BMJ Neurology Open

Classifying and quantifying changes in papilloedema using machine learning

Joseph Branco,[1] Jui-Kai Wang,[2,3,4] Tobias Elze,[5] Mona K Garvin,[2,4] Louis R Pasquale,[6,7] Randy Kardon,[2,3] Brian Woods,[7,8] David Szanto,[9] Mark J Kupersmith [10]

To cite: Branco J, Wang J-K, Elze T, *et al*. Classifying and quantifying changes in papilloedema using machine learning. *BMJ Neurology Open* 2024;**6**:e000503. doi:10.1136/bmjno-2023-000503

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjno-2023-000503).

Received 24 July 2023
Accepted 20 March 2024

For numbered affiliations see end of article.

Correspondence to
Dr Mark J Kupersmith; mark.kupersmith@mountsinai.org

ABSTRACT

Background Machine learning (ML) can differentiate papilloedema from normal optic discs using fundus photos. Currently, papilloedema severity is assessed using the descriptive, ordinal Frisén scale. We hypothesise that ML can quantify papilloedema and detect a treatment effect on papilloedema due to idiopathic intracranial hypertension.

Methods We trained a convolutional neural network to assign a Frisén grade to fundus photos taken from the Idiopathic Intracranial Hypertension Treatment Trial (IIHTT). We applied modified subject-based fivefold cross-validation to grade 2979 longitudinal images from 158 participants' study eyes (ie, the eye with the worst mean deviation) in the IIHTT. Compared with the human expert-determined grades, we hypothesise that ML-estimated grades can also demonstrate differential changes over time in the IIHTT study eyes between the treatment (acetazolamide (ACZ) plus diet) and placebo (diet only) groups.

Findings The average ML-determined grade correlated strongly with the reference standard (r=0.76, p<0.001; mean absolute error=0.54). At the presentation, treatment groups had similar expert-determined and ML-determined Frisén grades. The average ML-determined grade for the ACZ group (1.7, 95% CI 1.5 to 1.8) was significantly lower (p=0.0003) than for the placebo group (2.3, 95% CI 2.0 to 2.5) at the 6-month trial outcome.

Interpretation Supervised ML of fundus photos quantified the degree of papilloedema and changes over time reflecting the effects of ACZ. Given the increasing availability of fundus photography, neurologists will be able to use ML to quantify papilloedema on a continuous scale that incorporates the features of the Frisén grade to monitor interventions.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Original research using fundus photos taken from large studies has shown that fundus photos can be used as input in machine learning algorithms to characterise papilloedema in eyes affected by idiopathic intracranial hypertension. While studies have shown that machine learning can be used to distinguish normal eyes from those with papilloedema, no study has yet used artificial intelligence to assign an actual Frisén grade.

WHAT THIS STUDY ADDS

⇒ This study expands the application of machine learning by training a convolutional neural network to autonomously grade papilloedema in fundus photos using the ordinal Frisén scale. It also is the first to apply machine learning to detect a treatment effect of papilloedema due to idiopathic intracranial hypertension. The data show that machine learning can effectively be used as a tool to grade fundus photos of eyes with papilloedema without the requirement of an expert human grader.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The uses of machine learning to grade the extent of papilloedema in eyes with idiopathic intracranial hypertension are multifold. By tasking an artificial neural network to assign a Frisén grade to fundus photos, healthcare providers can more precisely track changes in papilloedema over time in response to treatment. Also, machine learning eliminates the burden of a reading centre in analysing and tracking papilloedema over time in large-scale studies. Finally, an artificial intelligence tool to assess papilloedema may make the assessment of idiopathic intracranial hypertension more feasible in remote settings or locations where an ophthalmologist may not be immediately available.

INTRODUCTION

Clinicians often use a descriptive ordinal scale, called the Frisén grade, to categorise and monitor papilloedema.[1 2] Recent work showed that machine learning (ML) can distinguish eyes with papilloedema from normal eyes.[3] To date, meaningful categorisation of the degree of papilloedema has been limited. Proper diagnosis of the presence and severity of papilloedema, whether due to idiopathic intracranial hypertension (IIH), a mass, hydrocephalus or dural venous sinus thrombosis, is essential for diagnosis, management and treatment. As severe Frisén grades are associated with a high risk of vision loss, the Frisén grade status is frequently used to decide the intensity of measures to reduce the elevated intracranial pressure.[4] Although ophthalmoscopy is not always performed and

expertise may be limited in emergency and neurology services,[5 6] the availability of non-mydriatic and handheld cameras capable of photographing the optic nerve head (ONH) should improve the capability of clinicians to view this important structure. ML may also be helpful when an accurate ophthalmoscopic diagnosis is not achievable or if other optic disc anomalies or abnormalities are present.[7–9]

Even though it remains the gold standard for grading papilloedema, the Frisén scale has limitations. First, it depends on the healthcare provider's recall and recognition of detailed features specific to each grade. Also, there is no set minimum number of features needed to definitively grade an optic disc by ophthalmoscopy or fundus photography. In clinical trials, where consensus on a Frisén grade is needed, typically two experts provide independent scores and a third expert serves to adjudicate the grade when there is no consensus. As currently used, the Frisén scale is an ordinal system that requires the evaluator to often over or undergrade papilloedema when features of more than one Frisén category are present.

In this study, we aimed to derive a supervised ML model capable of predicting Frisén grades from optic nerves photographed and classified in the Idiopathic Intracranial Hypertension Treatment Trial (IIHTT).[10] We hypothesised: (1) predicted Frisén grades would correlate with expert graded ground truths; (2) our ML model could identify a treatment effect using acetazolamide (ACZ) plus diet or placebo plus diet for the study eyes in the IIHTT; and (3) we could determine the earliest treatment effect of the Frisén grade by our ML model predictions and by expert grading.

METHODS

The methodologies used for every aspect of the National Eye Institute of the National Instute of Health-sponsored Neuro-Ophthalmology Disease Investigator Research Disease Consortium (NORDIC) IIHTT (NCT01003639) are published.[10 11] The IIHTT enrolled 165 participants with IIH (161 women and 4 men), ages 18–52 years, naïve to intervention and with a study eye defined as the eye with the worse perimetric mean deviation (PMD) of −2.00 to −7.00 dB. Recruitment began 1 March 2010, and the study was completed 30 June 2013 at 38 certified NORDIC sites in the USA and Canada. Stereoscopic digital photographs centred on the optic disc and macula were collected using certified personnel and photographic equipment from multiple sites.[12] At enrolment, participants were randomised to either ACZ or placebo and both were provided weight management. Site investigators and technicians were masked to study group assignments. The average age for the ACZ and placebo groups was similar (29.1 (range 18–46, SD 7.5) vs 28.3 (range 19–53, SD 8.0), p=0.72). Both groups contained only two men each. The average body mass index (BMI) for the ACZ and placebo groups was similar (40.0 (range 24.9–61.5, SD 8.5) vs 39.9 (range 26.9–71.2, SD 8.1), p=0.46). The percentage of overweight patients (BMI>30) was similar as well (11 patients (12.8%) for the ACZ group vs 10 patients (12.5%) for the placebo group, p=0.96). Both groups were racially diverse (online supplemental table 1). The trial results and outcome of PMD at 6 months or at treatment failure have been reported.[10] Patients with IIH from the University of Iowa (10 women and 2 men), ages 23–56, provided photos of Frisén grades 4 and 5 to supplement the low number of these grades found in the IIHTT.

This study and the analysis of fundus photos from the IIHTT were granted a waiver of informed consent as the data were collected with informed consent for the trial, and were approved by the Institutional Review Board (IRB) of the Icahn School of Medicine at Mount Sinai. All patients included in the data set from the University of Iowa signed an informed consent document for the use of their photos for research, which was approved by the IRB of the University of Iowa. This study was conducted in accordance with the regulations established by the Helsinki declaration.

Expert grading of fundus photos

Three neuro-ophthalmologists, masked to the treatment, used the modified Frisén grading system[2] to grade the fundus photos prior to this study. These grades were used as the reference, or expert-determined grades.

Fundus photos included in the data set

The entire data set included 5553 ONH-centred fundus photos taken of both eyes of 158 participants (316 eyes) enrolled in the IIHTT and 355 ONH-centred fundus photos, taken over approximately 12 months, of both eyes of eight de-identified patients (16 eyes) from the University of Iowa, for a total of 5908 fundus photos. Photos from the seven cases of treatment failure in the IIHTT were unavailable. We excluded photos from four patients from the University of Iowa due to optic atrophy. The majority of the fundus photos from the IIHTT were taken at monthly time points between the baseline visit and 6-month trial outcome, but some participants had photos taken at later time points after the outcome. Not all participants had photos taken at all monthly time points. The fundus photos from the University of Iowa were taken at the presentation and follow-up visits up to 1 year after the presentation. These photos were assigned ground truth Frisén grades by RK and MJK, and IIH was confirmed in these patients by lumbar puncture at presentation. All fundus photos from the IIHTT and the University of Iowa were obtained from dilated eyes, and no eyes had optic disc anomalies or other abnormalities. Fundus photos of left eyes were horizontally flipped to a right eye orientation, and all fundus photos were analysed in a right eye orientation. Photos were flipped to minimise unnecessary variability in this small data set, and training the model on flipped photos increased the accuracy of the model.

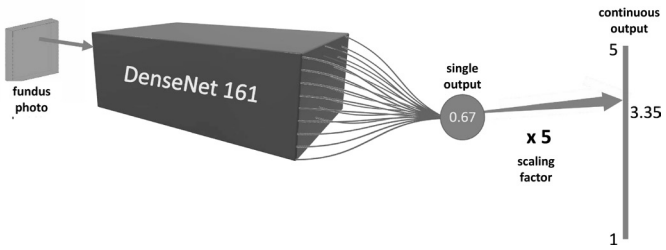

Figure 1 Diagram of our modified DenseNet-161 network, in which we used original fundus photos as input. To convert the original architecture into a regression, we replaced the 1000 outputs from the original DenseNet-161 classification model with a single output that we scaled by a factor of five to produce a continuous reading between one and five.

Between one and four photos were taken of each study eye at each time point.

Training and validation of the network

To predict the Frisén grade from fundus photos, we used a supervised convolutional neural network (CNN) based on a DenseNet-161 architecture[13] with several modifications. We first scaled image grades from 1 to 5 down to between 0 and 1. These scaled images are then entered into the model. Instead of using the pretrained output layer of DenseNet-161 as the final layer in our model, we replaced it with a linear layer. We then added a sigmoid layer and scaled the final output by a factor of five to match the established range of Frisén grades. The final output for each image in the validation set was a linear value between 0 and 5. We trained the network using fundus photos labelled with their expert-determined grades, and then had the network calculate ML-determined grades for fundus photos included in the validation set. A diagram of our CNN is shown below in figure 1. We trained the network using five different training/validation subset combinations as shown in online supplemental table 2.

We applied modified fivefold cross-validation in the neural network training and validation process. All 158 IIHTT subjects were split into five groups (groups I, II, III, IV and V); each subject in each group had multiple fundus photos over time from the study eye (with the worse MD) and the fellow eye (with the better MD). For the first fold validation, the neural network was trained using photos from IIHTT groups I, II, III and IV and photos from the University of Iowa. Then, the neural network was validated by using the photos of the study eyes from the untouched IIHTT group (group V). For the second fold validation, we then trained the neural network using the images from groups I, II, III and V and photos from the University of Iowa. Then, we validated the neural network using images of the study eyes from IIHTT group IV. We repeated this process until all the images of the study eyes from IIHTT groups I, II, III, IV and V were validated. It is important to note that (1) all the images from Iowa were always in the training set, and (2) the five IIHTT groups were divided by subjects, so the images in the validation set were always 'unseen' to the neural network in each fold during the cross-validation. The image numbers from the training and validation set in each fold are shown in online supplemental table 2.

Analysis of the validation set

The ML-determined Frisén grading of the validation set photos produced a numerical output on a continuous scale from <1 to 5, carried out to the 10th decimal place. We determined the correlation between the ML-determined Frisén grade values and the expert-determined ground truth grades using Spearman's rank correlation. We plotted the distribution of ML-determined Frisén grades for each ground truth value using box plots. We calculated the root mean squared error (RMSE) for predicted Frisén grade values. We determined the accuracy of our model by calculating F1 scores for fundus photos of each expert-determined Frisén grade.

Determination of treatment effect of acetazolamide plus diet versus placebo plus diet

We compared the distribution of baseline visit fundus photos by Frisén grade using the expert-determined and ML-determined Frisén grades between intervention groups using χ^2 analysis. We stratified the study eyes into each of the two intervention groups and calculated the average expert-determined and ML-determined Frisén grade at 1, 2, 3, 4, 5 and 6 months of follow-up for each group. For eyes with multiple photos taken at the same time point, we averaged the ML-predicted Frisén grades for those photos to calculate an average ML-predicted Frisén grade for each eye. We omitted data collected at 4 months from this calculation due to a small number of fundus photos collected at this time point. We omitted baseline photos from three participants due to poor photographic quality. We determined photos to be of poor quality if they were under or overexposed, were out of focus, or contained artefact. The number of photos and eyes included at each monthly time point are shown in table 1.

We explored how quickly the change in Frisén grade would show a reduction in papilloedema based on the intervention[10] using the Mann-Whitney U test, as there was not an equal number of photos available at each time point. We also correlated the expert-determined and ML-determined Frisén grade of fundus photos with the lumbar puncture opening pressure values for the IIHTT participants collected at baseline and at the 6-month outcome in the IIHTT using Spearman's rank correlation.

Determination of whether intermediate ML-determined Frisén grades reflect features of two grades

We attempted to determine whether ML-determined Frisén grades between two consecutive grade values indicated whether fundus photos contained features belonging to more than one Frisén category. For this analysis, we randomly chose 33 fundus photos with ML-determined values between two Frisén grades and masked their Frisén grade values. One author (MJK) then graded

Table 1 Number of photos and eyes used to calculate the average Frisén grade at each time point for the treatment group (acetazolamide (ACZ)+diet) and placebo group (placebo+diet) using both the expert-determined and machine learning-determined Frisén grades.

Month	ACZ+diet	Placebo+diet
0	275 (79 eyes)	280 (76 eyes)
1	239 (59 eyes)	240 (60 eyes)
2	197 (48 eyes)	205 (51 eyes)
3	188 (45 eyes)	176 (44 eyes)
4*	72 (19 eyes)	88 (21 eyes)
5	115 (26 eyes)	128 (31 eyes)
6	158 (49 eyes)	160 (50 eyes)
Total	**1244 (80 unique eyes)**	**1277 (78 unique eyes)**

*Data from month 4 were omitted due to the small sample size.

the photos according to the following rule: if a photo contained features of only one Frisén category, the photo was given that grade and if the photo contained features of two grades, the photo was given the average of the two grades. We rounded ML-determined Frisén grades for the 33 fundus photos to 0.5 intervals for comparison to Frisén grades assigned by MJK.

Determination of fundus photo regions of high gradient used in ML-determined Frisén grade calculation

We used SmoothGrad[14] to apply activation mapping to the fundus photos to reveal the regions our model used to calculate a Frisén grade. Areas of varying gradients indicated photographic regions of high weight and low weight in determining the predicted numerical Frisén grade.

RESULTS

Characterisation of ML-determined Frisén grades of fundus photos in the validation set

The expert-determined and ML-determined Frisén grades for study eyes were distributed from grades 1 to 5, with the fewest photos around grade 5. The ML-determined Frisén grades for study eyes correlated strongly with ground truths (r=0.76, p<0.001; RMSE=0.68). The median ML-determined values were lower than each ground truth value for all Frisén grades except for grade 1 (figure 2). F1 values (a measure of accuracy) for each Frisén grade are shown in table 2. The confusion matrix showing the distribution of ML-determined Frisén grades for fundus photos of each of the five expert-determined Frisén grades is shown in table 3.

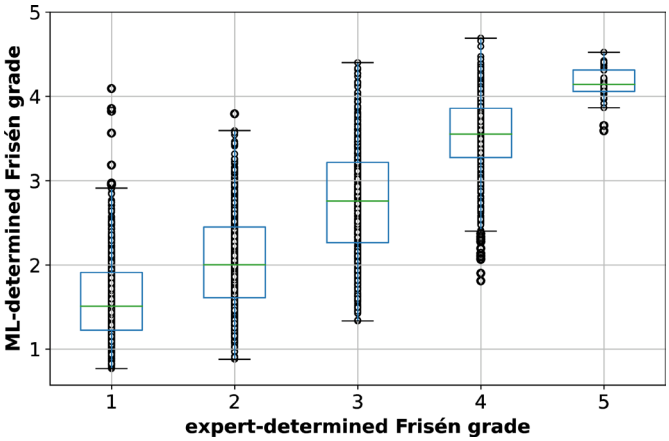

Figure 2 Box plots showing a range of machine learning-determined Frisén grades versus each expert-determined ground truth for fundus photos included in the validation set. ML, machine learning.

Determination of treatment effect of acetazolamide plus diet versus placebo plus diet

The distribution of expert-determined and ML-determined Frisén grades for study eye photos was similar between both intervention groups at enrolment (table 4).

The average expert-determined Frisén grade decreased significantly for study eyes treated with ACZ (change=−1.12, p<0.001) relative to placebo (change=−0.68, p<0.001) at the 6-month outcome. The average expert-determined Frisén grade of study eyes treated with ACZ (1.7, 95% CI 1.5 to 1.9) was lower than for study eyes treated with placebo (2.1, 95% CI 1.8 to 2.4) at 6 months (p=0.09). The average expert-determined Frisén grade was significantly lower for study eyes treated with ACZ relative to placebo at months 1, 3 and 5 (table 5, figure 3).

The average ML-determined Frisén grade decreased significantly for study eyes treated with ACZ (change=−1.03, p<0.001) relative to placebo (change=−0.56, p<0.001) at the 6-month outcome. The average ML-determined Frisén grade of study eyes treated with ACZ (1.7, 95% CI 1.5 to 1.8) was significantly lower than for study eyes treated with placebo (2.3, 95% CI 2.0 to 2.5) at 6 months (p=0.0003). The average ML-determined Frisén grade was also significantly lower for study eyes treated with

Table 2 F1 scores for ML-determined Frisén grades of fundus photos of each expert-determined Frisén grade

Expert-determined Frisén grade	F1 score for ML-determined Frisén grade
1	0.55
2	0.55
3	0.47
4	0.59
5	0.04
weighted average	**0.53**

ML, machine learning.

Table 3 Confusion matrix showing the distribution of machine learning (ML)-determined grades for fundus photos of each expert-determined grade

	Expert-determined grade				
ML-determined grade	**1**	**2**	**3**	**4**	**5**
1	386	200	16	0	0
2	372	632	225	21	0
3	32	232	335	169	0
4	4	5	77	230	39
5	0	0	0	3	1

ACZ relative to placebo at months 1, 2, 3 and 5 (table 5, figure 4).

The lumbar puncture opening pressure in 141 participants correlated with the expert-determined (r=0.31, p<0.001) and ML-determined (r=0.32, p<0.001) Frisén grades at baseline. The opening pressure in 58 participants at 6 months correlated with the ML-determined (r=0.39, p=0.002) but not expert-determined (r=0.04, p=0.77) Frisén grades at the 6-month outcome in the IIHTT.

Determination of whether intermediate ML-determined Frisén grades reflect features of two grades

The Frisén grade assigned by the masked observer and predicted by ML agreed for 29/33 (82%) fundus photos and differed by one Frisén grade for 4/33 (18%) fundus photos. Of the four photos that differed by one Frisén grade, the masked observer graded two as below and two as above the ML-determined Frisén grades. Frisén grades assigned by the masked observer and predicted by ML did not differ by more than one Frisén grade for any of the 33 fundus photos. The masked observer determined that 18/33 (55%) of fundus photos contained features belonging to more than one Frisén grade description. All the photos had principal features only of the adjacent two categorical Frisén grades and ML misclassified none.

Determination of fundus photo regions of high gradient used in ML-determined Frisén grade calculation

Activation mapping revealed that our ML model placed a high gradient over the region of the ONH and the peripapillary retina to produce a Frisén grade determination (figure 5).

DISCUSSION

This study shows that with a relatively small number of fundus photos of the optic disc region, supervised ML can satisfactorily classify papilloedema into Frisén grades. Furthermore, the ML-determined grading value for each optic disc can be expressed as a continuous variable, which simplifies potential correlations with optical coherence tomography (OCT) derived values,[15–19] visual performance measures and cerebrospinal fluid pressure. As the ML-determined grades appear to recognise characteristics from more than one grade, the clinician or evaluator may not need to decide between grades when a photo contains features for more than one grade. Additionally, ML-determined grading may be more sensitive to either worsening or improvement of papilloedema as descriptive grading requires a change of at least an entire grade. ML should help identify cases at risk for treatment failure or worse visual field deficit if the ML-determined grading can detect small changes leading to grade 4 and

Table 4 Number of eyes of each Frisén grade (1–5) at baseline as expert-graded and ML-predicted separated by treatment group with acetazolamide (ACZ)+diet or placebo+diet. ML-predicted Frisén grades were rounded to the nearest integer. Expert grading is based on the modified Frisén scale

Expert-determined				ML-determined			
	# of study eyes				# of study eyes		
Frisén grade	ACZ+diet (n=79)	Placebo+diet (n=76)	P value	Frisén grade	ACZ+diet (n=79)	Placebo+diet (n=76)	P value
1	12 (15%)	10 (14%)	0.86	1	13 (16%)	7 (9%)	0.19
2	26 (33%)	20 (32%)	0.89	2	18 (22%)	25 (33%)	0.13
3	15 (19%)	18 (24%)	0.45	3	29 (37%)	25 (33%)	0.60
4	24 (30%)	20 (26%)	0.58	4	18 (23%)	19 (25%)	0.77
5	2 (2%)	3 (4%)	0.47	5	0 (0%)	1 (1%)	0.37

ML, machine learning.

Table 5 Average expert-determined and machine learning (ML)-determined Frisén grade over time by treatment group with 95% CI

	Expert-determined Frisén grade					ML-determined Frisén grade				
	ACZ+diet (n=1309)		Placebo+diet (n=1278)			ACZ+diet (n=1309)		Placebo+diet (n=1278)		
Months	Average	95% CI	Average	95% CI	P value	Average	95% CI	Average	95% CI	P value
0	2.9	2.6 to 3.1	2.8	2.5 to 3.1	0.78	2.7	2.5 to 2.9	2.8	2.6 to 3.0	0.43
1	2.2	1.9 to 2.5	2.7	2.4 to 3.0	0.03	2.3	2.1 to 2.5	2.7	2.5 to 2.9	0.02
2	2.1	1.8 to 2.3	2.4	2.1 to 2.7	0.09	2.1	1.9 to 2.3	2.5	2.3 to 2.7	0.01
3	1.7	1.5 to 1.9	2.4	2.1 to 2.7	0.01	1.8	1.6 to 1.9	2.5	2.2 to 2.7	0.0002
5	1.8	1.5 to 2.1	2.3	2.0 to 2.7	0.05	1.8	1.7 to 2.0	2.3	2.1 to 2.6	0.04
6	1.7	1.5 to 1.9	2.1	1.8 to 2.4	0.09	1.7	1.5 to 1.8	2.3	2.0 to 2.5	0.0003

ACZ, acetazolamide.

5 papilloedema. The identification of swelling or Frisén feature reduction, particularly in the ACZ intervention group, is evidence of the potential utility of ML grading to monitor the effectiveness of therapy. These results mirror the ACZ treatment effect at the IIHTT 6-month outcome previously reported for Frisén grading for lay readers[10] and OCT.[20]

The IIHTT did not report whether a treatment effect was found using the Frisén grade at time points earlier than 6 months. In our present study, we detected an ACZ treatment effect in the expert-graded group beginning at 1 month and at all time points except 2 months. This suggests that improvement of papilloedema in patients treated with ACZ should be expected early, after the initiation of therapy. We also detected an ACZ treatment effect using the ML-determined Frisén grades beginning at 1 month and at all subsequent time points. The ability of our algorithm to detect a treatment effect could be useful in future clinical trials and reduce the need for expert descriptive grading.

Though OCT measurements correlate with the Frisén grade[15 16] and provide continuous values of thickness of the retinal nerve fibre layer (RNFL), the OCT machine algorithms do not reliably measure RNFL thickness when severe optic disc swelling is present. Furthermore, the features of the Frisén grade that may indicate a risk for vision loss are not identified by OCT. No prospective study of IIH has shown that an OCT measurement of the ONH region is a predictor of vision loss.

Milea *et al* used supervised ML to distinguish fundus photos of papilloedema from normal optic nerves.[3] This study pooled fundus photos of all Frisén grades into one training set class, such that only the presence or absence, but not the degree of papilloedema could be detected. Vasseneix *et al* used this method to distinguish severe from mild papilloedema, in a study designed to identify patients of high and low risk.[21] Though these two papers essentially opened the field for using ML to evaluate ONH swelling, neither study trained an artificial neural network to assign a numerical Frisén grade to fundus

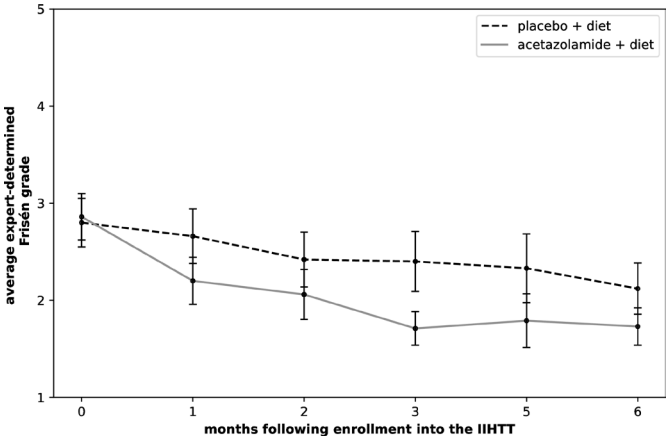

Figure 3 Average expert-determined Frisén grade values for placebo plus diet (dashed line) and acetazolamide plus diet (solid line) groups over time with 95% CIs at each time point. Data collected at 4 months were omitted due to a small sample size. IIHTT, Idiopathic Intracranial Hypertension Treatment Trial.

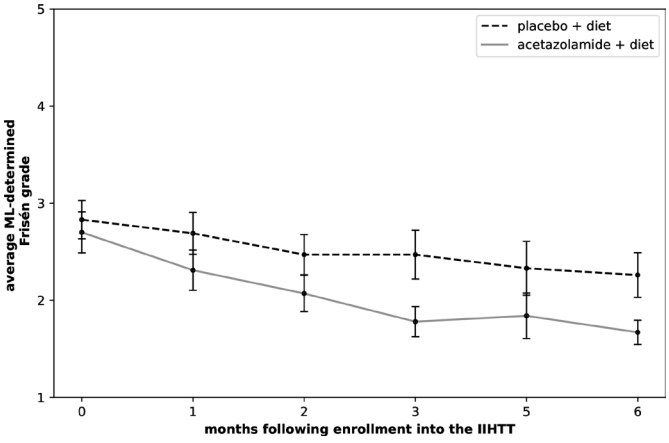

Figure 4 Average machine learning (ML) predicted Frisén grade values for placebo+diet (dashed line) and acetazolamide+diet (solid line) groups over time with 95% CIs at each time point. Data collected at 4 months were omitted due to a small sample size. IIHTT, Idiopathic Intracranial Hypertension Treatment Trial.

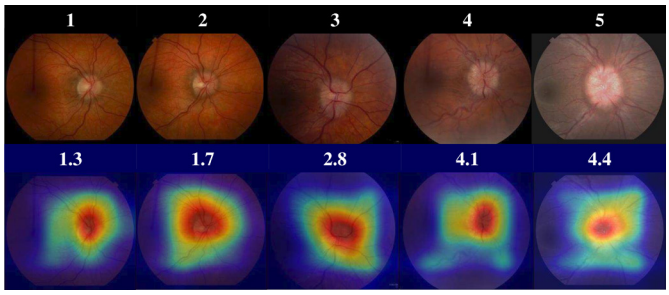

Figure 5 Activation mapping of fundus photos of the optic nerve region, with expert-determined ground truth Frisén grades (top row) and machine learning-determined Frisén grades (bottom row). Red-shifted regions correspond to areas of high gradient and blue-shifted regions correspond to areas of low gradient.

photos, a feature that would be required for the rating and monitoring of papilloedema in practice and clinical trials. Additionally, supervised ML to distinguish papilloedema from normal optic discs using fundus photos required over 14 000 images.[3] The features in each photo that contributed to the determination of the presence or absence of papilloedema were not known.[3] For our study, activation mapping revealed the ML model focused on the region of the ONH and peripapillary retina to assign a Frisén grade prediction. This algorithm is consistent with the established Frisén grade model which is based on specific characteristics located in the peripapillary retina and optic disc but may also extend it to features that might provide a more refined grading scheme.

As papilloedema is a sign of increased intracranial pressure, at times associated with a life-threatening illness or cause of severe vision loss or both, early intervention and reliable consistent monitoring are essential. Given the advancing technology of hand-held fundus digital photography, neurologists, ophthalmologists and even expert neuro-ophthalmologists will be more able to diagnose, document and track papilloedema. This should also help clinicians who may have less experience and training in ophthalmoscopy.[22] With future improvements to our current model facilitated by the expansion of the input data set to include more photos, a more accurate model can be used to autonomously grade fundus photos in scenarios where ophthalmic imaging departments are not available, such as in the treatment of bedridden patients or individuals with altered mental status. An autonomous method to grade photos will also be invaluable in the emergency room and intensive care unit settings. ML-determined Frisén grading of fundus photos may reduce the need for having to recall the detailed descriptive scale and expert observation by an individual skilled in ophthalmoscopy.

The applicability of our model may be limited by the inclusion of only individuals with mild vision loss in the IIHTT; however, due to the standardised manner in which fundus photos are taken, we expect our model to be generalisable to photos taken from individuals of different demographic characteristics than those included in the IIHTT. Our ML methodology will be tested to make sure that the images from other photography platforms can be analysed in a similar manner.

The updated expert Frisén grading was performed to improve the lay reader assignments by neuro-ophthalmologists, masked to the intervention, for an earlier study correlating the grade with disease severity.[4] The expert-determined and ML-determined Frisén grade values had a stronger correlation with the lumbar puncture opening pressure than did OCT measurements of the RNFL, ONH volume or total retinal thickness in the peripapillary region.[17]

There are limitations to our study. We did not have a normal control set containing fundus photos without papilloedema, as the intention of the study was not to develop a model to distinguish the many disorders that can elevate the ONH (often termed pseudopapilloedema). Therefore, our model cannot distinguish papilloedema from other causes of optic disc swelling or elevation such as optic neuritis or buried drusen in the ONH. However, this did not affect our ability to detect a treatment effect, which is the major goal of our study. In the future, if we wish to improve the accuracy of our model in assigning an ML-determined grade to a de novo fundus photo, we will need to include normal control data in our training set. Second, we did not include an external test set in our analysis because of the limited number of subjects in the IIHTT and therefore do not know if our model is generalisable beyond the IIHTT cohort. Furthermore, we had relatively few photos of eyes with grade 4 and 5papilloedema, even with the supplementation of our data set. This may explain why our F1 score for grade 5 photos was low; including more photos of eyes with grade 5 papilloedema will help improve our model's accuracy in the future. Because the IIHTT had a small number of participants, we used the IIHTT study eyes for the validation set as well as the ML predictions to show the ACZ treatment effect. Fortunately, the results were similar to the conventional Frisén grade results. Even though the ML-determined grades tended to be slightly lower than the expert grades, the capability of ML to detect clinically relevant changes in papilloedema is paramount. We anticipate that expanding our model by adding more cases should make the two methods coincide unless ML use of feature detection is actually more accurate. Additionally, we had a small number of IIHTT and clinical patient photos to use as the training set for our ML model. We do not know if the results apply to patients with long-standing papilloedema or atrophic papilloedema, as the IIHTT included only patients with untreated and likely early disease. We have not compared the ML-predicted Frisén grade model to the OCT results reported in the IIHTT because the latter was a substudy conducted on only 125 participants with only three visits during which data were collected.[16 17 20] Although the activation maps show a robust contribution to our model from the region of the ONH and peripapillary retina, the precise features for each grade assignment remain unknown.

One potential future endeavour involves the automated identification of significant visible image features associated with papilloedema in colour fundus photographs[23 24] accompanied by the predicted Frisén grade. The forthcoming system aims to not only output a single Frisén grade prediction for each input colour fundus photograph but also provide a feature list with estimated importance as supporting evidence for the grade prediction. Additionally, our efforts will encompass the estimation of some OCT-related parameters (such as optic disc volume, peripapillary RNFL thickness).[25] Although the process of gathering labels will be substantial, with continued efforts to expand our input data set and further develop our training model, we believe that our papilloedema assessment system will prove highly valuable in emergency rooms and primary care offices in the future.

Author affiliations
[1]New York Medical College, Valhalla, New York, USA
[2]Center for the Prevention and Treatment of Visual Loss, Iowa City VA Healthcare System, Iowa City, Iowa, USA
[3]Department of Ophthalmology and Visual Sciences, The University of Iowa Hospitals and Clinics, Iowa City, Iowa, USA
[4]Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa, USA
[5]Schepens Eye Research Institute, Harvard Medical School, Boston, MA, USA
[6]Ophthalmology, Icahn School of Medicine at Mount Sinai, New York, New York, USA
[7]New York Eye and Ear Infirmary of Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[8]Ophthalmology, Cork University Hospital, Cork, Ireland
[9]Renaissance School of Medicine, Stony Brook, New York, USA
[10]Neurology, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Contributors JB collected data and performed; ML, coauthored manuscript. TE guided ML and coauthored manuscript. MKG guided ML and coauthored manuscript. RK guided ML and coauthored manuscript. LRP guided ML and coauthored manuscript. DS guided ML and coauthored manuscript. J-KW guided ML and coauthored manuscript. BW guided ML and coauthored manuscript. MJK provided all data, supervised entire project, secured funding and coauthored manuscript, is the guarantor, accepts full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Funding The New York Eye and Ear Infirmary Foundation, New York, New York; NEI EY032522; Research to Prevent Blindness, Inc., New York, New York, unrestricted grant to the Department of Ophthalmology; NEI R01 EY015473; Department of Veteran Affairs (VA) Center for the Prevention and Treatment of Visual Loss, Rehabilitation Research and Development (RR&D) I50 RX003002; VA RR&D I01RX003797.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data may be obtained from a third party and are not publicly available. Photos used for ML can be accessed for reasonable research.

ORCID iD
Mark J Kupersmith http://orcid.org/0000-0003-0461-8839

REFERENCES

1 Frisén L. Swelling of the optic nerve head: a staging scheme. *J Neurol Neurosurg Psychiatry* 1982;45:13–8.
2 Scott CJ, Kardon RH, Lee AG, *et al*. Diagnosis and grading of papilledema in patients with raised intracranial pressure using optical coherence tomography vs clinical expert assessment using a clinical staging scale. *Arch Ophthalmol* 2010;128:705–11.
3 Milea D, Najjar RP, Zhubo J, *et al*. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med* 2020;382:1687–95.
4 Wall M, Falardeau J, Fletcher WA, *et al*. Risk factors for poor visual outcome in patients with idiopathic intracranial hypertension. *Neurology* 2015;85:799–805.
5 Roberts E, Morgan R, King D, *et al*. Funduscopy: a forgotten art? *Postgrad Med J* 1999;75:282–4.
6 Bruce BB, Thulasi P, Fraser CL, *et al*. Diagnostic accuracy and use of nonmydriatic ocular fundus photography by emergency physicians: phase II of the FOTO-ED study. *Ann Emerg Med* 2013;62:28–33.
7 Miller N, Newman N. *Walsh and Hoyt's clinical neuro-ophthalmology*. Baltimore, MD: Williams and Wilkins, 1998:1800–16.
8 Biousse V, Najjar R, Sathianvichitr K, *et al*. Deep learning can accurately distinguish between true papilledema and optic disc Drusen on ocular fundus photographs. *Neurology* 2022;98.
9 Echegaray S, Zamora G, Yu H, *et al*. Automated analysis of optic nerve images for detection and staging of papilledema. *Invest Ophthalmol Vis Sci* 2011;52:7470–8.
10 Wall M, McDermott MP, Kieburtz KD, *et al*. Effect of acetazolamide on visual function in patients with idiopathic intracranial hypertension and mild visual loss: the idiopathic intracranial hypertension treatment trial. *JAMA* 2014;311:1641–51.
11 Friedman DI, McDermott MP, Kieburtz K, *et al*. The idiopathic intracranial hypertension treatment trial: design considerations and methods. *J Neuroophthalmol* 2014;34:107–17.
12 Fischer WS, Wall M, McDermott MP, *et al*. Photographic reading center of the idiopathic intracranial hypertension treatment trial (IIHTT): methods and baseline results. *Invest Ophthalmol Vis Sci* 2015;56:3292–303.
13 Huang G, Liu Z, Pleiss G, *et al*. Convolutional networks with dense connectivity. 2019. *IEEE Trans Pattern Anal Mach Intell* 2019;44:8704–16.
14 Smilkov D, Thorat N, Kim B, *et al*. Smoothgrad: removing noise by adding noise. *ArXiv* 2017.
15 Wang JK, Kardon RH, Kupersmith MJ, *et al*. Automated Quantification of volumetric optic disc swelling in papilledema using spectral-domain optical coherence tomography. *Invest Ophthalmol Vis Sci* 2012;53:4069–75.
16 OCT Sub-Study Committee for the NORDIC IIHTT Study Group. Baseline optical coherence tomography (OCT) of participants in the idiopathic intracranial hypertension treatment trial: correlations and relationships to clinical features. *Invest Ophthalmol Vis Sci* 2014;55:3543.
17 Auinger P, Durbin M, Feldon S, *et al*. Baseline OCT measurements in the idiopathic intracranial hypertension treatment trial, part II: correlations and relationship to clinical features. *Invest Ophthalmol Vis Sci* 2014;55:8173–9.
18 Sheils CR, Fischer WS, Hollar RA, *et al*. The relationship between optic disc volume, area, and Frisén score in patients with idiopathic intracranial hypertension. *Am J Ophthalmol* 2018;195:101–9.
19 Echegaray S, Zamora G, Luo W, *et al*. Automated classification of papilledema using Frisen grading and OCT measurements. *Invest Ophthalmol Vis Sci* 2010;51:1775.
20 Optical Coherence Tomography Substudy Committee, NORDIC Idiopathic Intracranial Hypertension Study Group. Papilledema outcomes from the optical coherence tomography substudy of the idiopathic intracranial hypertension treatment trial. *Ophthalmology* 2015;122:1939–45.
21 Vasseneix C, Najjar RP, Xu X, *et al*. Accuracy of a deep learning system for classification of papilledema severity on ocular fundus photographs. *Neurology* 2021;97:e369–77.

22 Bruce BB. Nonmydriatic ocular fundus photography in the emergency department: how it can benefit Neurologists. *Semin Neurol* 2015;35:491–5.
23 Hernan A, Perdomo O, Daza L, *et al*. Unsupervised method to cluster color fundus eye images and text reports from patients with diabetic retinal lesions. *Invest Ophthalmol Vis Sci* 2020;61:2047.
24 Islam M, Wang J, Deng W, *et al*. Deep learning-based estimation of 3d optic nerve head shape from 2d color fundus photographs in cases of optic disc swelling. In: *Ophthalmic medical image analysis*. 2020.
25 Scott CJ, Kardon RH, Lee AG, *et al*. Diagnosis and grading of papilledema in patients with raised intracranial pressure using optical coherence tomography vs. clinical expert assessment using a clinical staging scale. *Arch Ophthalmol* 2010;128:705–11.

Race	ACZ + diet (n=85)	Placebo + diet (n=80)	P-value
White	47 (55.3%)	49 (61.3%)	0.44
Black	24 (28.2%)	16 (20.0%)	0.22
Hispanic	3 (3.5%)	6 (7.5%)	0.26
Asian	1 (1.2%)	0	0.33
Indian	1 (1.2%)	1 (1.3%)	0.95
other/more than one race	9 (10.6%)	8 (10.0%)	0.90
Sex			
Female	83 (97.6%)	78 (97.5%)	0.97
Male	2 (2.4%)	2 (2.5%)	0.97

Supplement Table 1: Demographic data for the treatment group (acetazolamide [ACZ] + diet) and placebo group (placebo + diet)

Run	Frisén Grade	Training Set	Validation Set
1	1	1390 (147 eyes)	168 (20 eyes)
	2	1704 (170 eyes)	208 (19 eyes)
	3	1017 (117 eyes)	108 (15 eyes)
	4	585 (88 eyes)	99 (13 eyes)
	5	94 (17 eyes)	8 (1 eye)
	Total	**4790 (270 unique eyes)**	**591 (31 unique eyes)**
2	1	1213 (146 eyes)	215 (19 eyes)
	2	1620 (169 eyes)	277 (21 eyes)
	3	1056 (117 eyes)	91 (14 eyes)
	4	647 (90 eyes)	70 (11 eyes)
	5	86 (17 eyes)	8 (1 eye)
	Total	**4622 (268 unique eyes)**	**661 (32 unique eyes)**
3	1	1433 (152 eyes)	130 (15 eyes)
	2	1655 (168 eyes)	237 (22 eyes)
	3	977 (118 eyes)	143 (18 eyes)
	4	568 (96 eyes)	90 (12 eyes)
	5	86 (17 eyes)	12 (1 eye)
	Total	**4719 (268 unique eyes)**	**612 (32 unique eyes)**
4	1	1265 (143 eyes)	184 (18 eyes)
	2	1731 (168 eyes)	188 (21 eyes)
	3	997 (118 eyes)	131 (12 eyes)
	4	680 (96 eyes)	52 (10 eyes)
	5	94 (16 eyes)	8 (2 eyes)
	Total	**4767 (268 unique eyes)**	**563 (32 unique eyes)**
5	1	1515 (153 eyes)	97 (17 eyes)
	2	1766 (165 eyes)	159 (21 eyes)
	3	865 (107 eyes)	180 (18 eyes)
	4	569 (84 eyes)	112 (15 eyes)
	5	94 (17 eyes)	4 (1 eye)
	Total	**4809 (270 unique eyes)**	**552 (31 unique eyes)**
	GRAND TOTAL	**5908 unique photos (332 unique eyes)**	**2979 (158 unique eyes)**

Supplemental Table 2**:** Distribution of fundus photos by expert-determined Frisén grade in training and validation sets for each of the five subsets.