NEJM
GROUP

TOPIC COLLECTION: ARTIFICIAL INTELLIGENCE AND OPHTHALMIC DIAGNOSIS

THE NEW ENGLAND JOURNAL OF MEDICINE

NEJM JOURNAL WATCH
Cardiology
Emergency Medicine
Gastroenterology
General Medicine
Hospital Medicine
Infectious Diseases
Neurology
Oncology and Hematology
Pediatrics and Adolescent Medicine
Physician's First Watch
Psychiatry
Women's Health

NEJM CAREERCENTER

NEJM KNOWLEDGE+

NEJM CATALYST

NEJM RESIDENT 360

Letter from the Editor

Papilledema detection by nonophthalmologists can be challenging. Conversely, the classification of anomalous optic discs as papilledematous may subject patients to anxiety regarding the diagnosis and to expensive and invasive testing, such as MRI and lumbar puncture, that ultimately proves unnecessary.

In the NEJM article "Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs," Milea and colleagues report on a deep-learning system developed to distinguish papilledematous discs from normal optic discs or discs with other abnormalities. A dilated fundus exam by a trained ophthalmologist remains the gold standard for papilledema diagnosis, and clinical diagnosis was the reference standard. The strength of this article is that it included fundus images from patients of different ethnic, racial, and geographical groups. The system could detect papilledema with greater than 95% sensitivity, but specificity was less high. Additionally, the system had more difficulty discriminating papilledema from other abnormalities than from normal optic discs. Nonetheless, it was unlikely to erroneously classify papilledema as normal. Such a system could help emergency department (ED) or other physicians exclude patients not needing an urgent work-up for possible intracranial pressure elevation or central nervous system mass. Its utility in distinguishing true disc swelling from other anomalies is less clear, and despite the large datasets used to train and evaluate the system, AI is still clearly challenged by this task. This data set was derived from dilated fundus photographs, and the ability to operate with such high fidelity on undilated eyes, as in an ED setting, remains to be shown. As in other image-intensive areas of medicine, such as radiology, adoption of AI based systems is likely to help in screening, especially for nonspecialists asked to make a diagnosis. However, the ability of such systems to finely discriminate among abnormalities is yet to be proven, at least where the ocular fundus is concerned. Finally, this trial reminds us that AI systems can only be as good as the training sets used to develop them, and misclassification of reference standard material can lead to erroneous function.

The NEJM Journal Watch summaries address related issues: Shahidi et al. point out the challenges of using AI to resolve differences between clinical and pathologic diagnoses; Lahham and colleagues demonstrate the diagnostic value and limits of ophthalmic ultrasound in the hands of nonspecialists; and McKinney et al. explore how AI might improve diagnostic accuracy in commonly performed procedures (mammography in this case).

Prem Subramanian, MD, PhD

Dr. Subramanian is Professor of Ophthalmology, Neurology, and Neurosurgery; Vice Chair for Academic Affairs; and Division Head, Neuro-Ophthalmology, at the Sue Anschutz-Rodgers UC Health Eye Center, University of Colorado School of Medicine, Aurora, CO.

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

Artificial Intelligence and Discrepant Histologic Diagnoses of Diminutive Colorectal Polyps

AI may help resolve diagnostic discrepancies between endoscopic and pathologic diagnoses of colorectal polyps ≤3 mm.

The histology of colorectal polyps can be accurately predicted during endoscopy using validated criteria, with pathology analysis as the gold standard. However, in a recent study, 15% of polyps ≤3 mm predicted to be adenomas with high confidence were interpreted as normal tissue on pathology (*Endoscopy* 2019 51; 221).

In the current study, investigators assessed whether a validated artificial intelligence (AI) clinical decision support tool could resolve such discrepancies. The dataset included 644 polyps ≤3 mm with previous high-confidence endoscopic diagnosis of adenoma. The results were as follows:

- The rate of discordance between endoscopic and pathologic diagnoses was 29%, with most polyps interpreted as hyperplastic polyp or normal mucosa.
- The AI diagnosis was discordant with endoscopic diagnosis for 11% of polyps, most of which were interpreted as normal mucosa.
- The AI system agreed with the endoscopic diagnosis in 90% of instances where endoscopy and pathology analyses were discordant.
- For the 99 polyps predicted as adenoma at endoscopy and reported as normal mucosa by pathology, AI agreed with the endoscopic diagnosis in 91%.

COMMENT

This study highlights a novel application of AI in colonoscopy. The higher concordance between endoscopic and AI-derived histologic diagnoses is not surprising, since both are based on the same visual criteria assessing polyp surface characteristics. An important observation is the substantial rate of discordance between endoscopic and pathologic diagnoses for polyps ≤3 mm, which the authors postulate could be due to errors in sample retrieval or processing. In such instances, the endoscopic diagnosis of adenoma should take precedence when determining postpolypectomy surveillance intervals.
— Charles J. Kahi, MD, MS

Shahidi N et al. Use of endoscopic impression, artificial intelligence, and pathologist interpretation to resolve discrepancies between endoscopy and pathology analyses of diminutive colorectal polyps. **Gastroenterology** *2019 Dec 18; [e-pub]. (https://doi.org/10.1053/j.gastro.2019.10.024)*

Ocular Ultrasound to Diagnose Posterior Chamber Pathology in the ED

In the hands of ED providers with varied ultrasound experience, ocular ultrasound was highly sensitive for retinal detachment and less accurate for vitreous hemorrhage and vitreous detachment.

Previous small studies have reported varying accuracy for emergency department (ED) point-of-care ultrasound (POCUS) to diagnose retinal detachment (*NEJM JW Emerg Med* Jul 2018 and *Acad Emerg Med* 2019; 26:16), but large studies assessing accuracy for multiple diagnoses have been lacking.

To determine the test characteristics of POCUS to diagnose retinal detachment, vitreous hemorrhage, and vitreous detachment, researchers enrolled patients being evaluated for these diagnoses at one of four EDs. Patients underwent POCUS performed by their ED provider and were then evaluated by an ophthalmologist blinded to the POCUS results, whose diagnosis was considered the criterion standard. The study involved 75 ED providers (20 attending physicians, 5 physician assistants, and 50 residents) with variable ultrasound experience. As part of the study, all the providers received a 30-minute lecture and 30 minutes of hands-on training in ocular ultrasound.

Among 225 enrolled patients, retinal detachment was diagnosed in 47 (21%), vitreous hemorrhage in 54 (24%), and vitreous detachment in 34 (15%). Sensitivity and specificity for POCUS were 97% and 88% for retinal detachment, 82% and 82% for vitreous hemorrhage, and 43% and 96% for vitreous detachment.

COMMENT

A strength of this study is that it likely included nonexperts in ultrasound. Although a POCUS exam is better than a nondilated fundoscopic exam, if your suspicion for retinal detachment is high, your patient needs emergent ophthalmologic evaluation, regardless of POCUS findings. In low-suspicion cases, a normal ultrasound (or one showing other vitreous pathology) might allow for discharge with next-day follow up, but ultimately POCUS is rarely going to change your management. ***— Benton R. Hunter, MD***

Lahham S et al. Point-of-care ultrasonography in the diagnosis of retinal detachment, vitreous hemorrhage, and vitreous detachment in the emergency department. **JAMA Netw Open** *2019 Apr 5; 2:e192162. (https://doi.org/10.1001/jamanetworkopen.2019.2162)*

Can Artificial Intelligence Improve Screening Mammography's Track Record?

AI's benefits in reading mammograms appear impressive, but broader testing is needed.

Even among expert radiologists, accuracy varies when interpreting screening mammograms. Artificial intelligence (AI) has shown promise in the setting of other medical imaging applications; thus, investigators used large U.K. and U.S. clinical datasets to create a "deep learning" model for identifying breast cancer (biopsy-confirmed cases diagnosed within 3 and 2 years of the index mammogram in U.K. and U.S. women, respectively), then compared the performance of their AI-based system with that of clinical radiologists in either country. In the U.K., screening is triennial, with each image interpreted by two clinicians; U.S. screening is biennial, with each image read by one clinician. To confirm generalizability of AI's utility, an AI system based on the U.K. data was tested with the U.S. data. Lastly, the overall performance of the AI system was compared with that of a subset of six independent U.S. radiologists.

Compared with U.K. and U.S. radiologists, the AI-based system reduced false positives by 1.2% and 5.7%, respectively, and reduced false negatives by 2.7% and 9.4% (*P*<0.005 for all comparisons). Compared with the six U.S. radiologists, the AI system performed significantly better (*P*=0.0002). Notably, one cancer case was identified with AI but missed by all six radiologists — and a second case was identified by all the radiologists but missed by AI.

COMMENT

Although the great majority of images in this analysis were generated with hardware from one manufacturer, these encouraging findings suggest AI might help triage mammograms by identifying those images requiring extra scrutiny by radiologists. However, computer-aided software that was introduced in the 1990s ultimately failed to improve mammogram performance despite initial enthusiasm (*NEJM JW Women's Health* Nov 2015 and *JAMA Intern Med* 2015 Nov; 175:1828). Before AI can be implemented clinically, its utility for interpreting screening mammograms should be assessed in broader populations of patients as well as radiologists, using all applicable imaging hardware. ***— Andrew M. Kaunitz, MD***

McKinney SM et al. International evaluation of an AI system for breast cancer screening. **Nature** *2020 Jan; 577:89. (https://doi.org/10.1038/s41586-019-1799-6)*