Invited Commentary | Diabetes and Endocrinology

Evaluating Artificial Intelligence Applications in Clinical Settings

Elaine O. Nsoesie, PhD

+ Related article

Author affiliations and article information are listed at the end of this article.

Artificial intelligence (AI)–based systems have been shown to reliably recognize cardiovascular disease risk[1] and diagnose conditions such as diabetic retinopathy[2,3] and melanoma[4] from medical images. These advances in image-based medical diagnosis have been widely publicized in the media and similar tools have been approved by the US Food and Drug Administration (FDA). In April of 2018, the FDA approved the first AI device to provide screening decision for a disease (ie, diabetic retinopathy) without assisted interpretation by a clinician.[5] Kanagasingam et al[6] evaluated a similar approach—a convolutional neural network algorithm, a deep learning method—for identifying diabetic retinopathy from medical images in a primary care setting in Midland, Western Australia. Their system correctly classified the 2 severe cases captured in the data (193 patients with diabetes), and misclassified 15 (false-positives) individuals as having diabetic retinopathy. The number of patients needing to be reviewed by an ophthalmologist was less than 10%. These findings demonstrate the potential for these systems to support efficient and improved care, while also highlighting the need for rigorous evaluation in clinical settings.

Most deep learning algorithms require large data sets for training, usually consisting of thousands or millions of images. Medical data sets of this magnitude are typically expensive to produce and annotate. Individuals developing AI diagnostic tools might therefore rely on whatever data are available to produce initial results. However, certain deficiencies might not be evident until an AI diagnostic tool is evaluated in a clinical setting because data sets used for training have been carefully curated to remove imperfect data samples. For example, a system trained only on high-quality images might provide incorrect diagnosis when classifying low-quality images or images affected by sheen or other defects present in real-world clinical settings, as observed by Kanagasingam and colleagues.[6] Also, evaluation of AI diagnostic tools in clinical settings will enable researchers and clinicians to ascertain its potential effect on patient outcomes and health care decisions. Problems identified can be corrected prior to deployment. Findings from these evaluations should also be published in peer reviewed literature to monitor progress and allow for comparison of different systems. There is currently a dearth of published studies on the evaluation of AI diagnostic tools used in clinical settings.

Of course, evaluating an AI diagnostic tool in a clinical setting does not guarantee generalizability of findings. The article by Kanagasingam et al[6] is based on a single algorithm evaluated in a single health care location. The authors acknowledge this limitation. Moving from good initial performance to a device that can be used across varied clinical settings might not be feasible in some cases. For example, AI diagnostic tools hold significant potential for improving health care in low-resource settings and regions where adequate medical infrastructure is lacking. However, observations made in a clinical context in a health care setting in a developed country might not be reproducible in a low-resource setting. This highlights the fact that different geographic regions and clinical settings might require tailored tools. Furthermore, training an AI diagnostic tool on a single data set or clinical setting might lead to outcomes that are dependent on a particular type of device used in capturing images or overrepresentation of a particular symptom or demographic group.[7]

Although multiple studies have demonstrated that AI can perform on par with clinical experts in disease diagnosis, most of these tools have not been evaluated in controlled clinical studies to assess their effect on health care decisions and patient outcomes. While AI tools have the potential to improve disease diagnosis and care, premature deployment can lead to increased strain on the health care system, undue stress to patients, and possibly death owing to misdiagnosis.

ARTICLE INFORMATION

Published: September 28, 2018. doi:10.1001/jamanetworkopen.2018.2658

Corresponding Author: Elaine O. Nsoesie, PhD, Institute for Health Metrics and Evaluation, University of Washington, 2301 Fifth Ave, Ste 600, Seattle, WA 98121 (onelaine@vt.edu).

Author Affiliation: Institute for Health Metrics and Evaluation, University of Washington, Seattle.

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0

2. 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(22):2402-2410. doi:10.1001/jama.2016.17216

3. Ting DSW, Cheung CY-L, 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(22):2211-2223. doi:10.1001/jama.2017.18152

4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

5. US Food and Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems: FDA news release. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm. Published April 11, 2018 Accessed August 7, 2018.

6. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M-L, Mehrotra A. Evaluation of an artificial intelligence–based grading of diabetic retinopathy in primary care. *JAMA Netw Open*. 2018;1(5):e182665. doi:10.1001/jamanetworkopen.2018.2665

7. Zou J, Schiebinger L. AI can be sexist and racist - it's time to make it fair. *Nature*. 2018;559(7714):324-326. doi:10.1038/d41586-018-05707-8