Doctopic: Analysis and Interpretation
tldigitalhealth-D-20-00425
PII:
Gold OA, awaiting license type

Deep learning models to detect hidden clinical correlates

See **Articles** page eXX

It is often said that a picture is worth a thousand words, a statement that is also true in the medical domain [A: edits ok for clarity?]. Over the past decade, deep learning algorithms applied to medical diagnostics have displayed the ability to accurately and precisely identify clinical phenotypes and diagnoses beyond their conventional usage. From fundoscopic images being used to evaluate blood pressure and cardiovascular risk to echocardiogram videos identifying a patient's age and sex, investigators have shown the ability for deep learning algorithms to capture systemic phenotypes that would be difficult for human experts to evaluate.[1,2] In *The Lancet Digital Health,* Joon-myoung Kwon and colleagues applied a deep learning algorithm to electrocardiograms (ECGs) obtained in the emergency department, out-patient department, and from admitted patients [A: edited in line with Article, ok?], and show good performance in identifying patients with anaemia.[3] With use of a convolutional neural network, the authors developed a deep learning algorithm with input from ECGs and demographic information to assess a regression task for haemoglobin concentration as well as a binary threshold for anaemia [A: edits ok?].

The work solidifies the claim that we can obtain additional value from the diagnostic testing and imaging that are already routinely obtained in medical care. Previous studies have shown that deep learning algorithms applied to ECGs can identify cardiovascular phenotypes, such as heart failure, hidden heart rhythms [A: term correct? Do you mean "hidden arrhythmia"?], and hypertrophic cardiomyopathy,[4–6] and the authors extend this direction to show deep learning on ECGs can predict systemic phenotypes beyond the heart. An important strength of this Article is the testing of the algorithm on an external dataset from another hospital, helping to avoid potential bias, confounding, or model overfitting, which can happen when an algorithm is only tested on an internal test dataset with the same characteristics as the training dataset. Additionally, the authors supported the claimed strength [A: ok?]of their deep learning algorithm by comparing their model with benchmark models on patient demographics and non-imaging characteristics.

Several challenges remain ahead to further understand and deploy artificial intelligence (AI) systems, such as the one proposed by Kwon and colleagues into routine clinical practice, as model performance and human interactions can vary in real-world settings. Similar to preclinical studies for pharmaceutical therapies, deep learning models are trained in pristine, controlled environments, which might not reflect real-world clinical settings. Real-world studies of AI systems have shown model performance can be significantly [A: is this statistical significance or can we edit to "substantially"?] affected by different hardware, seemingly trivial input settings, patient characteristics, and clinical workflows.[7] Particularly in resource-limited environments, understanding the routine clinical workflow, such as how patients will obtain treatment when obtaining additional diagnostic testing, is needed to establish the true effect of complicated interventions such as AI systems. We call upon the medical AI research community to release the code of their algorithms to allow for the broad testing of such systems in diverse settings to understand their limitations and biases. Ultimately, as with any important medical intervention, randomised controlled trials are needed to fully evaluate the effect of AI systems.

Equally important for widespread adoption of AI systems is increasing understanding of their decisions and the ability to interact, troubleshoot, and evaluate the diagnoses, as highlighted by the authors. Previous work has focused on human-interpretable models when applying AI systems for which the input is directly related to the diagnostic task,[8–9] however this [A: what does "this" refer to specifically?] is especially challenging when clinicians are asked to evaluate an AI system doing a task beyond the abilities of human experts. The requirement for understanding and trusting the determination of such an AI system is unavoidably higher. Although this question is still open, important corollaries include, "is this diagnosis biologically plausible?" and "is the input data necessary and sufficient to obtain the diagnosis?". Designing efficient human-AI interactions is an important and less explored component of this challenge.

Given the early stage in development of these algorithms, AI systems are often described in the setting of screening before confirmatory testing by conventional means. However, even incrementally

improving the diagnostic yield of a relatively inexpensive, routine medical test can have an impact on triage, costs, and the patient experience. The ability to screen for anaemia from ECG testing could minimise the need for painful phlebotomy and allow for earlier recognition of the need for blood transfusions.

The ongoing development of AI systems for early detection of disease is a promising avenue for obtaining additional value from medical testing. The pursuit of more precise and earlier diagnoses will allow for earlier detection of disease, more efficient triage of health-care resources, and help clinicians make complex decisions with additional diagnostic support.

We declare no competing interests. **[A: please confirm addition]**

*David Ouyang, James Zou

[A: please add email address for correspondence]

Department of Medicine, Stanford University, Stanford, CA, USA **[A: addition correct? Please add postcode for correspondence]** (DO); and Department of Biomedical Data Science, Computer Science, and Electrical Engineering **[A: add institution, city and country]** (JZ)

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:** 158–64.

2 Gorbhani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. *Digit Med* 2020; **3:** 10.

3 Kwon JM, Cho YH, Cho SH, Deep learning based artificial intelligence algorithm for detecting anemia using electrocardiogram. *Lancet Digital Health* 2020; **[Prod: to update]**

4 Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019; **25:** 70–74.

5 Ko WY, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol* 2020; **75:** 722–33.

6 Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019; **394:** 861–67.

7 Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020. https://dl.acm.org/doi/fullHtml/10.1145/3313831.3376718 (accessed May 26, 2020). **[A: ok addition of Article link?]**

8 De Fauw J, Ledsam JR, Romera-Paredes B. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Med 2018; **24:** 1342–50.

9 Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020; **580:** 252–56. **[A: updated with latest article info, ok?]**