research [27], and increasingly in PD-related studies. For example, the iPad-based Eye-Tracking Neurological Assessment (ETNA) platform has been shown to distinguish PD from HC, and later studies reported associations between oculomotor metrics and both cognition and disease severity in PD [28, 29]. Together, these studies suggest that smartphone- and

tablet-based systems can capture PD-relevant oculomotor patterns, including group differences and associations with cognition or disease severity.

However, portable consumer-device approaches have not been validated against a benchmark eye tracker in a way that establishes whether disease-relevant group-level effects are preserved. This question is important because consumer devices rely on visible-light RGB imaging, lower temporal resolution, and less specialized hardware than infrared clinical eye trackers, all of which can introduce measurement error in gaze estimation and downstream saccade metrics. Even modest subject-level error could attenuate true PD–HC differences or create spurious apparent differences in downstream group comparisons. At the same time, recent deep-learning eye-tracking models have shown increasingly strong performance on mobile devices, raising the possibility that, when paired with an accurate gaze-estimation pipeline, a consumer-device approach may still preserve disease-relevant group-level signatures despite these technological limitations. Thus, the key unresolved question is not simply whether a consumer tablet can track gaze reasonably well at the subject level, but whether it can reproduce the PD–HC conclusions obtained with a benchmark eye tracker.

We recently demonstrated that an iPad-based eye-tracking system using a deep-learning-based gaze-estimation model achieved subject-level agreement with the EyeLink 1000 Plus for temporal and spatial saccade metrics across multiple tasks [30]. Building on that work, the present study asks whether this validated iPad-based system is also accurate enough to preserve PD–HC differences at the group level, using the EyeLink 1000 Plus as the benchmark. Using simultaneous recordings in the same participants, we tested whether the iPad-based eye-tracking system preserves PD-related oculomotor signatures relative to this reference system. Conceptually, this study therefore shifts the question from subject-level device agreement to preservation of clinically relevant PD-related conclusions, which is a necessary step toward translational deployment of tablet-based PD assessment.

To this end, we recorded eye movements simultaneously with an iPad and the EyeLink 1000 Plus while 19 HC and 12 patients with PD performed a battery of PS, AS, MGS, and self-generated saccade (SGS) tasks. We first assessed whether statistically significant PD–HC differences and null results were concordant across devices for a set of pre-specified temporal and spatial metrics, and whether device-related deviations in those group differences were small relative to literature-based PD–HC benchmarks. We then evaluated whether iPad-derived oculomotor metrics could distinguish PD from HC using a prespecified regularized logistic-regression classifier, and performed a separate nested top-k analysis to determine how compact the discriminatory feature set could be while retaining strong performance. In this way, the present study serves as a proof-of-concept for how a validated iPad-based eye-tracking system might support scalable oculomotor screening and longitudinal monitoring in clinical and, ultimately, home-based settings.

Results

PD–HC differences across EyeLink and iPad recordings

Both the EyeLink and iPad data showed consistent between-group differences between PD and HC. In the AS task, participants with PD had a significantly higher directional error (DE) rate than HC participants on both systems (EyeLink: $\Delta = 0.18$, $p < 0.01$; iPad: $\Delta = 0.19$, $p < 0.01$) (Figure 1a; Table 1). Saccade gain was significantly lower in the PD group for AS and MGS on both devices (AS: EyeLink $\Delta = 0.15$, $p < 0.05$; iPad $\Delta = 0.18$, $p < 0.05$; MGS: EyeLink $\Delta = 0.20$, $p < 0.01$; iPad $\Delta = 0.19$, $p < 0.01$) (Figure 1c; Table 1). In the SGS task, participants with PD had a lower instantaneous primary saccade rate (IPSR) than HC participants on both EyeLink and iPad (EyeLink: $\Delta = 0.53\ s^{-1}$, $p < 0.05$; iPad: $\Delta = 0.53\ s^{-1}$, $p < 0.05$) (Figure 1d; Table 1).

Metrics without PD–HC differences

Metrics that did not differ between groups on the EyeLink also showed no significant difference on the iPad. Saccade latency in the PS, AS, and MGS tasks (Figure 1b; Table 1), saccade gain in the PS and SGS tasks (Figure 1c; Table 1), and inhibitory

error (IE) rate in the MGS task (Figure 1a; Table 1) were comparable between PD and HC on both devices.

Mixed-effects analysis of device agreement

To quantify device agreement in PD–HC effects, we fit saccade-level linear mixed-effects models with fixed effects of group, device, and their interaction and a random intercept for subject, as described in Statistical Analysis. Table 2 reports the estimated group × device interaction for each task and metric, along with 90% confidence intervals. For latency, group × device interaction estimates ranged from −1.7 to 7.0 ms across PS, AS, and MGS. For gain, interaction estimates were between −0.02 and 0.03 across PS, AS, MGS, and SGS. For AS DE rate and MGS IE rate, interaction estimates were within ±0.01 in proportion, and for SGS IPSR the interaction estimate was $-0.005\ s^{-1}$. These interaction estimates are later compared with the benchmark PD–HC differences derived from the literature (Expected saccade changes in PD) to contextualize device-related effects.

Classification of PD vs. HC using iPad metrics

We next examined whether iPad-derived oculomotor metrics could discriminate PD from HC at the subject level. Classification analyses were restricted to the 26 subjects with complete iPad-derived data across all candidate metrics (15 HC, 11 PD). We first evaluated three prespecified elastic-net logistic-regression models reflecting increasing biological constraint of the candidate feature space (Table 3). Our primary model excluded PS-derived and latency-derived metrics, whereas two broader sensitivity models either included all candidate metrics or excluded PS-derived metrics only.

In the biologically constrained primary model (no PS, no latency), nested LOSO elastic-net logistic regression yielded an AUC of 0.90 (95% CI [0.70, 1.00]), sensitivity of 0.73 (95% CI [0.42, 1.00]), specificity of 1.00 (95% CI [1.00, 1.00]), and accuracy of 0.89 (95% CI [0.73, 1.00]) at a fixed decision threshold of 0.5 (Table 3). Predicted probabilities were also well calibrated, with a Brier score of 0.12, calibration intercept of -0.01, and calibration slope of 0.93 (Table 4). Across outer LOSO folds, the most stable contributors in this primary model were AS DE rate (positive coefficient), AS gain (negative coefficient), and MGS gain (negative coefficient), with other coefficients substantially smaller or less stable.

We then assessed the robustness of classification performance to broader feature inclusion. Allowing all candidate metrics into the elastic-net model improved discrimination modestly, yielding an AUC of 0.94 (95% CI [0.82, 1.00]), sensitivity of 0.82 (95% CI [0.56, 1.00]), specificity of 1.00 (95% CI [1.00, 1.00]), and accuracy of 0.92 (95% CI [0.81, 1.00]) (Table 3). The intermediate model excluding PS-derived metrics only gave similar but slightly lower performance, with an AUC of 0.93 (95% CI [0.76, 1.00]), sensitivity of 0.82 (95% CI [0.55, 1.00]), specificity of 0.87 (95% CI [0.67, 1.00]), and accuracy of 0.85 (95% CI [0.69, 0.96]) (Table 3). Thus, although broader candidate sets provided somewhat higher numerical performance, the biologically constrained primary model remained highly discriminative and retained the clearest mechanistic interpretation.

To determine how many features were required for strong discrimination, we performed a separate nested top-k analysis using the full candidate feature space, with training-only feature ranking and ridge logistic regression for the resulting fixed subsets. Performance improved substantially from one to three features and then plateaued. Specifically, the top-1 model achieved an AUC of 0.70, sensitivity of 0.46, specificity of 0.80, and accuracy of 0.65; the top-2 model achieved an AUC of 0.82, sensitivity of 0.73, specificity of 0.73, and accuracy of 0.73; and the top-3 model achieved the best overall performance, with an AUC of 0.98 (95% CI [0.90, 1.00]), sensitivity of 0.91 (95% CI [0.56, 1.00]), specificity of 1.00 (95% CI [1.00, 1.00]), and accuracy of 0.96 (95% CI [0.89, 1.00]) (Table 5). Importantly, the top-3 feature set was stable across LOSO folds and always consisted of AS DE rate, AS gain, and MGS gain. These results indicate that most of the discriminatory signal in the iPad-derived feature space was captured by a compact three-metric model.

Figure 2 shows each participant's mean values for the two most informative metrics, AS DE rate and AS gain, with motor scores indicated in brackets for patients with PD. HC participants cluster in the bottom-right region, characterized by high gain and low DE rate, whereas participants with PD are more dispersed, exhibiting low gain, high DE rate, or both. One participant with PD (P12) was misclassified as HC, and their data point in Figure 2 resembles those of the HC group.

Expected saccade changes in PD

To interpret any iPad–EyeLink differences estimated by the mixed-effects models in the context of clinically meaningful disease effects, we reviewed prior work on PD–HC differences in saccadic eye movements, focusing our benchmark analysis on three tasks that span distinct domains of oculomotor control: PS as a reflexive, visually guided task; AS as a voluntary, visually guided task that additionally taxes inhibitory control and vector inversion; and SGS as a self-generated task in which saccades are voluntarily initiated but not directly driven by transient visual onsets. For each metric in these tasks, we extracted from the literature the smallest statistically significant PD–HC difference (by absolute value) and treated this value as a conservative benchmark for a clinically meaningful group effect. The full distributions of PD–HC differences reported in previous studies, together with these benchmark thresholds, are visualized in Figure 3. In each panel, the dashed vertical line marks zero (no PD–HC difference), and the solid blue line marks the benchmark threshold derived from prior work; note that the same study can appear multiple times if it examined different subgroups, conditions, or task paradigms. Red indicates the difference is statistically significant ($p < 0.05$). Although MGS is included in our primary analyses as a key memory-guided, voluntary task, we restricted the benchmark synthesis to PS, AS, and SGS to maintain a compact set of paradigms that collectively sample reflexive, visually guided voluntary, and self-generated saccadic control.

For AS DE rate, we used PD–HC DE differences reported in a recent meta-analysis of AS performance in PD that compiled DE rates from 34 studies and reported PD–HC effects across cohorts [31]. From this full set of effects, the minimal statistically significant PD–HC difference was 0.04 (i.e., a 4 percentage point absolute increase in errors for PD), which we adopted as our DE benchmark. Figure 3a plots the PD–HC DE differences from all included cohorts, with the dashed line at zero and the solid blue line at this 0.04 threshold.

For PS and AS latency, we used effect sizes from our previous study [30], in which PD–HC latency differences were derived from a curated subset of latency estimates taken from the same AS meta-analysis. In that analysis, the smallest statistically significant PD–HC differences were 19.0 ms for PS latency and 57.7 ms for AS latency. Figure 3b shows the PS and AS latency differences reported across studies, with dashed lines at zero and solid blue lines indicating these latency benchmarks.

For PS and AS gain, we again relied on our prior work [30], which originally reported PD–HC differences in saccade amplitude. Because the AS meta-analysis did not provide amplitude or gain values directly, we returned to the primary studies considered in that meta-analysis and extracted a subset of reported amplitude and gain effects for PS and AS, harmonizing them on an amplitude scale. In the present figure, we re-express these differences as gain by dividing amplitude by target eccentricity, and then identify the smallest statistically significant PD–HC differences in gain, yielding benchmarks of -0.06 for PS gain and -0.25 for AS gain. For SGS gain we drew on our earlier compilation of SGS studies in PD [30] and applied the same amplitude-to-gain conversion, yielding a smallest statistically significant PD–HC gain difference of -0.07. Figure 3c displays the PD–HC gain differences across studies for PS, AS, and SGS, with the dashed and solid blue lines marking zero and the corresponding gain benchmarks, respectively.

Finally, for SGS IPSR, we again relied on our previous work on SGS in PD [30], in which we compiled PD–HC differences in SGS rate across published cohorts and expressed them as IPSR. In that analysis, the smallest statistically significant reduction in SGS rate corresponded to a PD–HC difference of approximately –0.4667 s^{-1}. Figure 3d shows the PD–HC differences in SGS IPSR across studies, with the dashed line at zero and the solid blue line marking this benchmark reduction.

Taken together, Figures 3a–d therefore show, for each metric, the PD–HC difference estimates from all included studies, while the overlaid benchmark lines indicate the smallest statistically significant PD–HC differences that we use as reference thresholds when interpreting the mixed-effects device agreement analyses.

Discussion

This study demonstrates that an iPad-based eye-tracking system can reproduce PD-related group differences and null effects observed with a clinical-grade EyeLink 1000 Plus across multiple saccade tasks. Increases in AS DE rate, reductions in saccade gain for AS and MGS, and reductions in SGS IPSR were all detected by both devices with closely matched effect sizes. At the same time, metrics without significant PD–HC differences on the EyeLink, including PS gain, latency across PS, AS, and MGS, and MGS IE rate, also showed no differences on the iPad. Together, these findings indicate that the tablet-based system preserves patterns of PD-related oculomotor abnormalities without introducing obvious device-specific artifacts at the group level.

Our prior iPad–EyeLink study established that selected saccade metrics derived from the iPad agreed closely with those from the EyeLink at the subject/task level. The conceptual contribution of the present study is different and more clinically relevant. Here, the central question is not whether the two devices produce similar measurements in general, but whether the iPad preserves the downstream inferences that matter for PD assessment. Using simultaneous recordings in PD and HC participants, we therefore tested whether PD-related group-level conclusions were preserved across devices, whether group-by-device interaction terms remained small relative to literature-based benchmarks for clinically meaningful disease effects, and whether iPad-derived metrics supported compact subject-level PD discrimination. In this sense, the present work moves beyond device comparison toward translational validation by showing that residual tablet measurement error is small enough that the PD-related group-level conclusions examined here are preserved. This is an important step toward a deployable PD tool, because screening and longitudinal monitoring depend not on raw agreement alone, but on faithful preservation of disease-related signal under practical sensing constraints; the mixed-effects analysis provides a more formal test of that preservation.

The mixed-effects analysis, interpreted in light of expected PD-related changes from the literature, further supports the conclusion that any residual device-related discrepancy between the two systems is small relative to clinically meaningful disease effects. Saccade-level linear mixed-effects models yielded group × device interaction estimates that were all close to zero, with 90% confidence intervals that were narrow in absolute terms and centered near no difference (Table 2). When benchmarked against the smallest statistically significant PD–HC differences reported in prior work (Figure 3), these interaction estimates were consistently smaller than the disease effects of interest. For example, the AS DE benchmark was a 0.04 absolute increase in error rate in PD, whereas the AS DE interaction estimate was −0.007 with a 90% confidence interval of [−0.063, 0.050], indicating that any device-related distortion of the PD–HC DE effect is unlikely to exceed the range of the smallest disease effects reported in the literature. Latency interactions for PS, AS, and MGS had point estimates between roughly −2 and 7 ms, with 90% confidence intervals contained within approximately ±20 ms, whereas benchmark PD–HC latency differences were on the order of 20–60 ms. For gain and SGS IPSR, interaction point estimates were on the order of 0.02 or less in absolute value, compared with benchmark PD–HC gain differences of 0.06–0.25 and SGS IPSR differences of about 0.47 s^{-1}. Taken together, these comparisons suggest that any discrepancy between iPad- and EyeLink-derived PD–HC effects is unlikely to obscure or mimic clinically meaningful group differences.

This interpretation is consistent with our prior validation study, in which simultaneous iPad and EyeLink recordings showed that the iPad reproduced subject- and task-level temporal and spatial saccade measures with small error relative to the EyeLink. In particular, after averaging across saccades within a task, the mean latency error was about 2 ms, the mean amplitude error for the PS, AS, and MGS tasks was about 0.7°, and for the SGS task the mean IPSR error was about 0.003 s^{-1} with a mean amplitude error of about 1.6°. The present study extends those validation

results from the level of subject- and task-level measurement agreement to the level of clinically relevant downstream inference: despite the substantial hardware differences between a 60 Hz RGB tablet camera and a 1000 Hz infrared eye tracker, the residual iPad–EyeLink differences remain small enough that PD–HC group effects and null results are reproduced faithfully across devices. Put differently, the modest temporal and spatial measurement errors documented previously do not appear to materially distort the downstream group comparisons that are central to PD-related oculomotor assessment.

At the same time, these residual discrepancies should be interpreted relative to the disease effects they would need to preserve. For latency, although the interaction estimates were small relative to literature-derived PD–HC effects, the corresponding confidence intervals suggest that caution is warranted when interpreting very subtle latency abnormalities near the lower end of the reported range, particularly at the individual-subject level. For gain, the interaction estimates likewise remained small relative to benchmark PD–HC effects, suggesting that residual device-related differences are unlikely to obscure the larger gain abnormalities typically reported in PD. More broadly, these findings indicate that the current iPad-based system is well suited for detecting the robust PD-related abnormalities examined here and for supporting screening or longitudinal tracking at the group or cohort level, while very small effect sizes or subtle metric changes in isolation should be interpreted more cautiously.

The classification analysis extends these findings by showing that iPad-derived metrics can distinguish PD from HC under both biologically constrained and broader modeling assumptions. In our primary classifier analysis, we prespecified a biologically constrained feature space that excluded PS-derived and latency-derived metrics, motivated by prior literature suggesting relative preservation of reflexive PS in PD [7] and the possibility that latency abnormalities emerge later in the disease course [6] than the early/mid-stage disease represented in our cohort. Within this restricted feature space, nested elastic-net logistic regression still achieved strong discrimination, with an AUC of 0.90. The most stable contributors in this primary model were AS DE rate, AS gain, and MGS gain, indicating that much of the discriminatory signal was concentrated in metrics reflecting impaired inhibitory control and hypometric volitional saccades.

Broader sensitivity analyses supported the robustness of this signal. Allowing all candidate features into the elastic-net classifier modestly improved discrimination (AUC 0.94), and the intermediate model excluding PS-derived metrics only also performed well (AUC 0.93). However, the broader models did not alter the core biological interpretation: AS DE rate, AS gain, and MGS gain remained the dominant contributors. These results suggest that the strongest PD-related discriminatory information captured by the iPad lies in the same domains that showed concordant group-level abnormalities across devices.

The nested top-k analysis provided a direct answer to the question of model complexity. When all candidate iPad-derived features were allowed to compete and ranked within the training data only, performance improved markedly from one to three features and then effectively plateaued. The top-3 model consistently comprised AS DE rate, AS gain, and MGS gain in every outer LOSO fold and achieved an AUC of 0.98, sensitivity of 0.91, specificity of 1.00, and accuracy of 0.96. This finding is important for two reasons. First, it shows that the discriminatory signal does not depend on a large or unstable feature set; rather, it can be captured by a small and highly interpretable subset. Second, because the same three features emerged even when the full candidate feature space was considered, the compact classifier result is not simply an artifact of the biologically motivated pre-restriction.

Importantly, the metrics emphasized by the classifier align with both the mixed-effects and benchmark analyses. AS DE rate and AS gain reflect impaired inhibitory control and hypometric voluntary saccades, while MGS gain captures reduced scaling of memory-guided saccades; these metrics show robust PD–HC differences across devices and fall in domains where the literature indicates sizeable disease effects [4–7, 11, 31]. The fact that these same metrics both (i) exhibit concordant PD–HC differences on the iPad and EyeLink, and (ii) support strong subject-level discrimination, suggests that the iPad is not merely reproducing group means but is

also capturing subject-level variation in PD-related oculomotor dysfunction that is informative for discrimination [4–7].

By showing that the iPad reproduces disease-related group effects measured with the EyeLink 1000 Plus and that iPad-derived metrics also support strong PD–HC discrimination in this cohort, this work extends prior validation of pointwise measurement agreement and supports the use of consumer tablets as practical tools for quantitative oculomotor assessment in PD. At the same time, the present classification results should not be interpreted as defining a finalized operating threshold for real-world screening. In practice, the clinically acceptable balance between sensitivity and specificity would depend on the intended use case. In a front-end screening context, one would often prioritize sensitivity to reduce the chance of missed cases, whereas in a referral-support or triage setting a different balance may be appropriate. The high specificity observed here is encouraging, but because it was obtained in a small internally cross-validated cohort, it should not be assumed to generalize unchanged to broader clinical populations. Rather, these results should be viewed as proof-of-concept evidence that tablet-derived oculomotor metrics carry disease-relevant signal strong enough to support subject-level PD discrimination. Defining clinically appropriate operating thresholds will require prospective validation in larger and more heterogeneous independent cohorts. The most immediate clinical value of a validated tablet-based system may be as an accessible front-end tool for quantitative oculomotor assessment in settings where conventional infrared eye trackers are rarely available. These include primary care, general neurology clinics, and community-based settings. In such settings, tablet-based testing could help identify patients whose oculomotor profile warrants referral for specialist evaluation, while also providing a standardized and objective complement to bedside examination. In movement-disorders clinics, the same platform could support repeated measurements across routine follow-up visits using a low-burden, standardized protocol to track change over time. This type of repeated quantitative assessment could be especially useful for detecting gradual oculomotor change over time, which may be difficult to appreciate reliably from routine bedside examination alone. Longer term, its portability and low cost may also make it attractive for home-based longitudinal monitoring between clinic visits, although such use will require further validation under less constrained real-world conditions. Importantly, we view such a system not as a replacement for specialist diagnosis, but as a scalable quantitative tool that could extend access to oculomotor assessment and make repeated measurements more feasible across clinical contexts.

Several limitations should be acknowledged. The sample size is modest and drawn from a single center, and the cohort is relatively homogeneous, which may limit generalizability to broader and more heterogeneous clinical populations. All PD participants had established diagnoses; performance in prodromal or very early PD, where biomarkers are most needed, remains to be determined. Future studies should therefore evaluate the tablet-based system in earlier-stage and prodromal populations, where disease-related oculomotor abnormalities may be subtler and where a scalable front-end tool could be especially valuable for longitudinal risk stratification or early specialist referral. More broadly, future validation should extend to larger and more diverse cohorts to determine how well the present findings generalize across demographic, clinical, and disease-stage heterogeneity. The classifier was evaluated using internal cross-validation only, and the modest sample size and relative homogeneity of the cohort further limit the strength of any conclusions about generalizability. External validation in larger, more diverse independent cohorts will therefore be required before any clinical application. Accordingly, the present classification analyses should be interpreted as proof-of-concept evidence that tablet-based oculomotor metrics can support PD discrimination, rather than as a finalized screening model with clinically established operating thresholds. In addition, recordings were obtained under controlled conditions with head stabilization; future work should assess robustness of the tablet-based system under less constrained, more ecologically valid conditions and explore how performance is affected by variations in viewing distance, lighting, and comorbid ocular conditions. Future work should also evaluate whether more deployment-oriented machine-learning approaches can improve performance or calibration in real-world settings, provided that they remain interpretable and robust

to real-world variability. Finally, the present study focused on saccade-derived metrics, which had been the primary target of our prior validation work [30]. Fixation micro-movements and pupil-size metrics may also be relevant to PD and could provide complementary information about visuomotor and cognitive function, but they are currently difficult to estimate robustly from tablet RGB video, particularly given the low spatial resolution, visible-light imaging, and variability introduced by iris pigmentation and illumination. These therefore remain important directions for future work.

Despite these limitations, the present findings support consumer tablets as practical tools for scalable, quantitative oculomotor assessment in PD. The key result is not merely that the iPad approximates benchmark measurements, but that it preserves the PD-related group-level conclusions obtained with the EyeLink 1000 Plus across multiple reflexive and voluntary saccade tasks. The accompanying classification results further suggest that tablet-derived metrics capture subject-level variation in PD-related oculomotor dysfunction that is informative for discrimination, with a stable compact feature set centered on AS DE rate, AS gain, and MGS gain. Together, these findings provide proof-of-concept support for tablet-based PD screening and longitudinal monitoring and motivate further validation in larger and more diverse cohorts, including prodromal and early-stage disease, and under less constrained real-world testing conditions.

Methods

Participants

This study was approved by the Institutional Review Boards at Northwestern University and MIT (protocol number STU00221220), was conducted in accordance with the Declaration of Helsinki, and all participants provided written informed consent prior to participation. Initially, 33 individuals participated; however, two were excluded: one PD participant for repeatedly falling asleep and one HC participant for failing to produce a clean EyeLink gaze signal. Thus, 31 participants were included in the analysis: 12 with PD and 19 HC. For five participants (four HC, one PD), specific task recordings were excluded because EyeLink data were not saved due to technical issues (one participant) or because a clean gaze signal could not be obtained (four participants). Accordingly, one PS, one AS, one MGS, and four SGS task recordings were excluded. In these cases, reflections from participants' glasses and eye makeup were the most likely cause of unreliable EyeLink pupil detection.

Patients with PD were recruited from the Parkinson's Disease & Movement Disorders Center at Northwestern University and examined by a movement disorders neurologist. They were eligible if they: (1) satisfied the UK PD Society Brain Bank diagnostic criteria; (2) had normal or corrected-to-normal vision; (3) showed no eye-related abnormalities (e.g., blepharospasm, diplopia, eyelid-opening apraxia); (4) had no additional neurological disorders; and (5) could understand and complete the eye movement tasks. HC participants met the same vision- and task-related criteria, but reported no history of neurological disorders. Demographic and clinical characteristics are summarized in Table 6.

Experimental Setup and Procedure

The experimental setup and procedure have been described in detail previously [30]. Briefly, participants performed four horizontal eye movement tasks while being recorded simultaneously by an iPad-based system and an EyeLink 1000 Plus (SR Research) operating in remote mode. The EyeLink camera was positioned below the iPad. Participants were seated in a height-adjustable chair with their head stabilized on a chin rest. Viewing distance and device placement matched our prior validation study [30], with the iPad positioned approximately 35 cm from the eyes and the EyeLink camera approximately 55 cm from the eyes.

The iPad served as both the stimulus display and the front-facing RGB camera, recording video at 60 Hz, while the EyeLink recorded binocular gaze at 1000 Hz. Synchronization between the two systems was achieved using SR Research's WebLink. At each display refresh, the iPad app transmitted its current timestamp over a direct

Ethernet connection to the WebLink computer, which relayed it to the EyeLink host computer. This continuous timestamp exchange enabled estimation of clock offset and correction for drift throughout the experiment, yielding sub-millisecond alignment of the iPad and EyeLink data streams. The round-trip communication time of this synchronization framework was below 1 ms [30].

Before each task, both systems were calibrated using targets presented on the iPad screen. The EyeLink was calibrated and validated using WebLink's 5-point calibration and validation procedure for external displays, followed by a 5-point calibration of the iPad. Participants were instructed to fixate on known targets during both procedures. For the iPad, this calibration was used to fine-tune the deep-learning gaze-estimation model to the subject [30]. The mean validation accuracy of the EyeLink system across subjects was 0.75°, ranging from 0.67° to 0.85°, with a standard deviation of 0.05°.

Participants then completed four tasks (Figure 4): (1) Pro-Saccade (PS), (2) Anti-Saccade (AS), (3) Memory-Guided Saccade (MGS), and (4) Self-Generated Saccade (SGS), implemented as in [30]. The PS, AS, and MGS tasks each consisted of 40 trials, whereas the SGS task lasted 30 s. At the viewing distance used, target amplitudes were 17.2° in PS and AS, 34.4° in SGS, and 8.6° (near) and 17.2° (far) in MGS.

Eye Movement Metrics

Eye movement metrics were computed separately for iPad and EyeLink recordings using the same analysis framework described in our previous work [30]. Both systems provided point-of-gaze time series for each task. On the iPad, gaze vectors were estimated using Few-Shot Adaptive Gaze Estimation (FAZE) [32], a deep-learning-based gaze-estimation model that we selected because it is well matched to our acquisition setting. FAZE was trained on GazeCapture [33], a large iPad/iPhone-based dataset that closely mirrors our mobile front-facing camera setup, and it was specifically designed for person-specific adaptation from only a small number of calibration samples, which is important in our protocol because only a limited number of calibration points are available. In addition, FAZE uses a rotation-aware latent representation that explicitly accounts for gaze and head-pose variation, making it well suited to settings with natural head movement. In our experience across laboratory, clinical, and at-home recordings, FAZE-based gaze extraction has also proved relatively robust to sources of variation that are difficult to control in practice, including head movement and lighting variation.

The resulting gaze estimates were mapped to screen coordinates using the iPad camera intrinsics [34] and refined with a per-subject calibration. EyeLink gaze samples were used as provided by the EyeLink system after calibration. Figure 5 shows an example of raw iPad and EyeLink position signals for a typical PS trial. Both signals were then filtered for noise reduction using a method specifically designed for time series containing saccades and fixations [35].

We detected saccade onsets and offsets using the conventional 30°/s velocity threshold [3]. The velocity signal was derived from the filtered position data using the central finite difference method.

We focused on temporal and spatial saccade metrics because our prior validation study [30] showed that these measures can be estimated with sufficient accuracy from the iPad system relative to the EyeLink 1000 Plus. We selected this class of metrics because they have been studied extensively in the PD literature, are biologically motivated, and span reflexive, inhibitory, memory-guided, and internally generated oculomotor control.

Temporal and spatial metrics were derived from primary saccades for each task. In the PS, AS, and MGS tasks, the primary saccade of a trial was defined as the first saccade with amplitude exceeding 2° after peripheral stimulus onset, to avoid small fixational saccades and square-wave jerks [36]. In the SGS task, the primary saccade was defined as the first saccade with amplitude exceeding 2° in the direction opposite to the previous primary saccade.

We computed saccade latency for PS, AS, and MGS. Latency was defined as the interval between primary saccade onset and peripheral stimulus onset for PS and AS, and between primary saccade onset and central fixation offset for MGS. For AS, we computed the DE rate, defined as the fraction of trials in which the primary saccade was directed toward the peripheral stimulus. For MGS, we computed the IE rate,

defined as the fraction of trials in which the primary saccade was executed toward the remembered location within 250 ms of the flash (before the end of the retention period). For SGS, we computed IPSR, defined as the reciprocal of the interval between successive primary saccade onsets. For all tasks, we computed saccade gain as the ratio of primary saccade amplitude to target amplitude.

Saccade-level metrics were averaged within subjects for each device to yield subject-level summary measures.

Trial and Measurement Exclusion Criteria

We applied the same blink- and signal-loss-based exclusion criteria as in our previous validation study [30]. Trials in the PS, AS, and MGS tasks were excluded if blinks or EyeLink signal loss rendered the primary saccade unusable. Blinks were detected using the eye aspect ratio method [37]. For the SGS task, IPSR values immediately following blinks or unusable saccades were excluded.

For the present analysis, we additionally excluded PS, AS, and MGS trials if: (1) latency fell outside the 90–600 ms window (except for IE rate calculation), to remove anticipatory and markedly delayed responses; or (2) the absolute horizontal position of the saccade onset exceeded 5°, indicating poor central fixation. For latency and gain metrics, we included only correct-direction primary saccades. The same automated exclusion rules were applied to iPad and EyeLink data. These criteria were intended to exclude trials in which blinks, signal loss, or poor fixation quality would compromise reliable metric extraction.

Statistical Analysis

For each subject, task, and metric, we computed separate summary values from EyeLink and iPad data. To assess whether the iPad reproduced PD–HC differences, we compared groups using Welch's t-tests for each pre-specified metric and device (two-sided, significance threshold $p < 0.05$). In addition, for all metrics we fit linear mixed-effects models at the trial level with fixed effects of group (PD vs. HC), device (EyeLink vs. iPad), and their interaction, and a random intercept for subject. In this parameterization, the group × device interaction term quantifies the difference between the PD–HC effects estimated from the two devices and was used to assess whether disease-related group effects were consistent across EyeLink and iPad. For these interaction estimates, we report 90% confidence intervals to balance precision with the goal of detecting potentially non-negligible device-related differences relative to benchmark PD–HC effects.

For PD vs. HC classification using iPad-derived metrics, we restricted the analysis to subjects with complete data for all candidate metrics, yielding 26 subjects (15 HC, 11 PD). Candidate features were the prespecified subject-level iPad-derived metrics summarized in Table 1. These candidate features comprised PS latency and gain; AS latency, gain, and DE rate; MGS latency, gain, and IE rate; and SGS gain and IPSR. We prespecified three candidate feature spaces reflecting increasing biological restriction. Our primary model excluded PS-derived and latency-derived metrics because reflexive PS are generally normal or only mildly affected [7] and latency abnormalities may emerge later in the disease course [6], whereas our cohort was early/mid stage. We additionally evaluated two broader sensitivity models: one including all candidate metrics and one excluding PS-derived metrics only.

For each of these three feature spaces, we fit an elastic-net logistic-regression classifier and evaluated performance using nested leave-one-subject-out (LOSO) cross-validation. We chose elastic-net logistic regression for these analyses because the candidate oculomotor metrics were modest in number, biologically interpretable, and correlated, while the sample size was limited. In this setting, elastic net provided coefficient shrinkage to reduce overfitting while retaining a sparse and interpretable linear model. In each outer LOSO fold, one subject was held out for testing. Within the remaining training subjects, inner LOSO was used to select the regularization strength $C \in \{0.001, 0.01, 0.1, 1, 10\}$ and, for elastic-net models, the mixing parameter l_1 ratio $\in \{0.25, 0.5, 0.75\}$. Thus, the inner LOSO loop was used only to tune model hyperparameters within the training subjects, while the outer held-out subject remained completely unseen until final evaluation. The model selected in the inner loop was

then refit using all subjects other than the held-out test subject and used to generate a predicted PD probability for that held-out subject. Aggregating these outer-fold predictions yielded subject-level estimates of the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy at a fixed decision threshold of 0.5. We quantified uncertainty at the subject level via nonparametric bootstrapping over subjects (2,000 resamples) to derive 95% confidence intervals for the main classification performance measures reported below. For the primary classifier, calibration of predicted PD probabilities was assessed using the Brier score, defined as the mean squared difference between predicted probabilities and the observed outcomes (lower values indicate more accurate probabilistic predictions) and a logistic regression of the binary outcome on the logit of the predicted probability to estimate calibration intercept and slope.

To assess model complexity and determine how many features were required for strong discrimination, we performed a separate nested top-k analysis using the full candidate feature space. In each outer LOSO fold, candidate features were ranked using the outer-training subjects only according to their single-feature discriminatory ability, quantified as the absolute deviation of that feature's AUC from chance, $|\mathrm{AUC} - 0.5|$. Using this training-only ranking, we constructed models containing the top 1, top 2, top 3, or top 4 ranked features. For each subset size, a ridge logistic-regression classifier was fit within the training data, with inner-LOSO selection of $C \in \{0.001, 0.01, 0.1, 1, 10\}$, and then evaluated on the held-out subject. We used ridge logistic regression in this analysis because the subset size had already been fixed by the ranking step, and we therefore wanted a stable linear model rather than an additional sparse selector. This analysis was designed to determine whether classification performance plateaued as additional features were added, while avoiding information leakage from the held-out subject into feature ranking.

For the elastic-net models, we additionally examined feature stability across outer folds by recording the standardized regression coefficients from each fold and identifying the features that contributed most consistently across folds. For the top-k analysis, we assessed whether the highest-performing compact feature sets were stable across outer folds.

Although more flexible nonlinear classifiers could potentially improve internally cross-validated discrimination, they would also be more prone to overfitting in the present setting, given the modest sample size and correlated feature space. The goal here was therefore not to maximize predictive performance, but to test whether a compact and biologically interpretable set of tablet-derived metrics preserves disease-relevant signal. We therefore prioritized interpretability and robustness over model complexity.

All statistical analyses and visualizations were implemented in Python using the NumPy, pandas, statsmodels, scikit-learn, and Matplotlib libraries.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to participant privacy considerations but deidentified derived saccade metrics are available from the corresponding author on reasonable request.

Code Availability

The code is not publicly available but may be made available to qualified researchers upon reasonable request.

Acknowledgments

This study was funded, in part, by Analog Devices, Inc. through a partnership with the MIT Medical Electronic Device Realization Center (grant number: Not applicable), and by the MIT Aging Brain Initiative (grant number: Not applicable), the Northwestern University Department of Physical Therapy and Human Movement Sciences (grant number: Not applicable), and the Northwestern Medicine Enterprise Data

Warehouse (grant number: Not applicable). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author Contributions

J.K., T.H., and F.J.D. conceived and organized the research project. J.K., E.Z., J.A.K., C.P., R.G.M., and L.V.M. executed the research project. C.G.S., V.S., T.H., and F.J.D. supervised the research project. J.K., T.H., and F.J.D. designed the statistical analysis; J.K. executed it; J.A.K., C.P., R.G.M., L.V.M., C.G.S., V.S., T.H., and F.J.D. reviewed and critiqued the statistical analysis. J.K. wrote the first draft of the manuscript. All authors reviewed, critiqued, and approved the final version of the manuscript.

Competing Interests

The authors declare no competing financial or non-financial interests.

References

[1] Dorsey, E.R., Sherer, T., Okun, M.S., Bloem, B.R.: The emerging evidence of the Parkinson pandemic. Journal of Parkinson's disease **8**(s1), 3–8 (2018)

[2] Steinmetz, J.D., Seeher, K.M., Schiess, N., Nichols, E., Cao, B., Servili, C., Cavallera, V., Cousin, E., Hagins, H., Moberg, M.E., *et al.*: Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. The Lancet Neurology **23**(4), 344–381 (2024)

[3] Leigh, R.J., Zee, D.S.: The Neurology of Eye Movements. Oxford University Press, Oxford (2015). Chap. 4

[4] Pretegiani, E., Optican, L.M.: Eye movements in parkinson's disease and inherited parkinsonian syndromes. Frontiers in neurology **8**, 592 (2017)

[5] Antoniades, C.A., Spering, M.: Eye movements in parkinson's disease: from neurophysiological mechanisms to diagnostic tools. Trends in Neurosciences **47**(1), 71–83 (2024)

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

[7] Bronstein, A.M., Anderson, T., Kaski, D., MacAskill, M., Shaikh, A.: In: Gálvez-Jiménez, N., Korczyn, A.D., Lugo-Sanchez, R. (eds.) Oculomotor and Visual-Vestibular Disturbances in Parkinson's Disease, pp. 115–129. Cambridge University Press, Cambridge (2022)

[8] Spitzer, L., Mueller, S.: Using a test battery to compare three remote, video-based eye-trackers. In: 2022 Symposium on Eye Tracking Research and Applications, pp. 1–7 (2022)

[9] Ehinger, B.V., Groß, K., Ibs, I., König, P.: A new comprehensive eye-tracking test battery concurrently evaluating the pupil labs glasses and the eyelink 1000. PeerJ **7**, 7086 (2019)

[10] Aziz, S., Lohr, D.J., Komogortsev, O.: Synchroneyes: A novel, paired data set of eye movements recorded simultaneously with remote and wearable eye-tracking devices. In: 2022 Symposium on Eye Tracking Research and Applications, pp. 1–6 (2022)

[11] Weil, R.S., Schrag, A.E., Warren, J.D., Crutch, S.J., Lees, A.J., Morris, H.R.: Visual dysfunction in parkinson's disease. Brain **139**(11), 2827–2843 (2016)

[12] Pinkhardt, E.H., Kassubek, J.: Ocular motor abnormalities in parkinsonian syndromes. Parkinsonism & related disorders **17**(4), 223–230 (2011)

[13] Tsitsi, P., Benfatto, M.N., Seimyr, G.Ö., Larsson, O., Svenningsson, P., Markaki, I.: Fixation duration and pupil size as diagnostic tools in parkinson's disease. Journal of Parkinson's Disease **11**(2), 865–875 (2021)

[14] Tolosa, E., Garrido, A., Scholz, S.W., Poewe, W.: Challenges in the diagnosis of parkinson's disease. The Lancet Neurology **20**(5), 385–397 (2021)

[15] Zarkali, A., Thomas, G.E., Zetterberg, H., Weil, R.S.: Neuroimaging and fluid biomarkers in parkinson's disease in an era of targeted interventions. Nature Communications **15**(1), 5661 (2024)

[16] Ellmerer, P., Peball, M., Carbone, F., Ritter, M., Heim, B., Marini, K., Valent, D., Krismer, F., Poewe, W., Djamshidian, A., *et al.*: Eye tracking in patients with parkinson's disease treated with nabilone–results of a phase ii, placebo-controlled, double-blind, parallel-group pilot study. Brain Sciences **12**(5), 661 (2022)

[17] Yang, X., Krajbich, I.: Webcam-based online eye-tracking for behavioral research. Judgment and Decision making **16**(6), 1485–1505 (2021)

[18] Crutcher, M.D., Calhoun-Haney, R., Manzanares, C.M., Lah, J.J., Levey, A.I., Zola, S.M.: Eye tracking during a visual paired comparison task as a predictor of early dementia. American Journal of Alzheimer's Disease & Other Dementias® **24**(3), 258–266 (2009)

[19] Niehorster, D.C., Cornelissen, T.H., Holmqvist, K., Hooge, I.T., Hessels, R.S.: What to expect from your remote eye-tracker when participants are unrestrained. Behavior research methods **50**(1), 213–227 (2018)

[20] Culicetto, L., Cardile, D., Marafioti, G., Lo Buono, V., Ferraioli, F., Massimino, S., Di Lorenzo, G., Sorbera, C., Brigandì, A., Vicario, C.M., *et al.*: Recent advances (2022–2024) in eye-tracking for parkinson's disease: a promising tool for diagnosing and monitoring symptoms. Frontiers in Aging Neuroscience **17**, 1534073 (2025)

[21] Diotaiuti, P., Marotta, G., Di Siena, F., Vitiello, S., Di Prinzio, F., Rodio, A., Di Libero, T., Falese, L., Mancone, S.: Eye tracking in parkinson's disease: a review of oculomotor markers and clinical applications. Brain Sciences **15**(4), 362 (2025)

[22] Farrell, K., MacDougall, D.: An overview of clinical applications of virtual and augmented reality. Canadian Journal of Health Technologies **3**(3) (2023)

[23] Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., *et al.*: Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nature communications **11**(1), 4553 (2020)

[24] Kim, N.Y., He, J., Wu, Q., Dai, N., Kohlhoff, K., Turner, J., Paul, L.K., Kennedy, D.P., Adolphs, R., Navalpakkam, V.: Smartphone-based gaze estimation for in-home autism research. Autism Research **17**(6), 1140–1148 (2024)

[25] Clough, M., Bartholomew, J., White, O., Fielding, J.: Investigating the utility of the braineye smartphone eye tracking application and platform in concussion management. Sports Medicine-Open **11**(1), 24 (2025)

[26] Lai, H.-Y., Saavedra-Peña, G., Sodini, C.G., Heldt, T., Sze, V.: App-based saccade latency and directional error determination across the adult age spectrum. IEEE Transactions on Biomedical Engineering **69**(2), 1029–1039 (2021)

[27] Yoo, J.H., Kang, C., Lim, J.S., Wang, B., Choi, C.-H., Hwang, H., Han, D.H.,

Kim, H., Cheon, H., Kim, J.-W.: Development of an innovative approach using portable eye tracking to assist adhd screening: a machine learning study. Frontiers in Psychiatry **15**, 1337595 (2024)

[28] Villers-Sidani, É., Voss, P., Guitton, D., Cisneros-Franco, J.M., Koch, N.A., Ducharme, S.: A novel tablet-based software for the acquisition and analysis of gaze and eye movement parameters: a preliminary validation study in parkinson's disease. Frontiers in Neurology **14**, 1204733 (2023)

[29] Koch, N.A., Voss, P., Cisneros-Franco, J.M., Drouin-Picaro, A., Tounkara, F., Ducharme, S., Guitton, D., Villers-Sidani, É.: Eye movement function captured via an electronic tablet informs on cognition and disease severity in parkinson's disease. Scientific Reports **14**(1), 9082 (2024)

[30] Koerner, J., Zou, E., Karl, J.A., Poon, C., Verhagen Metman, L., Sodini, C.G., Sze, V., David, F.J., Heldt, T.: Towards scalable screening for the early detection of parkinson's disease: validation of an ipad-based eye movement assessment system against a clinical-grade eye tracker. npj Parkinson's Disease **11**(1), 233 (2025)

[31] Waldthaler, J., Stock, L., Student, J., Sommerkorn, J., Dowiasch, S., Timmermann, L.: Antisaccades in parkinson's disease: a meta-analysis. Neuropsychology Review **31**, 628–642 (2021)

[32] Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9368–9377 (2019)

[33] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)

[34] Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). Chap. 6. Camera Models

[35] Dai, W., Selesnick, I., Rizzo, J.-R., Rucker, J., Hudson, T.: Detection of normal and slow saccades using implicit piecewise polynomial approximation. Journal of vision **21**(6), 8–8 (2021)

[36] Munoz, M.J., Reilly, J.L., Pal, G.D., Metman, L.V., Rivera, Y.M., Drane, Q.H., Corcos, D.M., David, F.J., Goelz, L.C.: Medication adversely impacts visually-guided eye movements in parkinson's disease. Clinical Neurophysiology **143**, 145–153 (2022)

[37] Soukupová, T., Čech, J.: Real-time eye blink detection using facial landmarks. In: 21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia, pp. 1–8 (2016)

Figures

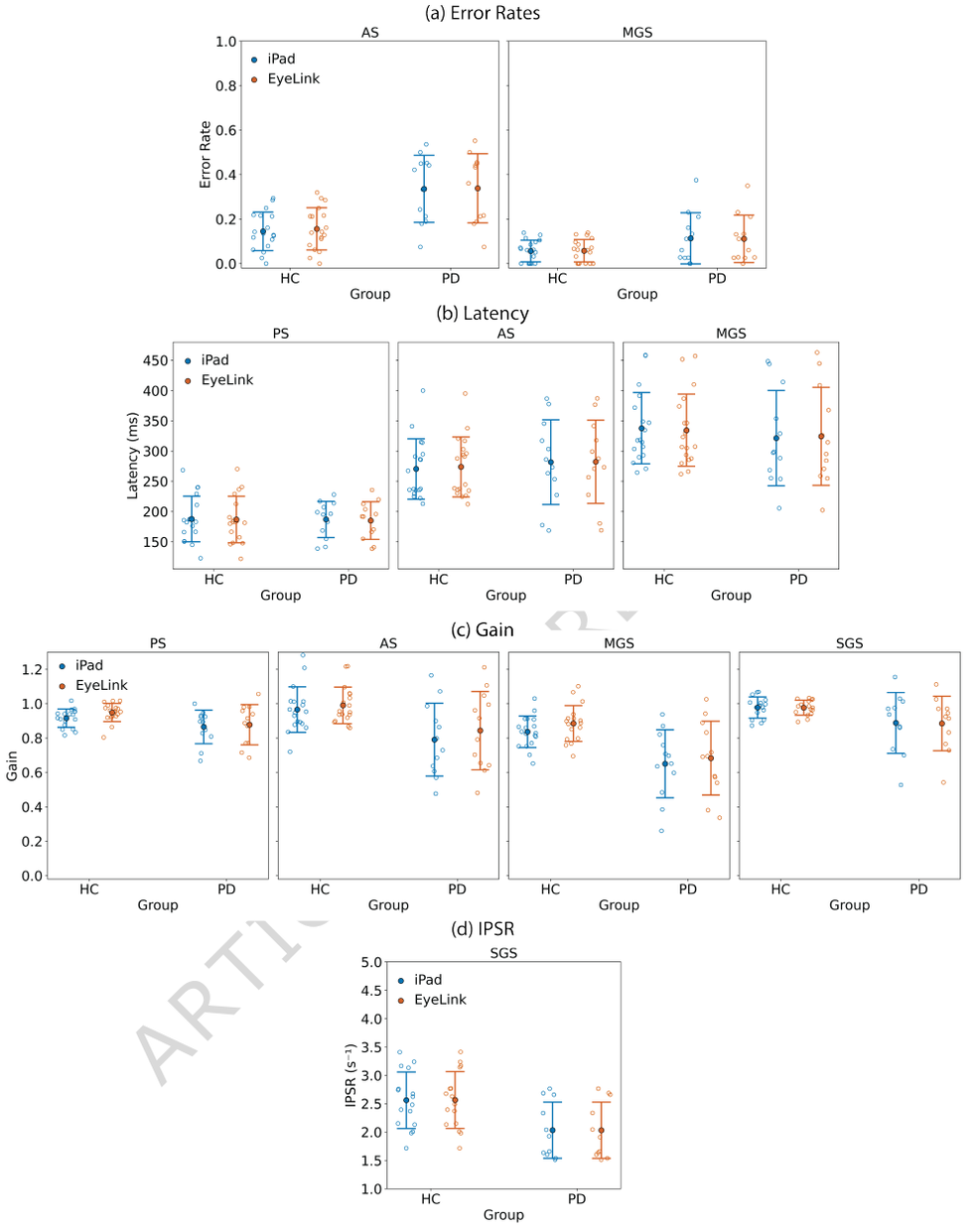

Fig. 1 Distributions of eye movement metrics for healthy control (HC) participants and participants with Parkinson's disease (PD) measured with the iPad and EyeLink. PS: pro-saccade; AS: anti-saccade; MGS: memory-guided saccade; SGS: self-generated saccade; IPSR: instantaneous primary saccade rate. (a) AS and MGS error rates; AS error reflects directional errors, and MGS error reflects inhibitory errors (early responses within 250 ms of the flash). (b) Saccade latency across PS, AS, and MGS tasks. (c) Saccade gain across PS, AS, MGS, and SGS tasks. (d) IPSR in the SGS task. Each point represents a subject-level mean; solid markers and vertical lines indicate the group mean ± SD for each Group × Device combination. iPad and EyeLink are distinguished using a blue–orange palette.

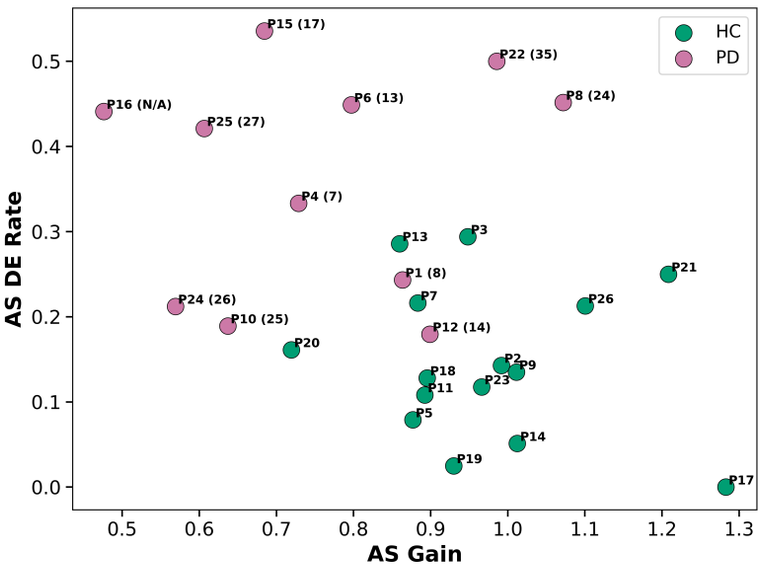

Fig. 2 Subject-level means for anti-saccade (AS) directional error (DE) rate (y-axis) versus AS gain (x-axis), computed from the iPad recordings. Points are colored by group, with healthy control (HC) participants shown in green and participants with Parkinson's disease (PD) shown in magenta. For participants with PD, the MDS-UPDRS Part III motor score is shown in brackets.

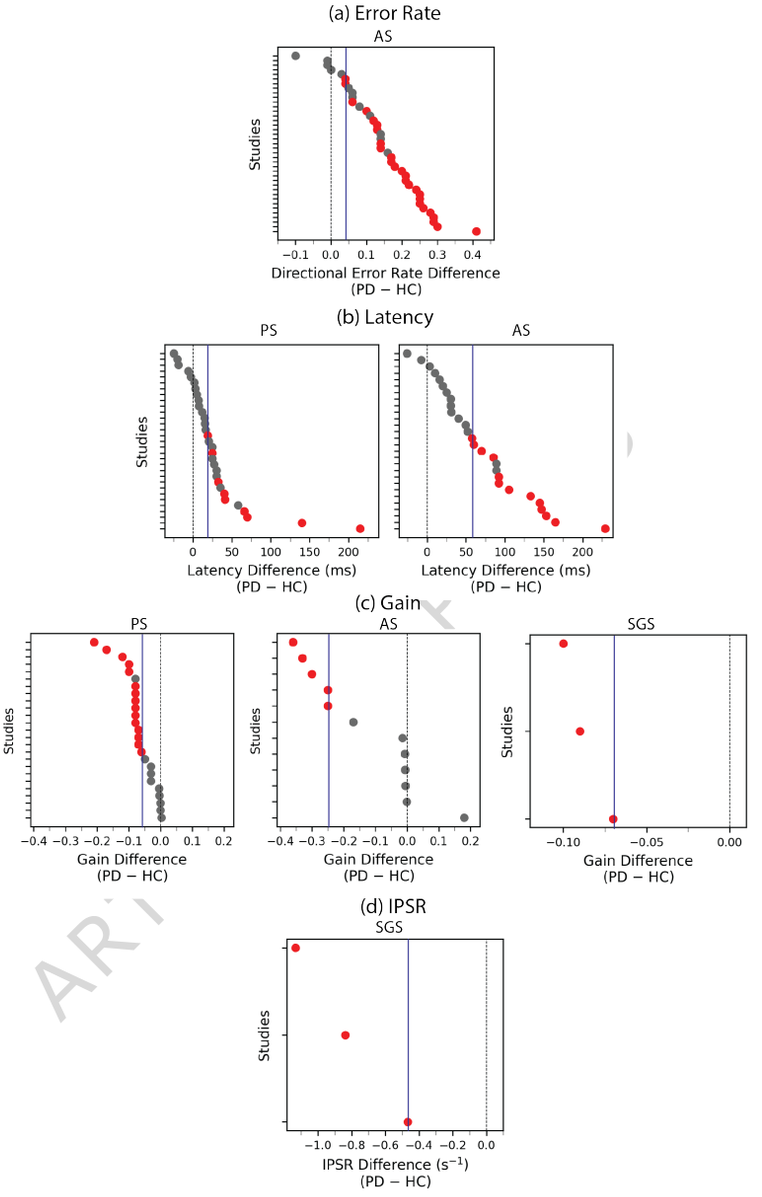

Fig. 3 Expected saccade changes in Parkinson's disease (PD) relative to healthy controls (HC) based on prior literature. AS: anti-saccade; PS: pro-saccade; SGS: self-generated saccade; DE: directional error; IPSR: instantaneous primary saccade rate. Each panel shows PD–HC differences reported in previous studies for (a) AS DE rate, (b) PS and AS latency, (c) PS, AS, and SGS gain, and (d) SGS IPSR. Points correspond to individual studies or cohorts. Values were obtained from prior work as follows: for AS DE, from the AS meta-analysis in [31]; for PS and AS latency, from our previous study [30], where PD–HC latency differences were derived from a curated subset of estimates in that meta-analysis; for PS and AS gain, from the same previous work, in which we re-examined the original studies included in the meta-analysis, extracted a subset of amplitude and gain effects, harmonized them on an amplitude scale, and re-expressed them here as gain; and for SGS gain and SGS IPSR, from our earlier compilation of SGS studies in PD in [30]. Red indicates a statistically significant difference ($p < 0.05$). The vertical dashed line marks zero PD–HC difference, and the vertical solid blue line indicates, for each metric, the smallest statistically significant PD–HC difference used here as the benchmark when interpreting the mixed-effects iPad–EyeLink agreement analyses.

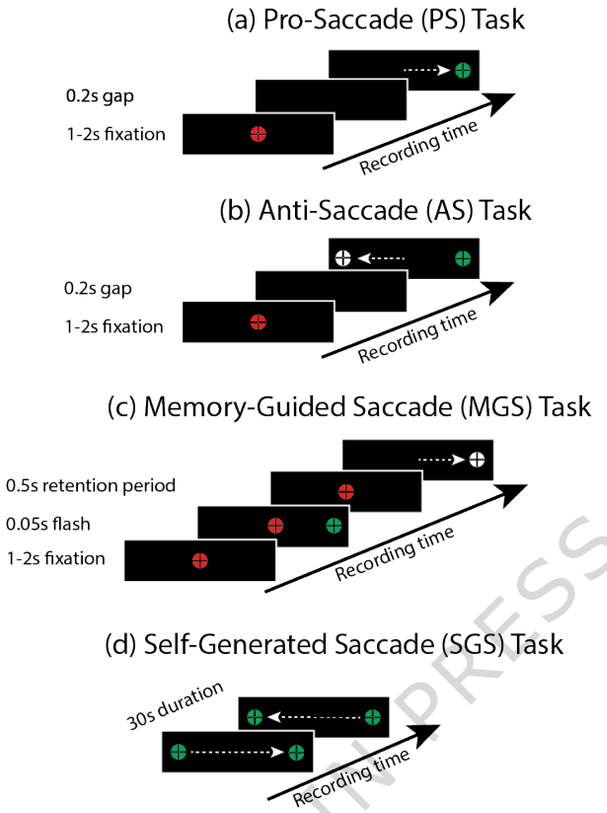

Fig. 4 The four eye movement tasks considered in this study, along with their timing parameters. PS: pro-saccade; AS: anti-saccade; MGS: memory-guided saccade; SGS: self-generated saccade. In the PS, AS, and MGS tasks, the central fixation target is shown in red, the peripheral stimulus in green, and the saccade target is shown in white when it differs from the peripheral stimulus (AS and MGS tasks). In the SGS task, the two continuously visible saccade targets are shown in green. The correct saccade direction is represented by the white arrow.

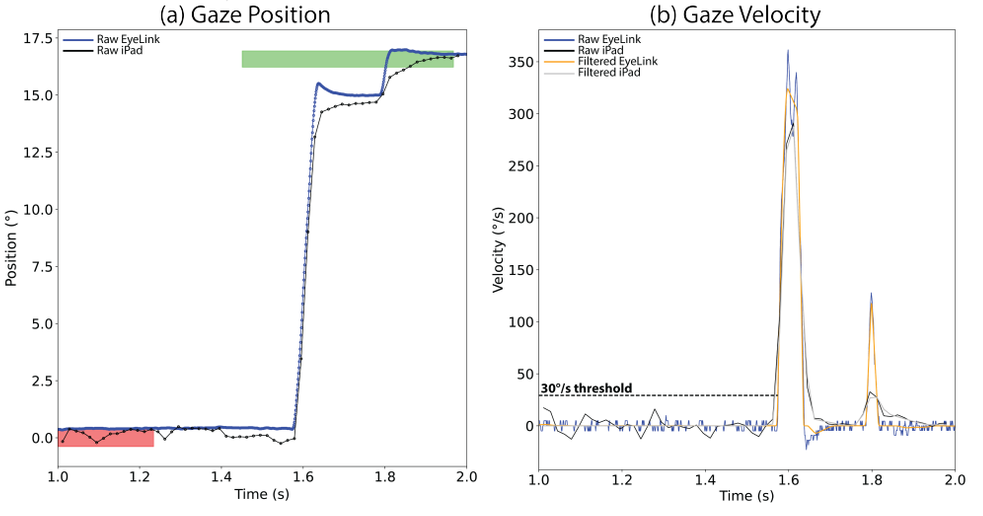

Fig. 5 Gaze position (left) and the corresponding velocity signals (right) produced by the iPad and EyeLink for a PS trial. The red and green horizontal bars represent the location and duration of the central fixation target and peripheral stimulus, respectively. Saccades are detected using a 30°/s velocity threshold.

Tables

Metric	Device	HC Mean	PD Mean	Δ (HC − PD)	p
AS DE Rate	EyeLink	0.156	0.338	-0.182	**0.002**
	iPad	0.145	0.336	-0.191	**0.001**
AS Gain	EyeLink	0.990	0.843	0.147	**0.048**
	iPad	0.966	0.790	0.175	**0.021**
AS Latency (ms)	EyeLink	273.647	282.219	-8.572	0.714
	iPad	270.284	281.617	-11.334	0.634
MGS Gain	EyeLink	0.884	0.683	0.202	**0.009**
	iPad	0.836	0.650	0.186	**0.009**
MGS IE Rate	EyeLink	0.196	0.262	-0.066	0.331
	iPad	0.196	0.268	-0.073	0.294
MGS Latency (ms)	EyeLink	334.326	324.162	10.164	0.714
	iPad	337.695	321.305	16.389	0.546
PS Gain	EyeLink	0.948	0.877	0.071	0.070
	iPad	0.915	0.864	0.051	0.122
PS Latency (ms)	EyeLink	186.808	185.078	1.730	0.893
	iPad	187.621	186.877	0.743	0.953
SGS Gain	EyeLink	0.976	0.884	0.092	0.087
	iPad	0.977	0.888	0.090	0.134
SGS IPSR (s^{-1})	EyeLink	2.566	2.031	0.534	**0.012**
	iPad	2.561	2.032	0.529	**0.013**

Table 1 Group-wise comparison of eye movement metrics across tasks and devices. HC: healthy control; PD: Parkinson's disease; PS: pro-saccade; AS: anti-saccade; MGS: memory-guided saccade; SGS: self-generated saccade; DE: directional error; IE: inhibitory error. Mean values for HC and PD groups are shown along with the difference (HC–PD) and associated p-values. Statistically significant p-values ($p < 0.05$) are shown in bold.

Task	Metric	Estimate	90% CI
PS	Latency (ms)	-1.126	[−10.104, 7.851]
AS	Latency (ms)	-1.690	[−12.282, 8.901]
MGS	Latency (ms)	7.030	[−9.682, 23.742]
PS	Gain	-0.022	[−0.043, −0.001]
AS	Gain	0.032	[−0.001, 0.065]
MGS	Gain	0.001	[−0.031, 0.033]
SGS	Gain	0.005	[−0.010, 0.020]
AS	DE Rate	-0.007	[−0.063, 0.050]
MGS	IE Rate	-0.004	[−0.043, 0.035]
SGS	IPSR (s^{-1})	-0.005	[−0.073, 0.063]

Table 2 Group × device interaction estimates from linear mixed-effects models. HC: healthy control; PD: Parkinson's disease; PS: pro-saccade; AS: anti-saccade; MGS: memory-guided saccade; SGS: self-generated saccade; DE: directional error; IE: inhibitory error. Positive values indicate a larger PD–HC effect for EyeLink than for the iPad.

Quantity	**Primary (no PS, no latency)**	**All features**	**No PS**
AUC	0.897 [0.700, 1.000]	0.939 [0.803, 1.000]	0.927 [0.762, 1.000]
Sensitivity	0.727 [0.417, 1.000]	0.818 [0.583, 1.000]	0.818 [0.545, 1.000]
Specificity	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.867 [0.667, 1.000]
Accuracy	0.885 [0.731, 1.000]	0.923 [0.808, 1.000]	0.846 [0.692, 0.962]

Table 3 Performance of the three prespecified elastic-net classifiers for Parkinson's disease (PD) versus healthy control (HC) classification using iPad-derived metrics. AUC: area under the receiver operating characteristic curve; PS: pro-saccade. The primary biologically constrained model excluded PS- and latency-derived features, the all-features model included all candidate metrics, and the no-PS model excluded PS-derived features only. All models were evaluated using nested leave-one-subject-out cross-validation in 26 subjects (15 HC, 11 PD). Values in brackets denote 95% confidence intervals obtained via nonparametric bootstrapping over subjects (2,000 resamples). A fixed decision threshold of 0.5 was used for sensitivity, specificity, and accuracy.

Quantity	**Estimate**
Brier score	0.115
Calibration intercept	-0.011
Calibration slope	0.927

Table 4 Calibration metrics for the primary biologically constrained elastic-net classifier for Parkinson's disease (PD) versus healthy control (HC) classification using iPad-derived metrics. PS: pro-saccade. This model excluded PS-derived and latency-derived features and was evaluated using nested leave-one-subject-out cross-validation in 26 subjects (15 HC, 11 PD).

	k=1	k=2	k=3	k=4
Metrics	AS DE	AS DE, AS gain	AS DE, AS gain, MGS gain	AS DE, AS gain, MGS gain, SGS IPSR
AUC	0.703	0.824	**0.976** ([**0.895, 1.000**])	0.958
Accuracy	0.654	0.731	**0.962** ([**0.885, 1.000**])	0.885
Sensitivity	0.455	0.727	**0.909** ([**0.556, 1.000**])	0.727
Specificity	0.800	0.733	**1.000** ([**1.000, 1.000**])	1.000

Table 5 Performance of nested top-k models using the full candidate iPad-derived feature space. In each outer leave-one-subject-out (LOSO) fold, candidate features were ranked using the training subjects only according to their single-feature discriminatory ability, quantified as $|\mathrm{AUC} - 0.5|$. Ridge logistic regression was then fit using the top 1, top 2, top 3, or top 4 ranked features, with inner-LOSO selection of the regularization parameter. The table reports the aggregated held-out performance across 26 subjects (15 healthy controls, 11 participants with Parkinson's disease). For the best-performing top-3 model, 95% confidence intervals obtained via nonparametric bootstrapping over subjects (2,000 resamples) are shown in parentheses. The listed $k = 4$ feature set is the most frequent top-4 set across outer folds. AS: anti-saccade; MGS: memory-guided saccade; SGS: self-generated saccade; DE: directional error; IPSR: instantaneous primary saccade rate; AUC: area under the receiver operating characteristic curve.

	HC	PD
Sex (M/F)	6/13	10/2
Age (mean ± SD, years)	59.3 ± 12.6	61.4 ± 4.9
MDS-UPDRS Part III (mean ± SD)	N/A	20.2 ± 8.9
Disease Duration (mean ± SD, years)	N/A	4.6 ± 5.8
Wearing Glasses (Y/N)	1/18	3/9

Table 6 Subject demographics and clinical characteristics. HC: healthy control; PD: Parkinson's disease; MDS-UPDRS Part III: Motor Examination component of the Movement Disorders Society–Unified Parkinson's Disease Rating Scale. Disease duration is defined as the time (in years) between diagnosis and the date of the experiment.