JAMA Ophthalmology | Original Investigation

Neurologic Dysfunction Assessment in Parkinson Disease Based on Fundus Photographs Using Deep Learning

Sangil Ahn, BS; Jitae Shin, PhD; Su Jeong Song, MD, PhD; Won Tae Yoon, MD, PhD; Min Sagong, MD, PhD; Areum Jeong, MD, PhD; Joon Hyung Kim, MD; Hyeong Gon Yu, MD, PhD

IMPORTANCE Until now, other than complex neurologic tests, there have been no readily accessible and reliable indicators of neurologic dysfunction among patients with Parkinson disease (PD). This study was conducted to determine the role of fundus photography as a noninvasive and readily available tool for assessing neurologic dysfunction among patients with PD using deep learning methods.

OBJECTIVE To develop an algorithm that can predict Hoehn and Yahr (H-Y) scale and Unified Parkinson's Disease Rating Scale part III (UPDRS-III) score using fundus photography among patients with PD.

DESIGN, SETTINGS, AND PARTICIPANTS This was a prospective decision analytical model conducted at a single tertiary-care hospital. The fundus photographs of participants with PD and participants with non-PD atypical motor abnormalities who visited the neurology department of Kangbuk Samsung Hospital from October 7, 2020, to April 30, 2021, were analyzed in this study. A convolutional neural network was developed to predict both the H-Y scale and UPDRS-III score based on fundus photography findings and participants' demographic characteristics.

MAIN OUTCOMES AND MEASURES The area under the receiver operating characteristic curve (AUROC) was calculated for sensitivity and specificity analyses for both the internal and external validation data sets.

RESULTS A total of 615 participants were included in the study: 266 had PD (43.3%; mean [SD] age, 70.8 [8.3] years; 134 male individuals [50.4%]), and 349 had non-PD atypical motor abnormalities (56.7%; mean [SD] age, 70.7 [7.9] years; 236 female individuals [67.6%]). For the internal validation data set, the sensitivity was 83.23% (95% CI, 82.07%-84.38%) and 82.61% (95% CI, 81.38%-83.83%) for the H-Y scale and UPDRS-III score, respectively. The specificity was 66.81% (95% CI, 64.97%-68.65%) and 65.75% (95% CI, 62.56%-68.94%) for the H-Y scale and UPDRS-III score, respectively. For the external validation data set, the sensitivity and specificity were 70.73% (95% CI, 66.30%-75.16%) and 66.66% (95% CI, 50.76%-82.25%), respectively. Lastly, the calculated AUROC and accuracy were 0.67 (95% CI, 0.55-0.79) and 70.45% (95% CI, 66.85%-74.04%), respectively.

CONCLUSIONS AND RELEVANCE This decision analytical model reveals amalgamative insights into the neurologic dysfunction among PD patients by providing information on how to apply a deep learning method to evaluate the association between the retina and brain. Study data may help clarify recent research findings regarding dopamine pathologic cascades between the retina and brain among patients with PD; however, further research is needed to expand the clinical implication of this algorithm.

+ Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: Won Tae Yoon, MD, PhD, Department of Neurology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, 29 Saemunan-Ro, Jongno-Gu, Seoul 03181, South Korea (wtyoon@gmail.com); Jitae Shin, PhD, Department of Electrical and Computer Engineering, Sungkyunkwan University, 2066, Seobu-Ro, Jangan-Gu, Suwon 16419, South Korea (jtshin@skku.edu).

P arkinson disease (PD) is the second most common neurodegenerative disorder, and, owing to the aging population, the number of patients with PD is expected to double by 2030.[1] Projected PD prevalence will be more than 1.6 million in the US with projected total economic burden surpassing $79 billion by 2037.[2] The diagnosis of PD is based on the clinical manifestations and findings from complex neuroimaging examinations, such as magnetic resonance imaging (MRI) or magnetic resonance angiography (MRA), which contribute to medical burden among patients with PD. Therefore, besides the established diagnostic method, many studies have been conducted to investigate the potential metrics for PD, translating these into clinically practical tools for the diagnosis and treatment of patients with PD.

The unique features of the retina make it a useful assessment tool for various systemic diseases. First, the retina is the only organ in which blood vessels can be examined noninvasively. Recently, in addition to predicting biometric data associated with cardiovascular risks, attempts to assess the direct cardiovascular disease risk by predicting the coronary calcium score based on fundus photographs have been made by several studies; these attempts have yielded encouraging results.[3-7] Cho et al[7] reported predicting coronary calcium score based on the fundus photographs with other biometrics showing an area under the receiver operating characteristic curve (AUROC) of 0.757 (95% CI, 0.73-0.79). Second, due to a common embryonic origin, the brain and retina share many biological features that play a prominent role in the pathogenesis of major neurodegenerative diseases, including PD.[8-18] Recent studies have reported that PD progression is associated with the structural changes in the retinal nerve fiber layer and suggest the potential role of the retina as a modality for assessing PD progression.[16-21] Recently, deep learning algorithms have been used in several medical fields. Even though deep learning algorithms are considered black box learning methods (ie, the machine learning model provides a result without explaining how that result was derived), these algorithms permit clinicians to seek new associations and findings that are beyond the extent of traditional analytic methods and human participants.[1-5] The application of various deep learning methods to predict cardiovascular disease risks or systemic biometric values using fundus photography is already well known.[6,7] Given the strong anatomic association between the retina and PD, it is natural for researchers of PD to study the possible role of fundus photographs using the Hoehn and Yahr (H-Y) scale and Unified Parkinson's Disease Rating Scale part III (UPDRS-III) evaluation metric. Therefore, in this prospective study, we developed a deep learning algorithm to predict both the H-Y scale and UPDRS-III score using fundus photography among patients with PD. Moreover, heat maps were evaluated in the macular area to confirm the anatomic association of our algorithm's function. Furthermore, external validation for the assessment function of this algorithm was conducted using 2 independent hospital data sets to test the function of our algorithm in different clinical settings.

Key Points

Question Is a deep learning system capable of performing neurologic dysfunction assessment for patients with Parkinson disease using fundus images?

Findings This decision analytical model including a total of 539 fundus images from 266 participants diagnosed with PD revealed an algorithm with high sensitivity above 80% for both the Hoehn and Yahr and Unified Parkinson's Disease Rating Scale part III scores assessment but a lower specificity ranging from 66% to 67%.

Meaning Study results suggest the potential ability of deep learning analysis of fundus photographs with demographic characteristics to assess neurologic dysfunction in patients with Parkinson disease.
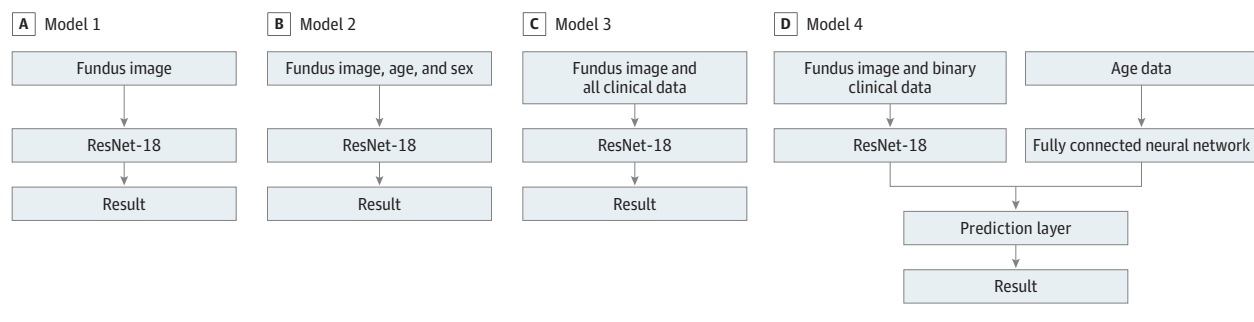
Methods

Study Design and Participants

This decision analytical model, conducted from May 1 to September 30, 2021, adheres to the tenets of the Declaration of Helsinki. The study protocol was reviewed and approved by the institutional review board of Kangbuk Samsung Hospital. Except for the fundus photographs used for external validation, all participants' written informed consent was acquired for this study. This study followed the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) reporting guidelines.

For external validation, we used the fundus photographs of patients with PD who visited the ophthalmology department of Seoul National University Hospital and Yeungnam University Hospital. Patients with PD were recruited from the Movement Disorders Clinic of the neurology department of Kangbuk Samsung Hospital from October 7, 2020, to April 30, 2021. A movement disorders specialist (W.T.Y.) determined the PD diagnosis based on the clinical diagnostic criteria of the UK PD Brain Bank Society's brain imaging study, which used brain MRI and MRA and fluoro-propyl-carbomethoxy-iodophenyl-tropane-ligand positron emission tomography findings. Fundus photographs were collected for the control group, which consisted of health screening participants from the same study period. Fundus photographs were taken with nonmydriatic fundus cameras, including EIDON (Centervue), TRC-NW300, TRC-50IX, TRC-NW200, and TRC-NW8 (Topcon).

Information on demographic characteristics included age, sex, age at PD onset, diabetes (types 1 and 2), hypertension, and cardiovascular diseases. In this study, the H-Y scale and UPDRS-III score were assessed to diagnose and assess the neurologic motor function of participants according to the criteria described by the Movement Disorders Society.[22,23] The participants who refused to undergo fundus photography or who had ungradable fundus photographs were excluded from the study. The ungradable fundus photographs were defined as follows: (1) images with

Figure 1. Four Types of Deep Learning Systems

baseline network based on a CNN and a fully-connected neural network. The purpose of this network was to improve the ability to predict PD severity by achieving synergy using both fundus images and clinical data (**Figure 1**).

Statistical Analysis

Among the clinical data, the *P* value for age was obtained by comparing the 2 groups (PD group vs non-PD group) through a *t* test, whereas the *P* values of sex, diabetes, and hypertension were obtained through a Pearson χ^2 test. A 2-sided *P* value < .05 was considered statistically significant. To determine the performance of our proposal, 4 metrics (sensitivity, specificity, AUROC, and accuracy) were calculated in the validation data set for the classification of both H-Y scale and UPDRS-III score.[26] Moreover, the sensitivity, specificity, AUROC, and accuracy were calculated in the external testing data set for the classification of H-Y scale to show the accuracy of our proposal for the new sample data. The 4 metrics are widely used statistics for explaining a diagnostic test. In particular, the proposed method is used to predict the presence or absence of PD, how well the classes of H-Y scale and UPDRS-III score are predicted, and to quantify how reliable the experiment is. Bootstrapping was used to estimate the 95% CIs of the performance metrics with the participants as the resampling unit. All analyses were performed using Python, version 3.6 with Scikit-learn, version 0.24.2 and SciPy, version 1.5.4 library metrics.[26]

We trained 5 well-known CNNs (ResNet-18, ResNet-101, VGG19, EfficientNet-b0, and EfficientNEt-b7) with binary cross-entropy loss function to verify that the normal and PD participant groups could be accurately distinguished by using the sigmoid function in the last layer of each model. We set the batch size to 32 and used the Adam optimizer; the learning rate was set to 1e-4 with 150 epochs. After selecting the ResNet-18 as the baseline CNN from which to build a multimodal network (because our goal was to predict the specific class of both the H-Y scale and UPDRS-III score), we used the cross-entropy loss function for our multimodal network with the softmax function in the last layer. We reset the learning rate and epochs into 1e-5 and 500 epochs, respectively. We used the implanted Pytorch framework, version 1.3 with NVIDIA GeForce TITAN XP 3 parallel graphics processing units.

third-generation branches not identifiable within 1 optic disc diameter of the macular area; (2) images with various artifacts; (3) images in which at least 1 of the following was incomplete: macula, optic disc, superior temporal arcade, or inferior temporal arcade; or (4) images in which the diagnosis could not be obtained due to various degradations.[24]

The enrolled participants were divided into the following 3 groups based on the H-Y scale: a control group with an H-Y scale of 0; an early-stage PD group with an H-Y scale of 1; and a moderate or advanced-stage PD group with an H-Y scale of 2 or greater. Moreover, the enrolled participants were classified into the following 3 groups according to the UPDRS-III score: a control group with a UPDRS-III score of 0; a group with a UPDRS-III score of 1 to 14; and a group with a UPDRS-III score of 15 or greater.

Data Preprocessing

To evaluate the participants' stage based on the deep learning system with fundus images, we applied general image-augmentation techniques, which include random rotation and random flip, to train the deep learning system. To prevent a possible problem with overfitting due to a small amount of data in the deep learning system, we additionally proposed a color-augmentation technique that changes the background color of the fundus images that can be observed in the actual fundus image (eFigure 1 in Supplement 1). Eventually, we generated the fundus images with various background colors, increasing the number of fundus images that could be used in the deep learning system.

Deep Learning System

Deep learning is a machine learning technique that consists of multilayered artificial neural networks (ANNs). Among the various types of ANNs, a convolutional neural network (CNN) extracts features from images, which are input values, using a convolutional filter. A CNN is particularly useful in using data to classify images by recognizing similar patterns. In other words, manually extracting features is not necessary. Because of these advantages, CNN is widely used in areas requiring object recognition or computer vision to find meaningful features from medical images.[13-16,25]

To establish a more accurate and discriminatory algorithm, we conducted a multimodal network composed of a

Table 1. Study Participants' Clinical Characteristics

Characteristics	Mean (SD)		P value
	Non-Parkinson (n = 349)	Parkinson (n = 266)	
Age	70.7 (7.9)	70.8 (8.3)	.82
Sex, No. (%)			
Male	113 (32.4)	134 (50.4)	<.001
Female	236 (67.6)	132 (49.6)	
Diabetes, No. (%)	66 (18.9)	62 (23.3)	.18
Hypertension, No. (%)	142 (40.7)	91 (34.2)	.10
H-Y scale	0 (0)	1.85 (0.7)	<.001
UPDRS-III score	0 (0)	26.58 (12.5)	<.001

Abbreviations: H-Y, Hoehn and Yahr; UPDRS, Unified Parkinson's Disease Rating Scale.

Results

Characteristics of the Data Set

A total of 615 participants were included in the study: 266 had PD (43.3%; mean [SD] age, 70.8 [8.3] years; 134 male individuals [50.4%]; 132 female individuals [49.6%]), and 349 had non-PD atypical motor abnormalities (56.7%; mean [SD] age, 70.7 [7.9] years; 236 female individuals [67.6%]; 113 male individuals [32.4%]). The study participants' clinical characteristics are summarized in **Table 1**. In this study, a total of 539 fundus images were obtained from 266 participants diagnosed with PD and demonstrated both H-Y scale and UPDRS-III score. Altogether, 349 participants (with 700 fundus images) with atypical motor abnormality were included as a control group. The 3 groups of participants categorized according to the H-Y scale were as follows: among 615 participants, 315 (57.0%) from the control group presented with only tremors not associated with PD and no other specific symptoms; 147 (23.9%) were diagnosed with early PD; and 117 (19.0%) presented with moderate or severe abnormalities affecting independent walking and activities of daily living. Regarding the UPDRS-III score, among the 615 participants, 351 (57.0%) demonstrated a UPDRS-III score of 0, 79 (12.8%) had UPDRS-III scores of 1 to 14; and 185 (30.0%) had UPDRS-III scores of 15 or greater. Data were randomly obtained from 80 people for the validation data set and the rest for the training data set.

Evaluation of the Deep Learning System for the Classification of Abnormalities

Initially, we evaluated if the deep learning system could distinguish PD based on the fundus images. Five well-known CNNs (ResNet-18, ResNet-101, VGG19, EfficientNet-b0, and EfficientNEt-b7) were used based only on the fundus images. As shown in the eTable in Supplement 1, the CNNs successfully discriminated between the groups with and without PD. ResNet-18, in particular, achieved the best performance, with a sensitivity of 77.05% (95% CI, 74.93%-79.18%), a specificity of 76.46% (95% CI, 74.75%-78.18%), an AUROC of 0.76 (95% CI, 0.74-0.78), and an accuracy of 76.75% (95% CI, 75.51%-77.99%) (eTable in Supplement 1). Therefore, we used the ResNet-18 as a baseline network for building our deep learning system.

Severity Classification Performance With Clinical Data in the Validation Data Set

A multimodality method was used to accurately classify the severity of symptoms and motor function of patients with PD using the H-Y scale and UPDRS-III score when used in conjunction with fundus images and clinical data. For this experiment, we constructed the following 4 models: model 1, fundus images only; model 2, fundus images with sex and age data; model 3, fundus images with sex, age, diabetes, and hypertension data; and model 4, fundus images with sex, age, diabetes, and hypertension data along with a multimodal method. The multimodal method is presented in **Figure 2**. The fourth model achieved the best performance, with a sensitivity of 83.23% (95% CI, 82.07%-84.38%), a specificity of 66.81% (95% CI, 64.97%-68.65%), an AUROC of 0.77 (95% CI, 0.75-0.79), and an accuracy of 73.38% (95% CI, 71.55%-75.21%) for the H-Y scale (**Table 2**). For UPDRS-III, the fourth model demonstrated a sensitivity of 82.61% (95% CI, 81.38%-83.83%), a specificity of 65.75% (95% CI, 62.56%-68.94%), an AUROC of 0.77 (95% CI, 0.75-0.79), and an accuracy of 73.38% (95% CI, 71.55%-75.21%) (**Table 3**).

Severity Classification Performance in the External Testing Data Set

In the external testing data set, we collected sex and age information along with the fundus image with H-Y scale of patients from Yeungnam University Hospital and Seoul National University Hospital. Because only 2 clinical data points, ie, sex and age information, were available from the external data set, we built a deep learning system by using the fourth method, which used fundus images and 2 clinical data points along with the multimodal method. The deep learning system discriminated the H-Y scale with a sensitivity of 80.32% (95% CI, 79.07%-81.56%), a specificity of 61.44% (95% CI, 58.02%-64.87%), an AUROC of 0.73 (95% CI, 0.70%-0.75%), and an accuracy of 65.42% (95% CI, 64.21%-66.62%) for the validation data set. A sensitivity of 70.73% (95% CI, 66.30%-75.16%), a specificity of 66.66% (95% CI, 50.76%-82.25%), an AUROC of 0.67 (95% CI, 0.55-0.79), and an accuracy of 70.45% (95% CI, 66.85%-74.04%) were obtained for the external testing data set.

Discussion

Predicting cognitive impairment among patients with PD is crucial not only for diagnostic purposes but also for risk

Figure 2. Performance Area Under the Receiver Operating Characteristic (AUROC) Curve of the Fundus Image and All Clinical Data Based on the Multimodality Method for the Hoehn and Yahr (H-Y) Scale and Unified Parkinson's Disease Rating Scale Part III (UPDRS-III) Score
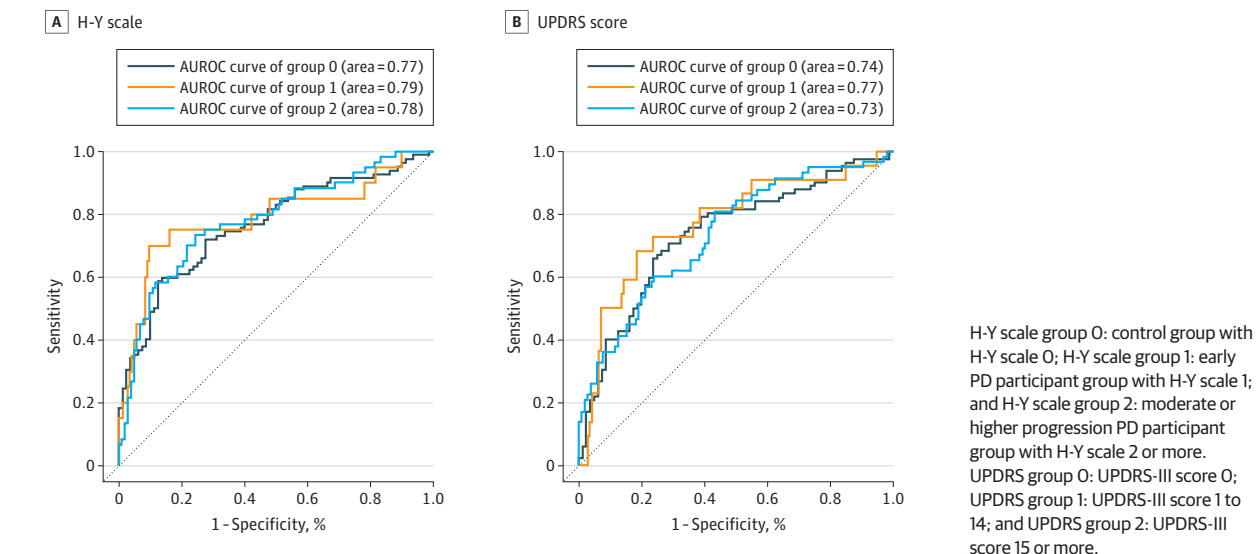

H-Y scale group 0: control group with H-Y scale 0; H-Y scale group 1: early PD participant group with H-Y scale 1; and H-Y scale group 2: moderate or higher progression PD participant group with H-Y scale 2 or more. UPDRS group 0: UPDRS-III score 0; UPDRS group 1: UPDRS-III score 1 to 14; and UPDRS group 2: UPDRS-III score 15 or more.

Table 2. Hoehn and Yahr (H-Y) Scale Prediction Results of the Deep Learning System

| Model | % (95% CI) | | AUROC (95% CI) | Accuracy, % (95% CI) |
	Sensitivity	Specificity		
1[a]	76.76 (75.39-78.12)	52.72 (48.89-56.55)	0.66 (0.63-0.69)	58.24 (56.17-60.32)
2[b]	79.48 (78.20-80.76)	58.81 (56.15-61.47)	0.71 (0.69-0.72)	64.80 (62.87-66.73)
3[c]	79.52 (78.59-80.45)	59.44 (57.46-61.42)	0.72 (0.70-0.73)	63.55 (61.58-65.53)
4[d]	83.23 (82.07-84.38)	66.81 (64.97-68.65)	0.77 (0.75-0.79)	73.38 (71.55-75.21)

Abbreviation: AUROC, area under the receiver operating characteristic curve.

[a] Model 1: fundus only.

[b] Model 2: fundus with sex and age data.

[c] Model 3: fundus with sex, age, diabetes, and hypertension data.

[d] Model 4: fundus image with sex, age, diabetes, and hypertension data with multimodal method.

Table 3. Unified Parkinson's Disease Rating Scale (UPDRS) Prediction Results of the Deep Learning System

| Model | % (95% CI) | | AUROC (95% CI) | Accuracy, % (95% CI) |
	Sensitivity	Specificity		
1[a]	77.16 (75.52-78.81)	54.75 (51.96-57.53)	0.67 (0.65-0.69)	59.43 (57.83-61.03)
2[b]	78.93 (77.84-80.03)	57.67 (55.68-59.66)	0.69 (0.68-0.71)	62.24 (60.43-64.05)
3[c]	79.54 (77.91-81.17)	59.21 (55.81-62.60)	0.70 (0.69-0.72)	65.31 (63.61-67.01)
4[d]	82.61 (81.38-83.83)	65.75 (62.56-68.94)	0.77 (0.75-0.79)	71.64 (70.19-73.09)

Abbreviation: AUROC, area under the receiver operating characteristic curve.

[a] Model 1: fundus only.

[b] Model 2: fundus with sex and age data.

[c] Model 3: fundus with sex, age, diabetes, and hypertension data.

[d] Model 4: fundus image with sex, age, diabetes, and hypertension data with multimodal method.

stratification of comorbidities, such as dementia. Therefore, identifying potential tools for the assessment of cognitive function that are readily accessible and reliable for patients with PD is of particular clinical importance. In this decision analytical model, our algorithm showed an AUROC of 0.77 for both the H-Y scale and UPDRS-III score based on our multimodal approach (Figure 2). A diverse score distribution for UPDRS-III in our data set may lead to a lower sensitivity and specificity for UPDRS-III, as compared with the H-Y scale in this study. Moreover, the lower AUROC of our

algorithm for the data at different hospitals was anticipated because there were no additional clinical data available (except for age and sex) for the external validation data set. Further studies are needed to add information on clinical data assessment function from fundus photography images to our algorithm, which will ensure a better cognitive impairment assessment function, regardless of data source.

The deep learning method enables us to investigate potential tools that are usually difficult to study through conventional investigation methods. Recently, using deep learning

methods, fundus photography has been used for cardiovascular disease risk assessment.[6,7] Several studies showed encouraging results for predicting not only conventional health-related variables (eg, age, sex, and smoking), but also for relatively new and specialized metrics (eg, coronary artery calcium score), which showed either compatible or significant results.[6,7] However, the application of deep learning on retinal imaging as a tool for assessing major cerebral or neurodegenerative disease (including PD) has been rarely used until now. This may be due to the difficulty in obtaining sufficient retinal images from patients with PD, as compared with patients with other systemic diseases. Our study results suggest that by modifying well-known open-source CNNs and using additional clinical information, we can conduct an algorithm with moderate to high assessment function when only small data are available, which is often the case in neurodegenerative diseases.

Results of this study further suggest the potential role of the retina as an assessment tool for neurologic dysfunction among patients with PD. Although dopaminergic cell loss is considered a pathologic change in both the macular area and brain, there is no direct evidence showing involvement of macular area in cognitive functioning among patients with PD. Our study results suggest that the CNN uses the macular area to assess the clinical severity of PD. This finding supports not only our hypothesis that the macular area is associated with neurologic function assessment among patients with PD but also supports black box models of deep learning processing (eFigure 2 in Supplement 1). However, the temporal or cause-effect association must be derived from future large longitudinal studies using the algorithm.

Limitations

This research has limitations that need to be addressed to help overcome future challenges in this field of study. First, due to the cross-sectional design of this study, the algorithm cannot predict the progression of the severity and neurologic dysfunction among patients with PD. Because PD is a chronic progressive disease, predicting the level of progression is an important factor for monitoring the diagnosis itself and evaluating treatment responses among patients. Again, large longitudinal data from patients with PD can magnify this algorithm's role in clinical practice. Second, because the predictive assessment function of the algorithm is based only on patients with PD, its application cannot be generalized to other neurodegenerative diseases, such as Alzheimer disease. The clinical importance of the algorithm can be significant if its use is just as effective as for other neurodegenerative diseases. Third, the algorithm is based on a standard open-source algorithm that is trained on 1 sample of data by adjusting coefficients; it may not work that well with new sources from other populations. We tried to minimize this by testing our algorithm function on 2 independent external hospital data sets. However, additional investigation among different ethnicities from different countries is needed to ensure the function of our algorithm.

Conclusions

In conclusion, this decision analytical model was a cross-disciplinary investigation reporting a deep learning algorithm that may predict the severity and neurologic dysfunction of patients with PD using fundus photography. Along with suggesting fundus photography as a potential tool for assessing PD patients, this study explores the role of deep learning method as a new modality in neurodegenerative disease research. Additional training using large longitudinal data for this algorithm is mandatory for the algorithm to be used in different clinical settings. Nevertheless, this study can be used as an example of an investigation that shows how a deep learning method can be applied using relatively small data that still yields promising results for neurodegenerative diseases.

REFERENCES

1. Dorsey ER, Constantinescu R, Thompson JP, et al. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology*. 2007;68(5):384-386. doi:10.1212/01.wnl.0000247740.47667.03

2. Yang W, Hamilton JL, Kopil C, et al. Current and projected future economic burden of Parkinson disease in the U.S. *NPJ Parkinsons Dis*. 2020;6:15. doi:10.1038/s41531-020-0117-1

3. Fetit AE, Doney AS, Hogg S, et al. A multimodal approach to cardiovascular risk stratification in patients with type 2 diabetes incorporating retinal, genomic, and clinical features. *Sci Rep*. 2019;9(1): 3591. doi:10.1038/s41598-019-40403-1

4. Rhee EJ, Chung PW, Wong TY, Song SJ. Relationship of retinal vascular caliber variation with intracranial arterial stenosis. *Microvasc Res*. 2016;108:64-68. doi:10.1016/j.mvr.2016.08.002

5. Nam GE, Han K, Park SH, Cho KH, Song SJ. Retinal vein occlusion and the risk of dementia: a nationwide cohort study. *Am J Ophthalmol*. 2021; 221:181-189. doi:10.1016/j.ajo.2020.07.050

6. Cheung CY, Xu D, Cheng CY, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nat Biomed Eng*. 2021;5(6): 498-508. doi:10.1038/s41551-020-00626-4

7. Cho S, Song SJ, Lee J, Song JE, Kim MS. Predicting coronary artery calcium score from retinal fundus photographs using convolutional neural networks. Paper presented at: Artificial Intelligence and Soft Computing, 19th International Conference; October 13, 2020; Zakopane, Poland.

8. Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. *Neurology*. 1967;17(5): 427-442. doi:10.1212/WNL.17.5.427

9. Goetz CG, Tilley BC, Shaftman SR, et al; Movement Disorder Society UPDRS Revision Task Force. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord*. 2008;23(15): 2129-2170. doi:10.1002/mds.22340

10. Ahn J, Shin JY, Lee JY. Retinal microvascular and choroidal changes in Parkinson disease. *JAMA Ophthalmol*. 2021;139(8):921-922. doi:10.1001/jamaophthalmol.2021.1728

11. Lee JY, Ahn J, Shin JY, Jeon B. Parafoveal change and dopamine loss in the retina with Parkinson disease. *Ann Neurol*. 2021;89(2):421-422. doi:10.1002/ana.25972

12. Sung MS, Choi SM, Kim J, et al. Inner retinal thinning as a biomarker for cognitive impairment in de novo Parkinson disease. *Sci Rep*. 2019;9(1):11832. doi:10.1038/s41598-019-48388-7

13. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med*. 2020;3(1):136. doi:10.1038/s41746-020-00341-z

14. Eskofier BM, Lee SI, Daneault JF, et al. Recent machine learning advancements in sensor-based mobility analysis: deep learning for Parkinson disease assessment. Paper presented at: the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 17, 2016; Orlando, Florida.

15. Grover S, Bhartia S, Akshama AY, et al. Predicting severity of Parkinson disease using deep learning. *Procedia Comput Sci*. 2018;132:1788-1794. doi:10.1016/j.procs.2018.05.154

16. Oliveira FP, Castelo-Branco M. Computer-aided diagnosis of Parkinson disease based on [(123)I]FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines. *J Neural Eng*. 2015;12(2):026008. doi:10.1088/1741-2560/12/2/026008

17. Satue M, Seral M, Otin S, et al. Retinal thinning and correlation with functional disability in patients with Parkinson disease. *Br J Ophthalmol*. 2014;98 (3):350–355.

18. Kaur M, Saxena R, Singh D, Behari M, Sharma P, Menon V. Correlation between structural and functional retinal changes in Parkinson disease. *J Neuroophthalmol*. 2015;35(3):254-258. doi:10.1097/WNO.0000000000000240

19. Mailankody P, Battu R, Khanna A, Lenka A, Yadav R, Pal PK. Optical coherence tomography as a tool to evaluate retinal changes in Parkinson disease. *Parkinsonism Relat Disord*. 2015;21(10): 1164–1169.

20. Garcia-Martin E, Larrosa JM, Polo V, et al. Distribution of retinal layer atrophy in patients with Parkinson disease and association with disease severity and duration. *Am J Ophthalmol*. 2014;157 (2):470–478.e2.

21. Bodis-Wollner I, Kozlowski PB, Glazman S, Miri S. α-synuclein in the inner retina in Parkinson disease. *Ann Neurol*. 2014;75(6):964-966. doi:10.1002/ana.24182

22. Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. *Neurology*. 1967;17(5): 427-442. doi:10.1212/WNL.17.5.427

23. Goetz CG, Tilley BC, Shaftman SR, et al; Movement Disorder Society UPDRS Revision Task Force. Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord*. 2008;23(15): 2129-2170. doi:10.1002/mds.22340

24. Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF. Automated assessment of diabetic retinal image quality based on clarity and field definition. *Invest Ophthalmol Vis Sci*. 2006;47(3):1120-1125. doi:10.1167/iovs.05-1155

25. Vasseneix C, Najjar RP, Xu X, et al; BONSAI Group. Accuracy of a deep learning system for classification of papilledema severity on ocular fundus photographs. *Neurology*. 2021;97(4):e369-e377. doi:10.1212/WNL.0000000000012226

26. Scikit-learn. Home page. Accessed April 5, 2022. https://scikit-learn.org/stable/

Invited Commentary

Deep Learning Using Images of the Retina for Assessment of Severity of Neurological Dysfunction in Parkinson Disease

Janan Arslan, MSc, MBiostat, PhD; Daniel Racoceanu, PhD; Kurt K. Benke, PhD

The deep learning (DL) approach is a recent development in research on the automation of medical diagnostics using machine learning. This type of statistical learning is not based on biophysical (mechanistic) models but on training (for parameter estimation) using a sample of population data.

Related article page 234

There are issues about robustness in performance when testing on new data sets in uncontrolled conditions separate from validation experiments. The use of DL in diagnostic modeling has been demonstrated in recent years with very encouraging results in the assessment of diabetic retinopathy, geographic atrophy in age-related macular degeneration, and cardiovascular disease when using images of the retina.[1]

The study by Ahn et al[1] in this issue of *JAMA Ophthalmology* provides insights and useful information on the application of DL to explore possible associations between changes in the retina and brain function. The reported work is interesting as it explores the use of DL approaches for assessment of cognitive function scores for disease severity in the case of Parkinson disease (PD), which is projected to be an increasing economic burden.[2] In the study, fundus photographs were collected from 615 participants who were outpatients at the neurology department of Kangbuk Samsung Hospital in South Korea (August 2020 to April 2021). To compensate for the typically small sample sizes often encountered in ophthalmology research, the authors applied a range of image augmentation techniques, which included random rotation and random flip, to increase the sample size.

Five common convolutional neural network (CNN) models were investigated to predict the Hoehn and Yahr Scale (H-Y) and Unified Parkinson's Disease Rating Scale part III (UPDRS) scores for PD,[3,4] with the training based on fundus photographs and patient demographic characteristics. The study echoed other works that have investigated the idea that changes manifested in the retina may provide indicators of neurological impairment due to PD.[5] The prospective case-control hospital-based study revealed that, for the internal validation data, the DL system had high sensitivity that was above 80% for both the H–Y and UPDRS scores. The specificity ranged from 66% to 67% for H-Y and UPDRS scores. For the external