Article

Artificial intelligence to detect papilledema from ocular fundus photographs

MILEA, Dan, BIOUSSE, Valérie, BONZAI, group & BONSAI Group

SANDA, Nicolae (Collab.), THUMANN, Gabriele (Collab.)

Abstract

BACKGROUND: Nonophthalmologist physicians do not confidently perform direct ophthalmos-copy. The use of artificial intelligence to detect papilledema and other optic-disk abnormalities from fundus photographs has not been well studied METHODS: We trained, validated, and externally tested a deep-learning system to classify optic disks as being normal or having papilledema or other abnormalities from 15,846 retrospectively collected ocular fundus photographs that had been obtained with pharmacologic pupillary dilation and various digital cameras in persons from multiple ethnic populations. Of these photographs, 14,341 from 19 sites in 11 coun-tries were used for training and validation, and 1505 photographs from 5 other sites were used for external testing. Performance at classifying the optic-disk ap-pearance was evaluated by calculating the area under the receiver-operating-characteristic curve (AUC), sensitivity, and specificity, as compared with a reference standard of clinical diagnoses by neuro-ophthalmologists RESULTS: The training and validation data sets from 6779 patients included 14,341 photo-graphs: 9156 of [...]

Reference

MILEA, Dan, BIOUSSE, Valérie, BONZAI, group & BONSAI Group, SANDA, Nicolae (Collab.), THUMANN, Gabriele (Collab.). Artificial intelligence to detect papilledema from ocular fundus photographs. *New England Journal of Medicine*, 2020, vol. 382, no. 18, p. 1687-1695

DOI : 10.1056/NEJMoa1917130
PMID : 32286748

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 APRIL 30, 2020 VOL. 382 NO. 18

Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs

D. Milea, R.P. Najjar, Z. Jiang, D. Ting, C. Vasseneix, X. Xu, M. Aghsaei Fard, P. Fonseca, K. Vanikieti, W.A. Lagrèze, C. La Morgia, C.Y. Cheung, S. Hamann, C. Chiquet, N. Sanda, H. Yang, L.J. Mejico, M.-B. Rougier, R. Kho, Thi H.C. Tran, S. Singhal, P. Gohier, C. Clermont-Vignal, C.-Y. Cheng, J.B. Jonas, P. Yu-Wai-Man, C.L. Fraser, J.J. Chen, S. Ambika, N.R. Miller, Y. Liu, N.J. Newman, T.Y. Wong, and V. Biousse, for the BONSAI Group*

ABSTRACT

BACKGROUND

Nonophthalmologist physicians do not confidently perform direct ophthalmoscopy. The use of artificial intelligence to detect papilledema and other optic-disk abnormalities from fundus photographs has not been well studied.

METHODS

We trained, validated, and externally tested a deep-learning system to classify optic disks as being normal or having papilledema or other abnormalities from 15,846 retrospectively collected ocular fundus photographs that had been obtained with pharmacologic pupillary dilation and various digital cameras in persons from multiple ethnic populations. Of these photographs, 14,341 from 19 sites in 11 countries were used for training and validation, and 1505 photographs from 5 other sites were used for external testing. Performance at classifying the optic-disk appearance was evaluated by calculating the area under the receiver-operating-characteristic curve (AUC), sensitivity, and specificity, as compared with a reference standard of clinical diagnoses by neuro-ophthalmologists.

RESULTS

The training and validation data sets from 6779 patients included 14,341 photographs: 9156 of normal disks, 2148 of disks with papilledema, and 3037 of disks with other abnormalities. The percentage classified as being normal ranged across sites from 9.8 to 100%; the percentage classified as having papilledema ranged across sites from zero to 59.5%. In the validation set, the system discriminated disks with papilledema from normal disks and disks with nonpapilledema abnormalities with an AUC of 0.99 (95% confidence interval [CI], 0.98 to 0.99) and normal from abnormal disks with an AUC of 0.99 (95% CI, 0.99 to 0.99). In the external-testing data set of 1505 photographs, the system had an AUC for the detection of papilledema of 0.96 (95% CI, 0.95 to 0.97), a sensitivity of 96.4% (95% CI, 93.9 to 98.3), and a specificity of 84.7% (95% CI, 82.3 to 87.1).

CONCLUSIONS

A deep-learning system using fundus photographs with pharmacologically dilated pupils differentiated among optic disks with papilledema, normal disks, and disks with nonpapilledema abnormalities. (Funded by the Singapore National Medical Research Council and the SingHealth Duke–NUS Ophthalmology and Visual Sciences Academic Clinical Program.)

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Wong at the Singapore National Eye Center, 11 Third Hospital Ave., Singapore 168751, Singapore, or at wong.tien.yin@singhealth.com.sg.

*A list of the members of the BONSAI Group is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Milea and Najjar and Mr. Jiang and Drs. Liu, Newman, Wong, and Biousse contributed equally to this article.

This article was published on April 14, 2020, and updated on May 6, 2020, at NEJM.org.

N Engl J Med 2020;382:1687-95.
DOI: 10.1056/NEJMoa1917130

Examination of the optic nerves is a fundamental component of the clinical examination, but direct ophthalmoscopy is usually avoided or poorly performed by general physicians and nonophthalmic specialists.[1-4] Detection of papilledema, defined as optic-nerve edema from intracranial hypertension, and the ability to determine that the optic disk is normal are valuable in the evaluation of patients with headache and other neurologic symptoms. The findings on ophthalmoscopy influence diagnostic strategy and treatment options.[3-13] Failure to detect papilledema may result in visual loss and neurologic complications.[2-8,13]

Digital ocular fundus photography has been used to obtain optic-disk images for the purpose of detecting papilledema and other optic-disk abnormalities in a variety of clinical settings, including emergency departments, urgent care centers, and neurologic and general adult and pediatric clinics.[1,4,7,12,14-18] In one study conducted in an emergency department,[12] 8.5% of patients presenting with headache had abnormal findings on fundus photographs. However, these photographs need to be interpreted by physicians onsite at the time of photography[15] or sent through tele-ophthalmology platforms for assessment by ophthalmologists or other experts.[17,19,20]

Artificial intelligence and deep learning have been developed for the automated detection of diabetic retinopathy and glaucomatous optic neuropathy from ocular fundus photographs.[21-30] We investigated whether a deep-learning system could aid in the diagnosis of optic-nerve abnormalities, particularly papilledema, from fundus photographs. We trained, validated, and externally tested a deep-learning system to identify and classify normal optic disks, disks with papilledema, and disks with other abnormalities from digital ocular fundus photographs collected from a large, international, multiethnic population.

Methods

Study Design and Oversight

We conducted a training, validation, and external-testing study on an artificial intelligence–based deep-learning system using digital color ocular fundus photographs, retrospectively collected by an international consortium (BONSAI: Brain and Optic Nerve Study with Artificial Intelligence) composed of neuro-ophthalmologists. (For details on study group organization and participating centers, see Section S1 in the Supplementary Appendix, available with the full text of this article at NEJM.org.)

We first trained and validated the deep-learning system using 14,341 fundus photographs obtained at 19 sites in 11 countries; we then externally tested the system on 1505 photographs obtained at 5 other centers in 5 countries. The study was approved by the centralized institutional review board of SingHealth, Singapore, and at each contributing institution and was conducted in accordance with the principles of the Declaration of Helsinki. Informed consent was exempted, given the retrospective nature of the data collection and the use of deidentified ocular fundus photographs.

Image Acquisition

Retrospectively collected fundus photographs were obtained from one or both eyes after pharmacologic pupillary dilation, with the use of various commercial digital fundus cameras. (For details on the cameras used in the study, see Section S2b and Table S1.) Images were centered on either the macula or the optic disk, but always including the optic disk, at various fields of view (subtending 20 to 45 degrees). Deidentified, unaltered images (size, 0.5 to 2 megabytes per image) were transferred to the Singapore Eye Research Institute for inclusion in the study.

Study Patients

The study included patients with optic-nerve disorders and healthy persons of multiple ethnic groups from 24 centers in 15 countries. The ocular fundus photographs, including those of normal optic nerves and a variety of common neuro-ophthalmic conditions affecting the optic nerves, were collected in each center by neuro-ophthalmologists who routinely obtain fundus photographs and who had access to the patients' medical records (principal investigators from each of these centers are authors of this article). In addition, photographs of normal optic disks were randomly selected from 3 centers, including Indian, Asian, and non-Asian patients, which provided large sets of photographs of normal optic disks, as determined by general ophthalmologists. (For patient characteristics, see Section S2a, Fig. S1, and Table S2.)

DEFINITION OF OPTIC-DISK ABNORMALITIES

Neuro-ophthalmologists provided a specific diagnosis, gathered retrospectively from medical records, for each fundus photograph at the time of clinical evaluation, considered for the purposes of this research to be the reference standard, on the basis of the appearance of the optic-nerve head as well as the medical evaluation, ancillary testing, and follow-up visits. All the patients seen by neuro-ophthalmologists underwent neuro-ophthalmologic evaluations, including visual-field and other tests, in order to obtain a final clinical diagnosis pertaining to each photograph, according to standard diagnostic criteria that could include brain imaging and lumbar puncture in some cases. (For details on the diagnostic process and reference standards, see Section S2a.) Patients from the three centers that provided photographs of normal fundi also underwent comprehensive evaluations by ophthalmologists.

Fundus photographs were classified by the study steering committee into three groups, consistent with the original reference diagnosis: normal optic disk; disk with papilledema due to proven intracranial hypertension; and disk with other abnormalities, including other visible abnormalities of the optic-nerve head such as anterior ischemic and inflammatory optic neuropathies, optic-disk drusen, optic atrophy, and congenital optic-nerve abnormalities. Patients with normal optic nerves were included only in the absence of any ocular conditions such as substantial media opacities, retinal disorders, or glaucoma. These three groups were considered reference standards for training, validation, and external testing.

DEVELOPMENT OF THE DEEP-LEARNING CLASSIFICATION MODEL

Our system consisted of a segmentation network (U-Net) to detect the location of the optic disk from fundus photographs and a classification network (DenseNet) to classify the optic disk into one of the three classes: normal disk, disk with papilledema, and disk with other abnormalities. To visualize optic-nerve abnormalities, we used a class-activation map (Fig. S2). A five-fold cross-validation was performed on the primary data set to differentiate among normal optic disks, disks with papilledema, and disks with other abnormalities (Fig. S3). With the use of the same thresholds as on the primary data set, the diagnostic performance of the three-class classification model was then assessed on the five independent external-testing data sets. (For details of the deep-learning system, see Section S2c, Fig. S4, and Table S3.[23,24])

STATISTICAL ANALYSIS

To determine performance characteristics, we used the one-versus-rest strategy and calculated the area under the receiver-operating-characteristic curve (AUC), sensitivity, specificity, and accuracy for the following three cases according to the results of our classification model: normal as compared with abnormal optic disks (including disks with papilledema and disks with other abnormalities), disks with papilledema as compared with those without papilledema (including normal disks and disks with nonpapilledema abnormalities), and disks with nonpapilledema abnormalities as compared with normal disks and disks with papilledema. Predictive values for the classification of papilledema and other optic-disk abnormalities were also calculated for each external-testing site. Bootstrapping was used to estimate 95% confidence intervals of the performance metrics, with the patient as the resampling unit. (For details on statistical and bootstrapping procedures, see Section S2d.)

RESULTS

CHARACTERISTICS OF THE DATA SETS

A total of 15,846 photographs (from 7532 patients [71.0% with photographs of both eyes, 17.6% with photographs of one eye, and 11.4% with repeat photographs during follow-up visits]; mean age, 48.6 years [range, 3 to 98]; 43.4% men or boys) were used to train, validate, and externally test the performance of the deep-learning system, after the exclusion of 153 photographs because of poor quality or poor centration of the photograph, with the optic disk being cut off at the edge. (For details on the inclusion and exclusion of photographs, see Section S2 and Fig. S1.)

The system was trained and validated on 14,341 photographs collected from 6779 patients in the first 19 sites of the BONSAI consortium, including 9156 images of normal optic disks, 2148 of disks with confirmed papilledema from proven intracranial hypertension, and 3037 of

Table 1. Summary of Training, Validation, and External-Testing Data Sets, According to Diagnosis of Fundus Images.

Location of Center	Normal Disks	Disks with Papilledema	Disks with Other Abnormalities*	Total
		number of images		
Primary training and validation data sets				
Angers, France	116	369	701	1186
Atlanta, GA, United States	441	1146	340	1927
Baltimore, MD, United States	295	104	49	448
Bologna, Italy	43	13	264	320
Bordeaux, France	19	25	26	70
Chennai, India	169	124	423	716
Coimbra, Portugal	61	28	244	333
Geneva, Switzerland	66	15	59	140
Grenoble, France	130	6	78	214
Guangzhou, China	27	0	91	118
Hong Kong, China	722	16	316	1054
Lille, France	330	0	0	330
London, United Kingdom	234	40	159	433
Manila, Philippines	17	17	39	73
Nagpur, India	1911	0	0	1911
Paris, France	152	89	53	294
Singapore, Singapore	4053	42	83	4178
Sydney, Australia	351	86	95	532
Syracuse, NY, United States	19	28	17	64
External-testing data sets				
Bangkok, Thailand	177	38	104	319
Copenhagen, Denmark	90	47	63	200
Freiburg, Germany	98	92	138	328
Rochester, MN, United States	92	95	97	284
Tehran, Iran	156	88	130	374
Total at all centers	9769	2508	3569	15,846

* Other optic-disk abnormalities included nonarteritic anterior ischemic optic neuropathy (760 images), anterior inflammatory optic neuritis (390), other causes of optic-disk swelling (164), optic-disk drusen (570), optic-disk congenital abnormalities (56), and optic atrophy (1629).

disks with other abnormalities. The percentage of images classified as being normal ranged across data sets from 9.8 to 100%; the percentage classified as having papilledema ranged across data sets from zero to 59.5%. A separate set of 1505 photographs that were collected from 5 other centers, including 613 images of normal disks, 360 of disks with papilledema, and 532 of disks with other abnormalities, was used for the external testing (Table 1).

CLASSIFICATION PERFORMANCE IN THE VALIDATION DATA SET

In the validation data set, the system discriminated normal from abnormal optic disks (including disks with papilledema and disks with other abnormalities) with an AUC of 0.99 (95% confidence interval [CI], 0.99 to 0.99) and discriminated disks with papilledema from all other optic disks (normal disks and disks with non-papilledema abnormalities) with an AUC of

Table 2. Classification Performance of the Deep-Learning System on the Primary Validation and External-Testing Data Sets.*

One-vs.-Rest Classification	Total	Normal	Papilledema	Other	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
		number					*percent*	
Primary validation data set†								
Normal vs. papilledema + other	14,341	9156	2148	3037	0.99 (0.99–0.99)	93.5 (92.9–94.1)	96.2 (95.5–96.9)	94.5 (94.0–94.9)
Papilledema vs. other + normal	14,341	9156	2148	3037	0.99 (0.98–0.99)	93.2 (91.8–94.5)	95.1 (94.7–95.6)	94.8 (94.4–95.3)
Other vs. normal + papilledema	14,341	9156	2148	3037	0.97 (0.97–0.97)	93.0 (91.9–94.0)	89.0 (88.3–89.8)	89.8 (89.2–90.5)
External-testing data set‡								
Normal vs. papilledema + other	1,505	613	360	532	0.98 (0.97–0.98)	86.6 (83.8–89.3)	95.3 (93.8–96.8)	91.8 (90.3–93.3)
Papilledema vs. other + normal	1,505	613	360	532	0.96 (0.95–0.97)	96.4 (93.9–98.3)	84.7 (82.3–87.1)	87.5 (85.5–89.3)
Other vs. normal + papilledema	1,505	613	360	532	0.90 (0.88–0.92)	85.7 (82.5–88.8)	78.6 (75.5–81.5)	81.1 (78.8–83.3)

* "Normal" indicates normal optic disks, "papilledema" indicates disks with papilledema, and "other" indicates disks with nonpapilledema abnormalities. AUC denotes area under the receiver-operating-characteristic curve.

† The mean age of the patients included in the primary training and validation data set was 49.1 years (95% CI, 48.7 to 49.6), on the basis of 94.5% of available patient demographic data. The male-to-female ratio in the primary training and validation data set was 0.79 (44.0% men or boys), on the basis of 94.4% of available patient demographic data.

‡ The mean age of the patients included in the external-testing data set was 44.4 years (95% CI, 43.1 to 45.8), on the basis of 99.7% of available patient demographic data. The male-to-female ratio in the testing data set was 0.61 (38.0% men or boys), on the basis of 99.6% of available patient demographic data.

0.99 (95% CI, 0.98 to 0.99), a sensitivity of 93.2% (95% CI, 91.8 to 94.5), and a specificity of 95.1% (95% CI, 94.7 to 95.6). The system also discriminated disks with nonpapilledema abnormalities from normal disks and disks with papilledema with an AUC of 0.97 (95% CI, 0.97 to 0.97) (Table 2 and Fig. S3).

CLASSIFICATION PERFORMANCE IN THE EXTERNAL-TESTING DATA SETS

In the external-testing data sets, the AUCs were 0.98 (95% CI, 0.97 to 0.98), 0.96 (95% CI, 0.95 to 0.97), and 0.90 (95% CI, 0.88 to 0.92) for the classification of normal disks, disks with papilledema, and disks with other abnormalities, respectively (Table 2 and Fig. 1). Across the five external-testing data sets, the AUCs ranged from 0.96 to 0.99 for the discrimination of normal from abnormal optic disks and from 0.93 to 0.98 for the discrimination of disks with papilledema from all other optic disks. (For details on the classification performance of the system on the individual external-testing data sets, see Table S4.)

The overall accuracies of our deep-learning system for the detection of normal disks, disks with papilledema, and disks with other abnormalities in the external-testing data sets were 91.8% (95% CI, 90.3 to 93.3), 87.5% (95% CI, 85.5 to 89.3), and 81.1% (95% CI, 78.8 to 83.3), respectively. In the five external-testing data sets, the trained system had an overall sensitivity and specificity of 96.4% (95% CI, 93.9 to 98.3) and 84.7% (95% CI, 82.3 to 87.1), respectively, for the detection of papilledema (Table 2). The mean estimated prevalence of papilledema in all the sets of data was 9.5% (Table S6), which resulted in an overall positive predictive value of the system for papilledema of 39.8% (95% CI, 36.6 to 43.2) and a negative predictive value of 99.6% (95% CI, 99.2 to 99.7) (Table 3). (The predictive values of the deep-learning system across a full prevalence range for the detection of normal discs, discs with papilledema, and discs with other abnormalities are provided in Fig. S5.)

ADJUDICATION OF CLASSIFICATION ERRORS

In a post hoc analysis, four expert neuro-ophthalmologists who were not involved in the original

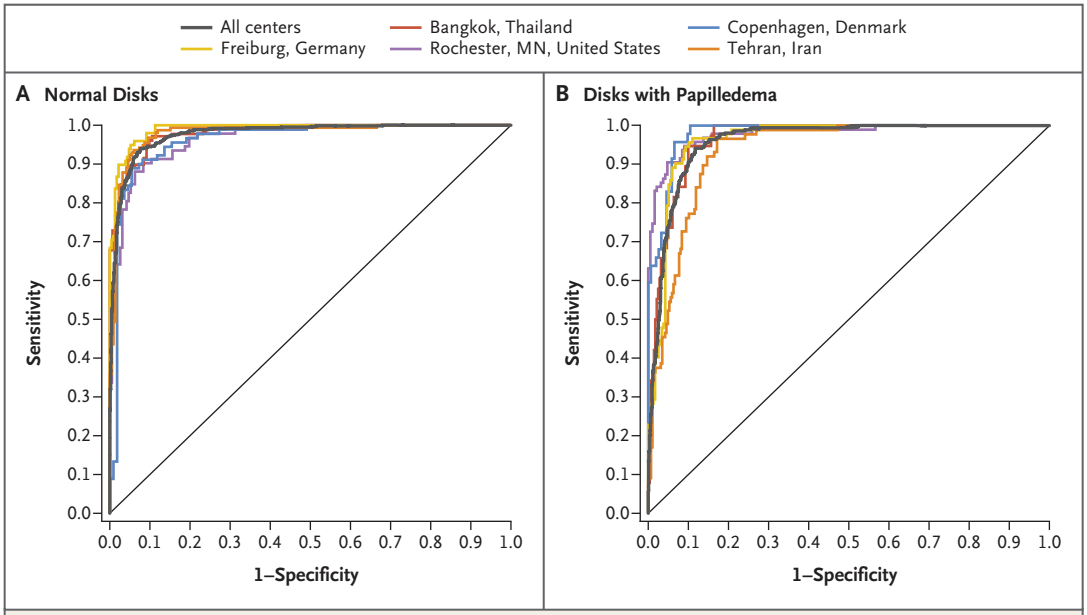

Figure 1. Performance of the Deep-Learning System for the Detection of Normal Disks and Disks with Papilledema in the External-Testing Data Sets.

The external-testing data sets included ocular fundus photographs from five centers with diverse ethnic backgrounds. As shown in Panel A, the deep-learning system discriminated normal optic disks from abnormal ones, with areas of the receiver-operating-characteristic curve (AUCs) that ranged from 0.96 to 0.99 and an overall AUC of 0.98 (95% CI, 0.97 to 0.98). As shown in Panel B, the deep-learning system discriminated disks with papilledema from normal disks and disks with nonpapilledema abnormalities, with AUCs that ranged from 0.93 to 0.98 and an overall AUC of 0.96 (95% CI, 0.95 to 0.97).

analyses and who were unaware of the initial reference-standard classification reviewed the 177 fundus photographs (11.8% of the 1505 photographs) in the external-testing data sets that had discordant findings between the reference standard by site expert neuro-ophthalmologists and the classification by the deep-learning system. This analysis showed that of the 360 disks with papilledema, 15 (4.2%) were misclassified by the system as disks with other abnormalities but never as normal optic disks. (For details on the 177 misclassified fundus photographs, see Sections S3a and S3b and Fig. S6A through S6C.) A review by the same neuro-ophthalmologists of the misclassified papilledema images at a patient level (i.e., both eyes of a patient viewed as a pair) disclosed only one patient with papilledema in both eyes missed by the system in the external-testing data sets. In 10 of the 177 fundus photographs for which the system provided a classification that differed from the reference standard, the four neuro-ophthalmologists, after review of the fundus photographs, agreed with the deep-learning system.

Subsequently, arbitration was performed by contacting the neuro-ophthalmologists at the applicable external-testing sites and requesting that they reevaluate their initial reference-standard diagnosis. In these 10 discordant cases, the classification of the deep-learning system was considered accurate, and the discrepancies were found to be the result of labeling errors by the site investigators. We performed a post hoc reanalysis of the corrected external-testing data set with the 10 reclassified images, which resulted in a slightly improved average AUC for the overall classification performance of the system, from 0.941 to 0.948. Subsequently, we requested that the neuro-ophthalmologists at each of the five centers used for the external-testing data sets recheck all diagnoses in their respective series of patients; this led to the identification of an additional 3 mislabeled photographs. However, all 3 remained in the category of disks with

Table 3. Predictive Values of the Deep-Learning System in the External-Testing Data Sets.*

Center and Ophthalmic Condition	Estimated Prevalence	Positive Predictive Value (95% CI)	Negative Predictive Value (95% CI)
		percent	
Bangkok, Thailand			
Papilledema	8.9	37.2 (30.9–43.9)	99.4 (97.7–99.8)
Other optic-disk abnormalities	63.3	89.7 (86.3–92.2)	72.7 (63.8–80.0)
Copenhagen, Denmark			
Papilledema	3.6	26.3 (18.3–36.2)	100 (100–100)
Other optic-disk abnormalities	14.3	33.4 (27.8–39.4)	98.1 (95.7–99.2)
Freiburg, Germany			
Papilledema	10.0	34.6 (29.2–40.5)	99.9 (98.9–100)
Other optic-disk abnormalities	40.0	78.6 (72.4–83.7)	90.6 (86.2–93.7)
Rochester, MN, United States			
Papilledema	17.2	55.9 (47.7–63.8)	99.2 (97.7–99.8)
Other optic-disk abnormalities	32.8	62.5 (56.8–67.8)	96.6 (92.4–98.5)
Tehran, Iran			
Papilledema	8.0	32.8 (27.1–38.9)	99.2 (98.3–99.6)
Other optic-disk abnormalities	32.0	60.6 (54.7–66.2)	87.9 (84.0–100)
All centers			
Papilledema	9.5	39.8 (36.6–43.2)	99.6 (99.2–99.7)
Other optic-disk abnormalities	36.5	69.7 (67.0–72.3)	90.5 (88.6–92.2)

* We calculated predictive values using the sensitivity and specificity of the deep-learning system in the five individual external-testing data sets and overall, after taking into account the estimated prevalence of papilledema and other optic-disk abnormalities at each site. (For details on the calculation of predictive values, see Section S2d in the Supplementary Appendix.)

nonpapilledema abnormalities and therefore did not change our results.

DISCUSSION

Our objective was to assess the performance of a deep-learning system to detect papilledema from fundus images taken at many international centers, from patients with a variety of ethnic backgrounds, types of fundus pigmentation, and ages and using a variety of commercially available digital fundus cameras. Our main finding was that an artificial-intelligence algorithm using deep-learning neural networks could discriminate among normal optic disks, disks with papilledema, and disks with other abnormalities. In our external-testing data sets, the sensitivity for detecting papilledema was 96.4% and the specificity was 84.7%. Negative predictive values were high, but positive predictive values were lower and varied considerably depending on the prevalence of papilledema and other optic-nerve conditions.

Several studies have suggested that direct ophthalmoscopy can be replaced by more user-friendly ocular fundus digital cameras that provide high-quality photographs of the optic nerve and retina, even without pharmacologic dilation of the pupils,[1,2,4,15,17,31,32] although our study used photographs taken after pupillary dilation. Most deep-learning research in ophthalmology has been for screening of retinal disorders and glaucoma.[24-30,33-35] Previous studies using fewer images than ours showed that deep-learning systems could recognize right from left optic disks in the presence of optic-nerve abnormalities on fundus photographs,[36] could discriminate disks with papilledema from normal disks with an average accuracy of 93% (similar to the value in

our study),[37] and could differentiate true optic-disk swelling from pseudo-swelling with an accuracy of approximately 95%.[38]

Our study has limitations. First, it was retrospective, since the photographs were collected retrospectively over a period of several years from a large number of centers. This resulted in an imbalance in class distribution among groups (i.e., differing prevalence of different optic-disk conditions), a mix of consecutive series of patients and convenience samples, and labeling errors.

Second, we chose as a reference standard the final diagnosis of the appearance of the normal optic-nerve head given by an expert neuro-ophthalmologist at each center, based on the clinical examination and other findings, including brain imaging and lumbar puncture when appropriate for patients with suspected papilledema and follow-up data. The final diagnosis of the appearance of the optic-nerve head in healthy persons was determined by neuro-ophthalmologists or ophthalmologists, on the basis of comprehensive ophthalmologic evaluations. A total of 10 labeling errors by the investigators were discovered and correctly identified by our deep-learning system. Relabeling the 10 of 1505 images in the external-testing data set improved the overall performance of the deep-learning system only slightly. Although our deep-learning system misclassified 15 of 360 photographs of disks with papilledema (4.2%), it labeled them as disks with other abnormalities and never as normal disks.

Third, the abnormal photographs were obtained after pharmacologic dilation of the pupils and may not reflect general practice. Fourth, our network was trained and calibrated primarily to identify normal optic nerves and those with papilledema. Therefore, the threshold for diagnosing papilledema was low, in order to avoid false negatives. Whether the results will be reproducible under other circumstances is not known.

We found that an artificial-intelligence, deep-learning algorithm that was trained on ocular fundus photographs had high sensitivity and specificity for discriminating between papilledema and normal optic nerves. Negative predictive values were high, but positive predictive values varied depending on the prevalence of papilledema in the population being studied. Further investigation is required in order to prospectively validate the use of deep-learning systems in various settings, which may have different prevalences of optic-disk abnormalities from those in our study.[39]

Supported by the Singapore National Medical Research Council (Clinician Scientist Individual Research Grant CIRG18Nov-0013) and the SingHealth Duke–NUS Medical School Ophthalmology and Visual Sciences Academic Clinical Program (Clinical Innovation Support Program Grant 05/FY2019/P2/06-A60).

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

APPENDIX

The authors' full names and academic degrees are as follows: Dan Milea, M.D., Ph.D., Raymond P. Najjar, Ph.D., Zhubo Jiang, M.Sc., Daniel Ting, M.D., Ph.D., Caroline Vasseneix, M.D., Xinxing Xu, Ph.D., Masoud Aghsaei Fard, M.D., Pedro Fonseca, M.D., Kavin Vanikieti, M.D., Wolf A. Lagrèze, M.D., Chiara La Morgia, M.D., Ph.D., Carol Y. Cheung, Ph.D., Steffen Hamann, M.D., Ph.D., Christophe Chiquet, M.D., Ph.D., Nicolae Sanda, M.D., Ph.D., Hui Yang, M.D., Ph.D., Luis J. Mejico, M.D., Marie-Bénédicte Rougier, M.D., Richard Kho, M.D., Thi H.C. Tran, M.D., Shweta Singhal, M.B., B.S., Ph.D., Philippe Gohier, M.D., Catherine Clermont-Vignal, M.D., Ching-Yu Cheng, M.D., Ph.D., M.P.H., Jost B. Jonas, M.D., Patrick Yu-Wai-Man, M.B., B.S., Ph.D., Clare L. Fraser, M.B., B.S., M.Med., John J. Chen, M.D., Ph.D., Selvakumar Ambika, D.O., D.N.B., Neil R. Miller, M.D., Yong Liu, Ph.D., Nancy J. Newman, M.D., Tien Y. Wong, M.D., Ph.D., and Valérie Biousse, M.D.

The authors' affiliations are as follows: the Singapore National Eye Center (D.M., D.T., S.S., C.-Y.C., T.Y.W.), Singapore Eye Research Institute (D.M., R.P.N., D.T., C.V., S.S., C.-Y.C., T.Y.W.), Duke–NUS Medical School (D.M., R.P.N., D.T., S.S., C.-Y.C., T.Y.W.), Institute of High Performance Computing, Agency for Science, Technology, and Research (Z.J., X.X., Y.L.), and Yong Loo Lin School of Medicine, National University of Singapore (S.S., T.Y.W.) — all in Singapore; Farabi Eye Hospital, Tehran University of Medical Science, Tehran, Iran (M.A.F.); the Department of Ophthalmology, Centro Hospitalar e Universitário de Coimbra, and the Coimbra Institute for Biomedical Imaging and Translational Research, University of Coimbra, Coimbra, Portugal (P.F.); the Department of Ophthalmology, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (K.V.); the Eye Center, Medical Center, University of Freiburg, Freiburg (W.A.L.), and the Department of Ophthalmology, Ruprecht Karl University of Heidelberg, Mannheim (J.B.J.) — both in Germany; IRCCS Istituto delle Scienze Neurologiche di Bologna, Unità Operativa Complessa Clinica Neurologica, and Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna, Bologna, Italy (C.L.M.); the Department of Ophthalmology and Visual Sciences, Chinese University of Hong Kong, Hong Kong (C.Y.C.), and Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou (H.Y.) — both in China; the Department of Ophthalmology, Rigshospitalet, University of Copenhagen, Glostrup, Denmark (S.H.); the Department of Ophthalmology, University Hospital of Grenoble-Alpes, and Grenoble-Alpes University, HP2 Laboratory, INSERM Unité 1042, Grenoble (C.C.), Service d'Ophtalmologie, Unité Rétine–Uvéites–Neuro-Ophtalmologie, Hôpital Pellegrin, Centre Hospitalier Universitaire de Bordeaux, Bordeaux (M.-B.R.), the Department of Ophthalmology, Lille Catholic Hospital, Lille Catholic University, and INSERM Unité 1171, Lille (T.H.C.T.), the Department of Ophthalmology, University Hospital Angers, Angers (P.G.), and Rothschild Foundation Hospital, Paris (C.C.-V.) — all in France; the Department of Clinical Neurosciences, Geneva University Hospital, Geneva (N.S.); the Department of Neurology, SUNY Upstate Medical University, Syracuse, NY (L.J.M.); the American Eye Center, Mandaluyong City, Philippines (R.K.); Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, University College London, London

(P.Y.-W.-M.), and Cambridge Eye Unit, Addenbrooke's Hospital, Cambridge University Hospitals, and Cambridge Centre for Brain Repair and Medical Research Council Mitochondrial Biology Unit, Department of Clinical Neurosciences, University of Cambridge, Cambridge (P.Y.-W.-M.) — all in the United Kingdom; the Save Sight Institute, Faculty of Health and Medicine, University of Sydney, Sydney (C.L.F.); the Department of Ophthalmology and Neurology, Mayo Clinic, Rochester, MN (J.J.C.); the Department of Neuro-ophthalmology, Sankara Nethralaya, Medical Research Foundation, Chennai, India (S.A.); the Departments of Ophthalmology, Neurology, and Neurosurgery, Johns Hopkins University School of Medicine, Baltimore (N.R.M.); and the Departments of Ophthalmology and Neurology, Emory University School of Medicine, Atlanta (N.J.N., V.B.).

REFERENCES

1. Bruce BB, Lamirel C, Wright DW, et al. Nonmydriatic ocular fundus photography in the emergency department. N Engl J Med 2011;364:387-9.
2. Mackay DD, Garza PS, Bruce BB, Newman NJ, Biousse V. The demise of direct ophthalmoscopy: a modern clinical challenge. Neurol Clin Pract 2015;5:150-7.
3. Golombievski E, Doerrler MW, Ruland SD, McCoyd MA, Biller J. Frequency of direct funduscopy upon initial encounters for patients with headaches, altered mental status, and visual changes: a pilot study. Front Neurol 2015;6:233.
4. Biousse V, Bruce BB, Newman NJ. Ophthalmoscopy in the 21st century: the 2017 H. Houston Merritt Lecture. Neurology 2018;90:167-75.
5. Rigi M, Almarzouqi SJ, Morgan ML, Lee AG. Papilledema: epidemiology, etiology, and clinical management. Eye Brain 2015;7:47-57.
6. De Luca GC, Bartleson JD. When and how to investigate the patient with headache. Semin Neurol 2010;30:131-44.
7. Thulasi P, Fraser CL, Biousse V, Wright DW, Newman NJ, Bruce BB. Nonmydriatic ocular fundus photography among headache patients in an emergency department. Neurology 2013;80:432-7.
8. Medical Advisory Secretariat. Neuroimaging for the evaluation of chronic headaches: an evidence-based analysis. Ont Health Technol Assess Ser 2010;10:1-57.
9. American College of Radiology. ACR appropriateness criteria: headache. Revised 2019 (https://acsearch.acr.org/docs/69482/Narrative/).
10. Tabatabai RR, Swadron SP. Headache in the emergency department: avoiding misdiagnosis of dangerous secondary causes. Emerg Med Clin North Am 2016;34:695-716.
11. National Hospital Ambulatory Medical Care Survey: 2016 emergency department summary tables (https://www.cdc.gov/nchs/data/nhamcs/web_tables/2016_ed_web_tables.pdf).
12. Sachdeva V, Vasseneix C, Hage R, et al. Optic nerve head edema among patients presenting to the emergency department. Neurology 2018;90(5):e373-e379.
13. Poostchi A, Awad M, Wilde C, Dineen RA, Gruener AM. Spike in neuroimaging requests following the conviction of the optometrist Honey Rose. Eye (Lond) 2018;32:489-90.
14. Wong TY, Mitchell P. Hypertensive retinopathy. N Engl J Med 2004;351:2310-7.
15. 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(1):28-33.e1.
16. Wong TY. Is retinal photography useful in the measurement of stroke risk? Lancet Neurol 2004;3:179-83.
17. Irani NK, Bidot S, Peragallo JH, Esper GJ, Newman NJ, Biousse V. Feasibility of a non-mydriatic ocular fundus camera in an outpatient neurology clinic. Neurologist 2020;25:19-23.
18. Ivan Y, Ramgopal S, Cardenas-Villa M, et al. Feasibility of the digital retinography system camera in the pediatric emergency department. Pediatr Emerg Care 2018;34:488-91.
19. Rathi S, Tsui E, Mehta N, Zahid S, Schuman JS. The current state of teleophthalmology in the United States. Ophthalmology 2017;124:1729-34.
20. Wedekind L, Sainani K, Pershing S. Supply and perceived demand for teleophthalmology in triage and consultations in California emergency departments. JAMA Ophthalmol 2016;134:537-43.
21. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-58.
22. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
23. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322:1806-16.
24. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations. Prog Retin Eye Res 2019;72:100759.
25. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39.
26. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
27. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211-23.
28. Bhaskaranand M, Ramachandra C, Bhat S, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. Diabetes Technol Ther 2019;21:635-43.
29. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 2018;125:1199-206.
30. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. JAMA Ophthalmol 2019;137:1353-60.
31. Zafar S, Cardenas YM, Leishangthem L, Yaddanapudi S. Opinion and special articles: amateur fundus photography with various new devices: our experience as neurology residents. Neurology 2018;90:897-901.
32. Bruce BB, Bidot S, Hage R, et al. Fundus Photography vs. Ophthalmoscopy Outcomes in the Emergency Department (FOTO-ED) phase III: Web-based, in-service training of emergency providers. Neuroophthalmology 2018;42:269-74.
33. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1:40.
34. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. Ophthalmology 2019;126:513-21.
35. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digital Health 2019;1(6):e271-e297.
36. Liu TYA, Ting DSW, Yi PH, et al. Deep learning and transfer learning for optic disc laterality detection: Implications for machine learning in neuro-ophthalmology. J Neuroophthalmol 2019 August 22 (Epub ahead of print).
37. Akbar S, Akram MU, Sharif M, Tariq A, Yasin UU. Decision support system for detection of papilledema through fundus retinal images. J Med Syst 2017;41:66.
38. Ahn JM, Kim S, Ahn KS, Cho SH, Kim US. Accuracy of machine learning for differentiation between optic neuropathies and pseudopapilledema. BMC Ophthalmol 2019;19:178.
39. Keane P, Topol E. Reinventing the eye exam. Lancet 2019;394:2141.