Neurology®

The most widely read and highly cited peer-reviewed neurology journal
The Official Journal of the American Academy of Neurology

AMERICAN ACADEMY OF NEUROLOGY®

The Role of OCT Criteria and Machine Learning in Multiple Sclerosis and Optic Neuritis Diagnosis

Rachel C Kenney, PhD[1,2]; Mengling Liu, PhD[2]; Lisena Hasanaj, BA[1]; Binu Joseph, MBBS[1]; Abdullah Abu Al-Hassan, MD[3]; Lisanne J. Balk, PhD[4]; Raed Behbehani, MD[3]; Alexander Brandt, MD[5, 6]; Peter A. Calabresi, MD[7]; Elliot Frohman, MD, PhD[8]; Teresa C. Frohman, PA-C[8]; Joachim Havla, MD[9,10]; Bernhard Hemmer, MD[11,12]; Hong Jiang, MD, PhD[13]; Benjamin Knier, MD[11]; Thomas Korn, MD[11,12]; Letizia Leocani, MD[14]; Elena Hernandez Martinez-Lapiscina, MD, PhD[15]; Athina Papadopoulou, MD[5,16]; Friedemann Paul, MD[5,17]; Axel Petzold, MD, PhD[18]; Marco Pisa, MD[14]; Pablo Villoslada, MD[15]; Hanna Zimmermann, M.Eng[5,17]; Lorna E Thorpe, MPH, PhD[2]; Hiroshi Ishikawa, MD[19]; Joel S Schuman, MD[20, 21, 22, 23]; Gadi Wollstein, MD[20]; Yu Chen, MPH, PhD[2]; Shiv Saidha, MBBCh, MD, MRCPI[7]; Steven Galetta[1, 20]; Laura J Balcer, MD[1, 2, 20]

Corresponding Author:
Laura J Balcer, laura.balcer@nyulangone.org

Neurology® Published Ahead of Print articles have been peer reviewed and accepted for publication. This manuscript will be published in its final form after copyediting, page composition, and review of proofs. Errors that could affect the content may be corrected during these processes.

Affiliation Information for All Authors:

1 - Department of Neurology, New York University Grossman School of Medicine, New York, NY, USA

2 - Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA

3 - Al-Bahar Ophthalmology Center, Ibn Sina Hospital, Kuwait City, Kuwait

4 - Mulier Institute, Centre for Research on Sports in Society, Utrecht, Netherlands

5 - Experimental and Clinical Research Center, Max Delbrueck Center for Molecular Medicine and Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

6 - Department of Neurology, University of California, Irvine, CA, USA

7 - Department of Neurology, Johns Hopkins University, Baltimore, MD, USA

8 - Laboratory of Neuroimmunology, of Professor Lawrence Steinman, Stanford University School of Medicine, Palo Alto, CA, USA

9 - Institute of Clinical Neuroimmunology, LMU Hospital, Ludwig Maximilians Universität München, Munich, Germany

10 - Data Integration for Future Medicine consortium (DIFUTURE), Ludwig-Maximilians University, Munich, Germany

11 - Department of Neurology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

12 - Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

13 - Bascom Palmer Eye Institute, Department of Neurology, University of Miami Miller School of Medicine, Miami, FL, USA

14 - Vita-Salute University & Hospital San Raffaele, Milano, Italy

15 - Center of Neuroimmunology and Department of Neurology, Hospital Clinic of Barcelona, Institut d'Investigacions Biomèdiques August Pi Sunyer (IDIBAPS), University of Barcelona, Villarroel 170, Barcelona ES 08036, Spain

16 - Neurologic Clinic and Policlinic, MS Center and Research Center for Clinical Neuroimmunology and Neuroscience (RCN2NB) Basel, University Hospital Basel and University of Basel, Basel, Switzerland

17 - NeuroCure Clinical Research Center, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

18 - Moorfields Eye Hospital, City Road, London; The National Hospital for Neurology and Neurosurgery, Queen Square, UCL Institute of Neurology, London, United Kingdom; Dutch Neuro-ophthalmology Expertise Centre, Amsterdam UMC, The Netherlands

19 – Oregon Health and Science University, Portland, OR, USA

20 - Department of Ophthalmology, New York University Grossman School of Medicine, New York, NY, USA

21 - Departments of Biomedical Engineering and Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, NY, USA

22 - Center for Neural Science, NYU, New York, NY, USA

23 - Neuroscience Institute, NYU Langone Health, New York, NY, USA.

Equal Author Contribution:

Contributions:

Rachel C Kenney: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataMengling Liu: Drafting/revision of the manuscript for content, including medical writing for content; Study concept or design; Analysis or interpretation of dataLisena Hasanaj: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of dataBinu Joseph: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of dataAbdullah Abu Al-Hassan: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataLisanne J. Balk: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataRaed Behbehani: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataAlexander Brandt: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataPeter A. Calabresi: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataElliot Frohman: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of data

Teresa C. Frohman: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataJoachim Havla: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataBernhard Hemmer: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataHong Jiang: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataBenjamin Knier: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataThomas Korn: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataLetizia Leocani: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataElena Hernandez Martinez-Lapiscina: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataAthina Papadopoulou: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataFriedemann Paul: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataAxel Petzold: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataMarco Pisa: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataPablo Villoslada: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataHanna Zimmermann: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataHiroshi Ishikawa: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataJoel S Schuman: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Analysis or interpretation of dataGadi Wollstein: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataYu Chen: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataShiv Saidha: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataLorna E Thorpe: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataSteven Galetta: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of dataLaura J Balcer: Drafting/revision of the manuscript for content, including medical writing for content; Major role in the acquisition of data; Study concept or design; Analysis or interpretation of data

Figure Count:
3

Table Count:
4

Search Terms:
[41] Multiple sclerosis, [43] Optic neuritis; see Neuro-ophthalmology/Optic Nerve (S), [111] Diagnostic test assessment, [188] Optic nerve, Optical Coherence Tomography, [323] Class III

Acknowledgment:

Abstract:

Background and Objectives: Recent studies have suggested that inter-eye differences (IEDs) in peripapillary retinal nerve fiber layer (pRNFL) or ganglion cell+inner plexiform (GCIPL) thickness by spectral-domain optical coherence tomography (SD-OCT) may identify people with a history of unilateral optic neuritis (ON). However, this requires further validation. Machine learning classification may be useful for validating thresholds for OCT IEDs and for examining added utility for visual function tests, such as low-contrast letter acuity (LCLA), in the diagnosis of people with multiple sclerosis (PwMS) and for unilateral ON history.

Methods: Participants were from 11 sites within the International Multiple Sclerosis Visual System (IMSVISUAL) consortium. pRNFL and GCIPL thicknesses were measured using SD-OCT. A composite score combining OCT and visual measures was compared individual measurements to determine the best model to distinguish PwMS from controls. These methods were also used to distinguish those with history of ON among PwMS. ROC curve analysis was performed on a training dataset (2/3 of cohort), then applied to a testing dataset (1/3 of cohort). Support vector machine (SVM) analysis was used to assess whether machine learning models improved diagnostic capability of OCT.

Results: Among 1,568 PwMS and 552 controls, variable selection models identified GCIPL IED, average GCIPL thickness (both eyes), and binocular 2.5% LCLA as most important for classifying PwMS vs. controls. This composite score performed best, with AUC=0.89 (95% CI 0.85, 0.93), sensitivity=81% and specificity=80%. The composite score

ROC curve performed better than any of the individual measures from the model (p<0.0001). GCIPL IED remained the best single discriminator of unilateral ON history among PwMS (AUC=0.77, 95% CI 0.71,0.83, sensitivity=68%, specificity=77%). SVM analysis performed comparably to standard logistic regression models.

Conclusions: A composite score combining visual structure and function improved the capacity of SD-OCT to distinguish PwMS from controls. GCIPL IED best distinguished those with history of unilateral ON. SVM performed as well as standard statistical models for these classifications.

Classification of Evidence: The study provides Class III evidence that SD-OCT accurately distinguishes multiple sclerosis from normal controls as compared to clinical criteria.

Introduction:

Recent studies have suggested that an inter-eye difference (IED) in peripapillary retinal nerve fiber layer (pRNFL) thickness of 5 microns and a 4-micron IED in ganglion cell inner+plexiform layer (GCIPL) thickness by spectral-domain optical coherence tomography (SD-OCT) optimizes identification of people with multiple sclerosis (PwMS) who harbor a history of unilateral optic neuritis[1-3].

The diagnostic utility of IEDs by OCT to detect history of optic neuritis (ON) requires further validation. Machine learning techniques have developed significantly over the past several years and may be useful for validating IED thresholds by OCT. Such models may examine the added utility of visual function tests, such as low-contrast letter acuity (LCLA), for confirming history of acute ON.

Support vector machine (SVM) and other machine learning classifier (MLC) methods have yielded prediction models with high degrees of accuracy, sensitivity and specificity. In one study, the MLC method out-performed general ophthalmologists for accurate diagnosis of glaucoma[4]. These models utilize areas under the curves (AUC) for Receiver Operating Characteristic Curve (ROC) methods to measure model performance. ROC analysis has shown that pRNFL thickness is an effective discriminator for glaucoma versus healthy eyes[5, 6].

There is a paucity of studies aimed at evaluating utility of MLCs for distinguishing PwMS from controls, or PwMS with ON history vs. those without. One small study by DelPalomar et al. found that MLCs with swept-source (SS-) OCT showed high levels of ability to differentiate PwMS vs. controls for pRNFL thickness; this was not the case for GCIPL thickness[7]. A recent study by Cavaliere et al. used SVM classifiers to distinguish control subjects from PwMS

without ON, and demonstrated sensitivity=89%, specificity=92% and AUC=0.97 when applying machine-learning techniques for retinal neurodegeneration. In contrast to the study by Del Palomar et al., these authors found total macular GCIPL thickness to be one of the most discriminant variables[8].

Further studies incorporating inter-eye thickness differences using SD-OCT are needed to investigate the ability of MLCs to distinguish PwMS vs. healthy controls and ON vs. non-ON eyes in PwMS in larger cohorts within a wider range of ages and countries. The primary research question being addressed in the study was to assess the utility of OCT in MS and ON diagnosis. The purpose of the current investigation was to 1) determine the best diagnostic predictors using SD-OCT, low- and high-contrast visual acuity and demographics for distinguishing PwMS from controls; 2) determine best diagnostic predictors to identify those with history of unilateral ON within a cohort of PwMS; 3) examine the ability of MLCs to classify these groups using best determined predictors. We hypothesized that adding visual parameters and demographics would improve the ability of OCT to detect MS and ON history, and that MLCs would also perform well in classifying these groups.

Methods:

Study Cohort

Participants for this study included a convenience sample of 546 control subjects with no history of neurological or ophthalmological disease and 1,568 people with MS (PwMS) who fulfilled the 2010 McDonald criteria for diagnosis. All participants were ≥18 years of age and had spherical diopter refractive error between -6-+6. Age, race, and gender were included where available. Participants were part of an 11-site collaboration (International Multiple Sclerosis Visual System [IMSVISUAL] consortium) in the United States, Europe and the Middle East. Data collection methods have been previously described[2]. Subjects with ocular comorbidities were excluded. Patients with an episode of acute ON within 3 months prior to data collection, or those with other demyelinating diseases such as neuromyelitis optica spectrum disorder or myelin oligodendrocyte glycoprotein antibody disease (MOGAD), were excluded (**Figure 1**).

Standard Protocol Approvals, Registrations and Patient Consents

Study protocols were approved by Institutional Review Boards (IRBs), and all participants provided written informed consent.

Visual Assessments

LCLA was measured monocularly and binocularly at 2.5% and 1.25% contrast levels using low-contrast Sloan letter charts (Precision Vision, LaSalle, IL) in a retro-illuminated light-box[9]. Scores are recorded as numbers of letters read correctly/70 per chart. The Early Treatment Diabetic Retinopathy Study (ETDRS) visual acuity charts or Snellen acuity charts were used to measure high-contrast visual acuity (HCVA). Scores were normalized and converted to the ETDRS scale. LCLA and HCVA were not performed at every site (**Figure 1**); as such, only subjects with these measurements were included in analyses. Refraction and visual field assessments were performed at one site (NYU).

Optical Coherence Tomography

Spectralis SD-OCT (Heidelberg Engineering, Heidelberg, Germany) or Cirrus HD-OCT (Carl Zeiss Meditec, Dublin, CA) scans were obtained at each site by trained technologists[2]. pRNFL thickness was measured using a 3.4 mm diameter circle centered on the optic disc. Macular GCIPL thickness was also assessed. GCIPL measurements were obtained from automated macular volume cube scans, with measurement area of a 4x5 mm annulus surrounding the fovea[10] on Cirrus OCT, and from the macular volume scan encompassing 6x6 mm cylinder surrounding the fovea on Spectralis OCT. OCT scans from Cirrus and Spectralis devices were reviewed and OSCAR-IB criteria for quality control were followed[11]. Scans not meeting quality control criteria, including not being centered properly, were excluded. OCT results were reported in concordance with APOSTEL 2.0 guidelines[12, 13]. Since two different devices, Spectralis and Cirrus, were used based on site protocols, Spectralis SD-OCT measurements were converted to the Cirrus scale using equations established by concurrent work (unpublished data):

[Cirrus= -5.0 + 1.0*Spectralis pRNFL]

[Cirrus= -4.5+0.9*Spectralis GCIPL]

These conversion equations were developed using structural equation modeling (R^2=0.996), accounting for clustering, with control data (n=346) from one site where participants were scanned on both devices on the same day.

Optic Neuritis

History of acute ON was defined as an exacerbation consistent with inflammatory demyelinating event affecting the optic nerve, self-reported by the patient and verified by medical record review. Clinical features considered in ON diagnosis were pain on eye movement, subacute visual loss, reduced visual acuity, afferent pupillary defect and reduced

color vision. The episode of ON was categorized as unilateral (n=488) or bilateral (n=132). ON refers to any instance of optic neuritis, and, where applicable, will be classified as unilateral or bilateral. People with uncertain history of ON (n=74), either because the data were not available or because the patient was not sure of their history, were excluded from the analyses of IED thresholds.

Statistics

Descriptive statistics were used to report continuous variables (mean±SD), assuming normality of distribution, with categorical variables as frequencies and percentages. PRNFL and GCIPL IEDs were calculated by subtracting right eye from left eye values and reporting absolute value. Average pRNFL and GCIPL thicknesses were calculated as averages between the two eyes. Conversion equations, described above, were used to pool data from Spectralis and Cirrus OCT devices. Each potential diagnostic criterion (**Table 1**) was standardized to have a mean of 0 and a standard deviation of 1. A composite score combining the above diagnostic criteria was developed to assess for improvements in areas under the receiver operating characteristic (ROC) curves (AUCs). The composite score was developed using logistic regression with the dichotomous outcome of MS vs. control diagnosis.

Accuracies and AUCs for several models were compared to find the best model for distinguishing PwMS from controls. Forward and backward selection using logistic regression were used to determine variables above that had the most significant contributions to the models. Classification and regression tree (CART) models were used to determine variable importance. The CART model was developed using a training dataset (2/3 of cohort) and tested on a testing dataset (1/3 of cohort). We compared the following five models: 1) a full model with all variables; 2) a partial model with variables with high numbers of missing data excluded (disease duration, binocular acuities), 3) models with variables from forward and backward selection; 4) a model with the three most important variables from the CART model; and 5) a model with the most clinically relevant variables. Composite scores for the four best models were developed and compared using ROC curves for the training dataset (2/3 of cohort), and then applied to the testing dataset (remaining 1/3 of cohort). Training and testing datasets were chosen randomly.

Comparisons of the AUCs were performed in a pair-wise fashion accounting for within-patient, inter-eye correlations and modeled after DeLong et al.'s nonparametric approach, accounting for the correlated nature of the data[14]. "Cut-points," or thresholds for the composite score, were chosen based on the 1-sum of squares method. Since HCVA

can be measured using numbers of letters correct or logMAR units, the composite scores incorporated both scoring methods.

A second composite score was developed based on the results of a logistic regression model with the dichotomous outcome of presence or absence of history of ON and included weights based on beta effect estimates for significant diagnostic criterion for classifying unilateral ON history in PwMS. Sensitivity and specificity for the overall pRNFL and GCIPL thinning, inter-eye and overall LCLA scores, and the composite scores were calculated and compared. Sensitivity and specificity for pRNFL and GCIPL IEDs have been previously analyzed and published[2]. Lastly, a third composite score for history of ON by eye was developed and compared. A binary composite score was also developed based on the optimal cut-points identified on ROC curve analysis for classification of MS history, unilateral ON history by patient and ON history by eye. Analyses were performed using Stata 16.0 and R.

Support vector machine (SVM) analysis was used to assess diagnostic capability of OCT to distinguish PwMS from controls and to assess potential incremental capacity by AUCs compared vs. logistic regression models. SVM is a supervised machine learning method that seeks to define a hyperplane to separate two classes (i.e., MS vs. controls) with maximized margins. A margin is the distance between the nearest data point in either set and the hyperplane. SVM models input data and can transform it using mathematical functions to non-linear models using various kernels. Linear, radial, polynomial and sigmoid kernels were evaluated in these analyses. Only complete cases with no missing data were included in SVM models. To determine best cost and gamma parameters, 10-fold cross-validation was used. Models evaluated included a full model with all variables (**Table 1**), and a model with a subset of variables used to develop the composite score. SVM models were developed using a training data set (2/3 of cohort) and independent testing data set (1/3 of cohort). AUC were compared to assess the best model fit. Our data complied with the OSCAR-AI (RASCO) criteria[15] where applicable.

RESULTS

Study Cohort

For the 2,120 participants (1,568 PwMS, 552 controls, **Figure 1**) , characteristics are shown in **Table 2**. History of ON (unilateral or bilateral) was reported in 39.5% of PwMS (n=620/1,568). The MS cohort had a larger proportion of females than controls (70% vs 55%, p<0.001). As would be expected, PwMS, particularly those with ON history, had lower pRNFL and GCIPL thicknesses, lower (worse) LCLA and HCVA scores, and larger IEDs for all measures.

Composite Score for MS Disease Classification

Variable selection models using logistic regression identified GCIPL IEDs, average GCIPL thicknesses for both eyes, pRNFL IEDs, binocular 2.5% contrast LCLA scores, and binocular HCVA scores as most important measures for distinguishing PwMS vs. controls. The CART model identified GCIPL IED, average GCIPL thickness for both eyes, and binocular 2.5% LCLA as the most important variables. The two models (logistic regression variable selection and CART) were compared and the model with fewer variables performed equally well (p=0.42). Therefore, GCIPL IED, average GCIPL thickness for both eyes, and binocular 2.5% LCLA were combined to form a composite score. This composite score had the best performance among all of the models tested for distinguishing PwMS vs. controls. AUC was 0.89 (95% CI 0.85, 0.93), with sensitivity 81%, specificity 80%, and accuracy 81%. **Figure 2** shows the ROC curve for this model and for each individual measure contributing to the composite score. Comparisons of ROC curves for composite score to those for each individual measure showed that the composite score was a better model (p<0.0001 for inter-eye GCIPL and GCIPL average of both eyes; p=0.0002 for binocular 2.5% LCLA). Sensitivity, specificity and AUCs for each OCT measure and for the composite score are shown in **Table 3**. A model substituting HCVA for LCLA in the composite score was evaluated since HCVA measures may be more available clinically than LCLA. This model, with GCIPL IED, average GCIPL thickness of both eyes, and binocular HCVA also performed very well, with AUC 0.89 (95% CI 0.85, 0.93), sensitivity 86%, specificity 78% and accuracy 84% (**Table 4**). Scores greater than 0 indicate higher likelihood of MS for all composite scores (**Figure 3**).

The cut-points identified for each measure, where sensitivity and specificity are optimized using ROC curve analysis, were as follows: GCIPL IED ≥3 microns, average GCIPL thickness for both eyes ≤80 microns, and binocular LCLA at 2.5% contrast ≤42 letters (20/40 Snellen). When considering these binary outcomes, if a patient met all 3 criteria, the specificity for MS diagnosis was 99.5%, with a sensitivity of 11.4%. For comparison, the sensitivities and

specificities for each individual binary measure alone were as follows: 70.6% and 73.4% for GCIPL IED ≥3 microns, 40.7% and 81.3% for average GCIPL thickness for both eyes ≤80 microns, and 33.8% and 83.3% for binocular LCLA at 2.5% contrast ≤42 letters (20/40 Snellen). The threshold of GCIPL IED ≥4 microns was previously published (ref) and is discussed in the following paragraph for ON classification. We substituted this threshold instead of GCIPL IED ≥3 microns in the binary composite score for MS classification to compare results and found sensitivity=9.9% and specificity=98.9%. Additionally, GCIPL IED ≥4 microns was evaluated individually and found to have sensitivity=44.9% and sensitivity=96.7%, with an AUC=0.75.

Composite Score for ON Classification

GCIPL IED, average GCIPL thickness for both eyes, and pRNFL IED were identified as best for identifying a history of unilateral ON in PwMS. The composite score combining these measures did not improve the model (AUC=0.77, 95% CI 0.71, 0.83) compared to GCIPL IED alone (AUC=0.77, 95% CI 0.71, 0.83, sensitivity=68%, specificity=77%, **Table 3**). The binary composite score (GCIPL IED ≥4 microns, average GCIPL thickness ≤75 microns and pRNFL IED ≥5 microns) had high specificity (97.9%) and low sensitivity (7.1%).

For classifying ON history by eye, instead of by patient, GCIPL IED, average GCIPL thickness for both eyes, overall pRNFL thickness (by eye), average pRNFL thickness of both eyes and age performed well. When these measurements were combined into a composite score, the composite had AUC=0.72 (95% CI 0.67, 0.76) and optimized sensitivity (58%) and specificity (77%). Sensitivity, specificity and AUCs for each separate OCT measure and for the composite scores are shown in **Table 3**. The composite score did not perform better than any individual OCT measure for identifying history of ON in that eye. Sensitivity, specificity, and AUCs for the 4 best models are shown in **Supplementary Table 1**.

Support Vector Machine

Among 482 participants with complete data who were included in the SVM analysis, the training set included n=321, and the testing dataset included n=161. The linear kernel, which can be used when data are linearly separable, performed best when compared to radial, sigmoid and polynomial kernels. The linear model had 58 support vectors when used on the testing dataset. SVM and logistic regression both performed best, as would be expected, when using all of the potential variables in the model (**Table 1**). The SVM method yielded sensitivity=86%, specificity=79%, accuracy=84% and AUC=0.95 in correctly identifying PwMS vs. controls. Using the logistic regression model, there was

sensitivity=86%, specificity=89%, accuracy=88% and AUC=0.96. Using the subset of variables identified for the composite score performed comparably. The SVM method yielded sensitivity=83%, specificity=90%, accuracy=88% and AUC=0.93. The logistic regression model yielded sensitivity=81%, specificity=87%, accuracy=85% and AUC=0.93. Similar results were seen for classification of ON history within PwMS. The linear kernel performed best and the AUC did not improve using SVM compared to logistic regression models. SVM showed an AUC=0.89, while logistic regression yielded AUC=0.91 using the full model with all variables.

Classification of Evidence (COE)

Classification of Evidence: The study provides Class III evidence that SD-OCT accurately distinguishes multiple sclerosis from normal controls as compared to clinical criteria.

Discussion:

Results of this investigation demonstrate that a composite score combining GCIPL IED, average GCIPL thickness of both eyes, and 2.5% LCLA improved the capacity of single-eye SD-OCT measurements alone to distinguish PwMS from healthy controls. GCIPL IED was the most informative measure for identifying unilateral ON among PwMS. SVM analyses were equally effective vs. standard statistical models. All study objectives were met and hypotheses were confirmed to demonstrate that OCT and visual function measures are able to distinguish PwMS vs. controls, as well as to identify those with history of ON.

Individual SD-OCT measures did not perform as well as the composite score in distinguishing PwMS vs. controls. This could be because not all PwMS have optic nerve disease or retinal thinning. In addition, even if a PwMS does have ON history, degrees of retinal thinning could have OCT values in normal ranges. The composite score for classification of PwMS incorporates GCIPL IEDs, average GCIPL of both eyes, and binocular 2.5% LCLA. GCIPL is a measure of visual structure, while LCLA and HCVA measure function. Including three potential measures that incorporate both increases SD-OCT sensitivity and specificity. The model that incorporates the same GCIPL variables, combined with HCVA, also performed well in distinguishing PwMS from controls. Since HCVA is more common in clinical practice, the prediction equation with HCVA may be more broadly applicable. Models incorporating monocular acuity did not perform as well. The GCIPL IED component of the composite score would argue for an asymmetric optic neuropathy while a reduced binocular 2.5% LCLA score would imply a visual deficit. It could be that having an

asymmetric GCIPL could produce a binocular vision deficit. For clinical applications, it is important to consider the context. For example, if a patient came in with a first demyelinating episode suggestive of MS, and then was found to have these OCT and visual deficits, it could argue strongly for the optic nerve as an involved site. However, if a patient had no other neurological complaints, then it would be not enough to make a diagnosis or to start MS treatment. However, it could most certainly prompt further investigation. Similarly, if a patient had other symptoms or a clinical history suggestive of optic neuritis, a GCIPL IED on OCT could confirm the diagnosis. In the context of the MS McDonald diagnostic criteria, GCIPL IED could potentially satisfy requirements for dissemination in time or space as an optic nerve event.

A composite score did not improve the capacity for OCT to distinguish people (or eyes) with a unilateral ON history within the MS cohort compared to previously published results using GCIPL IED alone[2]. A recent study by Coric et al.[3] showed a diagnostic sensitivity of 70% for GCIPL inter-eye difference (as a percentage difference), with a specificity of 97% for distinguishing people with unilateral ON vs. healthy controls. The specificity in that study may be higher since healthy controls were the comparison group; in the present study, PwMS without ON history represented the comparison group. Our results confirm high degrees of capacity for GCIPL IED to distinguish eyes and people with an ON history. Therefore, to identify ON history within an MS cohort, the simplest model that uses GCIPL IED is sufficient, with a threshold of 4-microns. However, caution should be used when interpreting results in a clinical setting; other mimickers of MS and ON should be considered before making a diagnosis given the moderate sensitivity (68%) and specificity (77%) of this measure.

Binary composite scores (utilizing the measures in the composite score, such as GCIPL IED, and incorporating a normal/abnormal component above or below the cut-points for each measure), for both identification of PwMS vs. controls and for distinguishing people with vs. without ON history, had high levels of specificity (99.5% and 97.9%, respectively) with low levels of sensitivity (11.4% and 7.1%, respectively). High specificity correlates with low false positive rate, and lower degrees of IED may be useful to identify those who do not have disease (MS or ON). Low levels of sensitivity may indicate a cautious test with a high likelihood of false negatives. For tests that are candidates for diagnostic criteria, high levels of specificity are important to avoid false positives[16]. In the context of adding a test to the diagnostic criteria for MS, or to identify ON history, a test with high specificity may be useful. People that are classified as "positive" on the test are very likely to have MS/ON history; so, coupled with other diagnostic tests, the OCT results

could potentially improve overall diagnostic criteria. In situations for which high levels of sensitivity and specificity are preferred, the composite score with continuous variables would be more appropriate. In clinical practice, if a patient has abnormal LCLA, a large GCIPL IED, and GCIPL thinning in one or both eyes, this increases likelihood of an optic neuropathy, potentially MS-related, although other causes of optic nerve injury should be ruled out before confirming a diagnosis.

The linear SVM model performed best, compared to non-linear SVM models. However, logistic regression models performed just as well; this method may produce more interpretable results compared to SVM models. The comparability in performance between SVM and logistic regression models could also mean that machine learning models could be useful in certain contexts, for example when no neuro-ophthalmologist is available to interpret results. Similar degrees of performance between logistic regression and linear SVM models suggest that variables in the models are likely to be linear. Non-linear models may not be needed in this case. Machine learning methods work best with large sample sizes and large numbers of feature variables; the present analyses have a relatively small sample sizes and relatively small numbers of features variables compared to traditional machine learning cohorts. This may help explain why machine learning models did not improve diagnostic accuracy for this particular study cohort. It is reassuring that human-driven design and interpretation of statistical models, which should be planned ahead of research collection, can be just as effective as high-performing machine learning methods in certain instances. Further studies could explore other supervised machine learning methods, such as Random Forest or Gradient Tree Boosting, which have performed well in glaucoma diagnostics. Deep learning/artificial intelligence models using SD-OCT images may be useful for future studies of diagnosis for MS and for investigation of the role for ON and optic nerve lesions as MS diagnostic criteria.

Limitations of this study include the cross-sectional design. Since longitudinal data for these investigations are not yet available, predictions of change over time, such as the capacity to predict future ON or future conversion to MS based on current diagnostic criteria, are yet to be determined. Another limitation is the fact that GCIPL thickness and HCVA were not measured for all sites. In addition, 74 participants did not have information on ON history. These limitations were addressed in detail in our initial publication on inter-eye differences in OCT measures[2]. We would not expect bias to be introduced from the small amount of missing data. Participants with no GCIPL, HCVA, or ON history data were similar in all baseline characteristics compared to those with these measures included. Participants who did not have LCLA had slightly longer disease durations than did those with LCLA measurements (10.2 vs. 6.7 years); however,

distributions of Expanded Disability Status Scale scores (EDSS), disease subtypes, pRNFL and GCIPL values were similar. It is unlikely that this subgroup had more severe optic nerve disease and, therefore, we would not expect the results to be biased by these characteristics. The healthy control cohort included 55% females while the PwMS cohort was 70% female. In a previous study utilizing these data, sensitivity analyses were performed to compare the cohorts; it was determined that no significant bias would come from these discrepancies in gender distribution[2]. Although there could be site-related effects, site was not included as a covariate in the statistical models. Site would not be interpretable in the composite score or machine learning models. To overcome potential site differences, the conversion equation described was used to normalize data. Further, other factors such as age, gender, and race/ethnicity were included in the models. Given these adjustments, we do not expect any bias was introduced from site-related effects.

The data were normalized to be on the scale of Zeiss devices; however, these data can be converted to Spectralis scale using the conversion equations provided. Inter-eye differences across Zeiss and Spectralis are comparable and not subject to the overall thickness of the retina or the device used based on our previous study[2]. We would expect this to be true for all OCT devices; therefore, IED could be used as a measure regardless of device used. However, the average GCIPL measure incorporated into the composite score for MS diagnosis could be affected by device differences. If other devices are used, a conversion equation would need to be developed to convert to that device's scale, or a consistent segmentation algorithm used on both devices. Similarly, macular grid size of 1,3,6mm were used to measure GCIPL thickness. If an alternate grid size was used, a comparison would need to be made to assess the level of adjustment needed.

A composite score incorporating GCIPL IED thickness, average GCIPL thickness for both eyes, and binocular 2.5% LCLA (or HCVA) improved diagnostic accuracy of SD-OCT for differentiating PwMS vs. controls. A GCIPL IED thickness of ≥4 microns is the best measure to distinguish ON history in PwMS. Results of this study provide strong preliminary data upon which further studies could be designed to evaluate the utility of IEDs and of the composite score to identify optic nerve lesions as a 5th anatomical lesion site for the next revision of the McDonald criteria for MS. For the composite score to be added to the MS diagnostic criteria, a subsequent study may be done showing the capacity of the score to potentially increase sensitivity and specificity of the current diagnostic criteria. This may require substitution of the vision composite score with current criteria or could be done by developing a protocol using data from patients with clinically-isolated syndromes followed over time. Such studies would incorporate the vision composite score at baseline

to see if it would correctly capture the patients who progress to clinically definite MS. GCIPL IEDs could also be

considered for inclusion in the MS diagnostic criteria as evidence of a potentially MS-related optic neuropathy and as a

proxy for an optic nerve lesion. A similar study as described for the composite score could be designed to evaluate the

capacity of this measure to improve MS diagnosis. Thus, IED could be used to identify a 5th lesion site, while the

composite score could be most helpful to distinguish MS patients from healthy controls. As such, both could be additive

to the diagnostic criteria. Machine learning models performed as well as logistic regression, suggesting that these models

should be explored further studies of MS and ON, perhaps incorporating raw SD-OCT image analyses.

Data Availability Statement: Anonymized data may be made available by reasonable request from a qualified

investigator.

Table 1. Potential variables used in classification models.

Demographics	Age
	Sex
	Race/ethnicity
	Disease duration (within MS cohort only)
SD-OCT	Inter-eye pRNFL thickness difference
	Inter-eye GCIPL thickness difference
	Average pRNFL thickness of both eyes
	Average GCIPL thickness of both eyes
	pRNFL thickness (for each eye)
	GCIPL thickness (for each eye)
Visual Acuity	LCLA 2.5% monocular acuity
	LCLA 2.5% binocular acuity
	LCLA 2.5% average of both eyes
	LCLA 2.5% inter-eye difference
	LCLA 1.25% monocular acuity
	LCLA 1.25% binocular acuity
	LCLA 1.25% average of both eyes
	LCLA 1.25% inter-eye difference
	HCVA monocular acuity
	HCVA binocular acuity
	HCVA average of both eyes
	HCVA inter-eye difference

Table 2. Baseline demographics for study cohort [2008-2017]

	Healthy Controls n=552	MS Cohort		
		All PwMS n=1568	PwMS without ON n=874	PwMS with Unilateral ON n=488*
Age, years, mean (SD)	39.4 (14.6)	41.9 (11.2)	42.9 (11.5)	40.5 (11.0)
Sex, n (% female)	303 (55%)	1,097 (70%)	602 (69%)	333 (68%)
History of unilateral ON, n(%)	-	488 (36%)	0 (0%)	488 (100%)
Disease duration, median, IQR	-	7.0 (3.0-12.5)	6.5 (3.0-11.7)	6.5 (3.0-12.8)
pRNFL thickness, microns, mean (SD)	93.7 (9.9)	84.7 (14.8)	88.4 (13.4)	81.5 (15.5)
PRNFL IED, microns, mean (SD)	3.0 (2.9)	7.1 (8.5)	4.6 (5.6)	11.4 (10.7)
GCIPL thickness, microns, mean (SD)	84.6 (6.7)	76.0 (10.5)	78.3 (9.9)	74.8 (10.5)
GCIPL IED, microns, mean (SD)	1.4 (2.3)	5.3 (6.4)	3.0 (4.3)	8.9 (7.9)
Binocular HCVA, number of letters correct of 70	62.7 (5.2)	57.5 (8.1)	57.9 (7.2)	57.4 (7.8)
Inter-eye HCVA, letters	3.1 (3.2)	4.8 (9.0)	3.5 (5.7)	6.7 (12.3)
Binocular HCVA (logmar)	-0.15 (0.10)	-0.05 (0.16)	-0.06 (0.14)	-0.05 (0.16)
Inter-eye HCVA (logmar)	0.06 (0.06)	0.10 (0.20)	0.07 (0.12)	0.14 (0.27)
Binocular 2.5% LCLA (number letters correct of 70)	42.9 (6.4)	33.2 (11.8)	34.1 (11.6)	33.6 (12.0)
Inter-eye 2.5% LCLA	4.5 (4.5)	8.3 (8.8)	6.4 (6.5)	11.1 (11.0)

MS = multiple sclerosis, PwMS = people with MS, ON=optic neuritis, pRNFL = peripapillary retinal nerve fiber layer, GCIPL = ganglion cell+inner plexiform layer, HCVA=high-contrast visual acuity, LCLA=low-contrast letter acuity, SD=standard deviation.
*132 participants had bilateral optic neuritis, and 74 had an unknown history of optic neuritis.
Study cohort derived from the IMSVISUAL consortium.

Table 3. Performance of classification models.

Classification	Model	AUC	Sensitivity	Specificity
MS vs. healthy control	Inter-eye GCIPL	0.75	53	91
	Inter-eye pRNFL	0.70	36	33
	GCIPL Thickness, average of both eyes	0.77	28	34
	pRNFL Thickness, average of both eyes	0.65	57	67
	Composite Score*	0.89	81	80
Unilateral optic neuritis by patient in the MS cohort	Inter-eye GCIPL[a]	0.77	68	77
	Inter-eye pRNFL[a]	0.74	71	65
	GCIPL thickness, average of both eyes	0.64	29	48
	pRNFL Thickness, average of both eyes	0.66	30	42
	Composite Score**	0.77	73	72
History of optic neuritis by eye	Inter-eye GCIPL	0.67	53	71
	Inter-eye pRNFL	0.62	50	65
	GCIPL Thickness	0.73	19	46
	pRNFL Thickness	0.73	27	36
	Composite Score***	0.72	58	77

AUC = area under the receiver operating characteristic curve; GCIPL = ganglion cell+inner plexiform layer; pRNFL = peripapillary retinal nerve fiber layer

*Composite score combined inter-eye GCIPL, average GCIPL thickness of both eyes, and binocular 2.5% low-contrast acuity.

**Composite score combined inter-eye GCIPL, overall GCIPL thickness, and inter-eye pRNFL measurements.

***Composite score combined inter-eye GCIPL, average GCIPL thickness of both eyes, overall pRNFL thickness, average pRNFL thickness of both eyes and age.

[a]This data has been previously published[2].

Table 4. Composite scores for MS disease classification model.

Classification	Composite Score
MS vs. Healthy Control	113 + 4*Inter-eye GCIPL – GCIPL thickness average of both eyes – Binocular LCLA
	133 + 4*Inter-eye GCIPL – GCIPL thickness average of both eyes – Binocular HCVA (ETDRS)
	70 + 4*Inter-eye GCIPL – GCIPL thickness average of both eyes – 4.5*Binocular HCVA (logmar)

MS=multiple sclerosis, GCIPL = ganglion cell+inner plexiform layer, HCVA=high-contrast visual acuity, LCLA=low-contrast letter acuity (letters correct out of 70), ETDRS= Early Treatment Diabetic Retinopathy Study (letters correct out of 70)

Figure 1. Patient Inclusion Diagram

Figure 1 Legend: This figure describes subjects assessed for eligibility in the study and shows excluded participants, and numbers included for each SD-OCT and visual acuity assessment.
anti-MOG = anti–myelin oligodendrocyte protein; GCIPL = ganglion cell + inner plexiform layer; HCVA = high-contrast visual acuity; LCVA = low-contrast visual acuity; NMO = neuromyelitis optica; OCT = optical coherence tomography; ON = optic neuritis; pRNFL = peripapillary retinal nerve fiber layer.
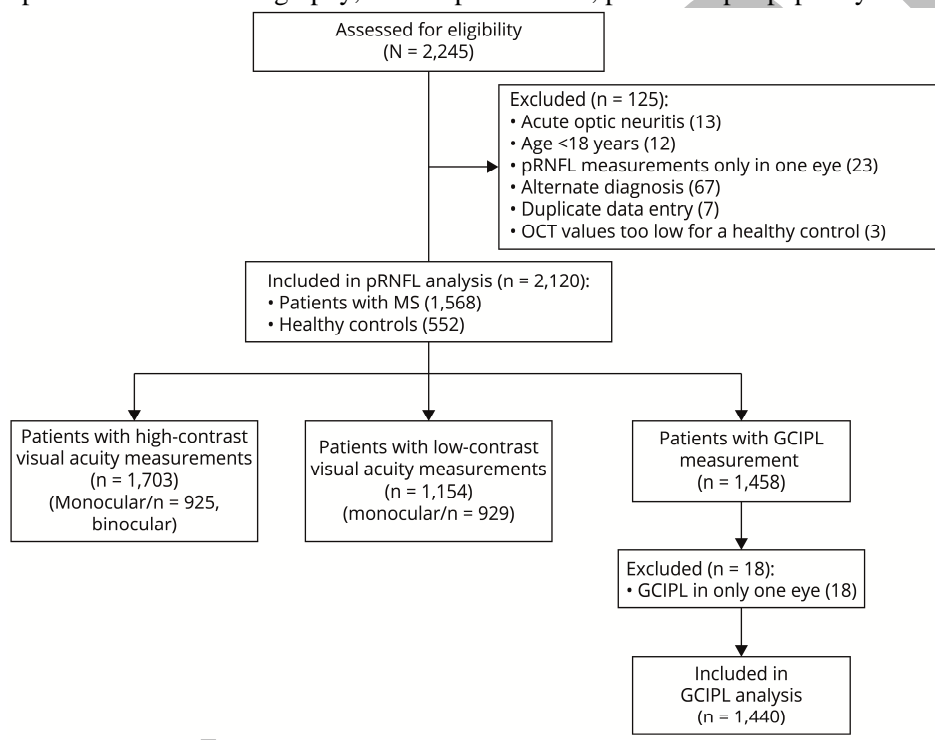
Figure 2. ROC Curve analysis for MS disease classification using composite score.

Figure 2 Legend: This figure illustrates the ROC curve analysis using the composite score (113 + 4*Inter-eye GCIPL – GCIPL thickness average of both eyes – Binocular LCLA) to classify PwMS from healthy controls. A composite score of 0 is the cut-point where sensitivity and specificity are optimized; a score greater than 0 is associated with a higher likelihood of MS disease status. The area under the ROC curve (0.89) demonstrates excellent discriminatory power. The ROC curves for the individual components of the composite score are plotted for comparison.
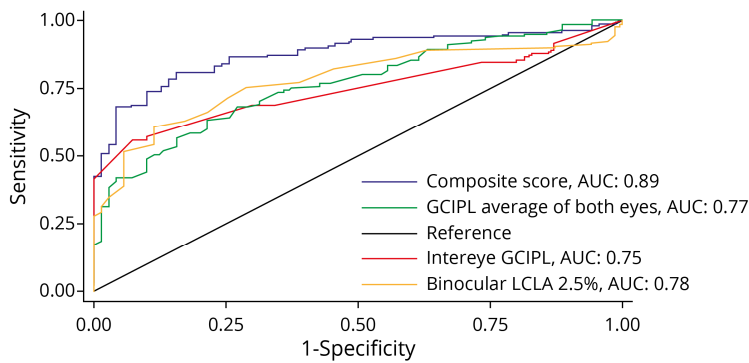

Figure 3. Boxplot of composite score for classification of MS diagnosis.

Figure 3 Legend: This figure illustrates a box plot for the composite score (113 + 4*Inter-eye GCIPL – GCIPL thickness average of both eyes – Binocular LCLA) for PwMS a healthy controls. A score of 0 indicates a higher likelihood of MS disease status (red reference line). The lower boundary of the box indicates the 25th percentile. The upper boundary of the box indicates the 75th percentile. The line within the box indicates the median value of the composite score for that group. Dots above the plot indicate outliers.

eTable1 -- http://links.lww.com/WNL/C156
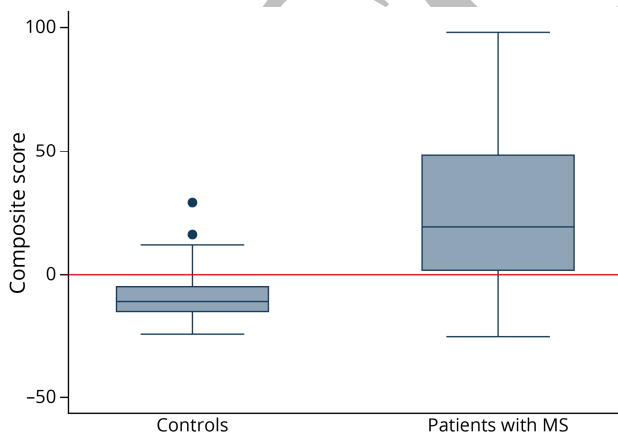
References

1.   Nolan RC, Galetta SL, Frohman TC, et al. Optimal Intereye Difference Thresholds in Retinal Nerve Fiber Layer Thickness for Predicting a Unilateral Optic Nerve Lesion in Multiple Sclerosis. Journal of neuro-ophthalmology : the official journal of the North American Neuro-Ophthalmology Society 2018.

2.   Nolan-Kenney RC, Liu M, Akhand O, et al. Optimal intereye difference thresholds by optical coherence tomography in multiple sclerosis: An international study. Ann Neurol 2019;85:618-629.

3.   Coric D, Balk LJ, Uitdehaag BMJ, Petzold A. Diagnostic accuracy of optical coherence tomography inter-eye percentage difference for optic neuritis in multiple sclerosis. Eur J Neurol 2017;24:1479-1484.

4.   Shigueoka LS, Vasconcellos JPC, Schimiti RB, et al. Automated algorithms combining structure and function outperform general ophthalmologists in diagnosing glaucoma. PloS one 2018;13:e0207784.

5.   Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. PloS one 2017;12:e0177726.

6.   Silva FR, Vidotti VG, Cremasco F, Dias M, Gomi ES, Costa VP. Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using Spectral Domain OCT and standard automated perimetry. Arq Bras Oftalmol 2013;76:170-174.

7.   Perez Del Palomar A, Cegonino J, Montolio A, et al. Swept source optical coherence tomography to early detect multiple sclerosis disease. The use of machine learning techniques. PloS one 2019;14:e0216410.

8.   Cavaliere C, Vilades E, Alonso-Rodriguez MC, et al. Computer-Aided Diagnosis of Multiple Sclerosis Using a Support Vector Machine and Optical Coherence Tomography Features. Sensors (Basel) 2019;19.

9.   Balcer LJ, Raynowska J, Nolan R, et al. Validity of low-contrast letter acuity as a visual performance outcome measure for multiple sclerosis. Multiple sclerosis (Houndmills, Basingstoke, England) 2017;23:734-747.

10.   Oberwahrenbrock T, Weinhold M, Mikolajczak J, et al. Reliability of Intra-Retinal Layer Thickness Estimates. PloS one 2015;10:e0137316.

11.   Tewarie P, Balk L, Costello F, et al. The OSCAR-IB consensus criteria for retinal OCT quality assessment. PLoS One 2012;7:e34823.

12.   Cruz-Herranz A, Balk LJ, Oberwahrenbrock T, et al. The APOSTEL recommendations for reporting quantitative optical coherence tomography studies. Neurology 2016;86:2303-2309.

13.   Aytulun A, Cruz-Herranz A, Aktas O, et al. APOSTEL 2.0 Recommendations for Reporting Quantitative Optical Coherence Tomography Studies. Neurology 2021;97:68-79.

14.   DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-845.

15.   Petzold A, Albrecht P, Balcer L, et al. Artificial intelligence extension of the OSCAR-IB criteria. Ann Clin Transl Neurol 2021;8:1528-1542.

16.   Gordis L. Epidemiology, Fifth edition. ed. Philadelphia, PA: Elsevier/Saunders, 2014.

Neurology®

This information is current as of June 28, 2022

Updated Information & Services	including high resolution figures, can be found at: http://n.neurology.org/content/early/2022/06/27/WNL.0000000000200883.full
Citations	This article has been cited by 1 HighWire-hosted articles: http://n.neurology.org/content/early/2022/06/27/WNL.0000000000200883.full##otherarticles
Subspecialty Collections	This article, along with others on similar topics, appears in the following collection(s): **Class III** http://n.neurology.org/cgi/collection/class_iii **Diagnostic test assessment** http://n.neurology.org/cgi/collection/diagnostic_test_assessment_ **Multiple sclerosis** http://n.neurology.org/cgi/collection/multiple_sclerosis **Optic nerve** http://n.neurology.org/cgi/collection/optic_nerve **Optic neuritis; see Neuro-ophthalmology/Optic Nerve** http://n.neurology.org/cgi/collection/optic_neuritis
Permissions & Licensing	Information about reproducing this article in parts (figures,tables) or in its entirety can be found online at: http://www.neurology.org/about/about_the_journal#permissions
Reprints	Information about ordering reprints can be found online: http://n.neurology.org/subscribers/advertise

AMERICAN ACADEMY OF NEUROLOGY®