Invited Commentary

# Leveraging Large Clinical Data Sets for Artificial Intelligence in Medicine

David Ouyang, MD; Christine M. Albert, MD, MPH

**The last decade** has seen extraordinary progress in the development and refinement of artificial intelligence (AI)–enabled tools for clinical applications.[1,2] In cardiovascular medicine, deep learning algorithms using data from echocardiography and electrocardiography (ECG) have demonstrated superhuman performance in tasks typically done by physicians[1,3] and identified subtle phenotypes invisible to a clinician's routine review.[4,5] Progress has been particularly quick in AI interpretation of ECGs, where the input data are well structured and the medical infrastructure for data acquisition and storage is relatively standardized across health systems. Already, there have been randomized clinical trials[4] showing the effects of ECG-based AI algorithms on patient care. Compared with other forms of medical imaging, ECG waveforms routinely undergo signal processing and standardization that makes generalization across systems and devices easier. The ongoing development of AI systems for interpretation of ECGs is a promising avenue for extracting additional value from this readily available medical test.

+
Related article

In this issue of *JAMA Cardiology*, Hughes and colleagues[6] implemented and tested a deep learning neural network for interpreting 12-lead ECGs to classify 38 common electrophysiological diagnoses. Extending beyond prior work using single-lead ECGs,[7] the authors leveraged a large data warehouse of 12-lead ECGs and labels from clinical cardiologist interpretations within an academic medical center for model training. From diagnoses associated with rhythm and conduction (such as atrial fibrillation, bundle branch blocks, and pacing) to anatomic characterization (such as left ventricular hypertrophy to presence of infarct), the authors show their AI model performs with an area under the curve greater than 0.96 in most diagnoses commonly made from 12-lead ECGs. Recognizing the variation in clinician interpretations, the authors performed a second validation study using gold-standard diagnoses made by a committee of expert electrophysiologists. In these analyses, the range of aggregate F1 scores in the AI model interpretation (0.60-0.88) was superior to that of both mean clinician overreads (0.42-0.77) and initial interpretations from the MUSE Cardiology Information System (GE Healthcare) (0.24-0.69) for most diagnoses.

This important work[6] exemplifies the significant value that can be obtained from applying deep learning algorithms to large data sets of extant clinical information. Because deep learning algorithms require vast amounts of data to learn the associations between input and diagnosis, leveraging large clinical data sets and existing physician interpretations alleviates the need for costly prospective data labeling by clinicians. However, a limitation to the use of historical clinical data for model training is the variation in clinician interpretation. Multiple prior works applying AI to medical data both inside and outside of cardiology have shown that expert clinician interpretations exhibit discordance, necessitating the use of consensus committees or blinded retesting.[1,8] While AI models can capture subtle patterns that identify disease, this potential is bounded by the quality of expert labels (clinician interpretations) used during training. This limitation is again evident in the study by Hughes and coworkers,[6] such that 20% of the time the unedited MUSE diagnosis agreed better with the expert consensus committee than the overreading clinical cardiologist. Human attention is fatigable and can be prone to error, highlighting the need and opportunity for accurate automated approaches to the interpretation of medical data.

Several challenges remain before we can deploy deep learning models into clinical practice, including understanding model interpretations, evaluating generalizability to real-world settings, and eventual randomized studies. To overcome the challenge of understanding how deep learning models come to its conclusions, the authors[6] applied local interpretable model agnostic explanations,[9] an advanced model interpretation technique, to identify key features on the ECG waveform that resulted in the model interpretation. Understanding how AI models come to their conclusions is critical for widespread adoption of AI systems, since increasing understanding of AI decisions allows for humans to interact with, troubleshoot, and evaluate the proposed diagnoses.

The need for understanding and trusting AI decision support systems is critical, because medical decisions are on the line and optimal human-AI interactions is an important but less-explored component of deployment. With conventional decision support already present in ECG acquisition systems, novel AI-based systems need to integrate conventional infrastructure and human clinician input when needed to provide the most accurate diagnoses together. In the study by Hughes and coworkers,[6] the performance of the AI system was relatively weaker for sinus rhythm (despite being the most common label in their data set), and some rule-based tasks (measurement of QTc and identification of axis) might be better suited to conventional decision support systems. Understanding where AI should start and clinicians should end will be an iterative process as we learn how to harness the potential of new technology. As with any important medical intervention, randomized clinical trials are needed to fully evaluate the effects of AI systems, which should be encouraged in the US Food and Drug Administration approval process for AI technologies.[10]

The pursuit of more precise and earlier diagnosis through advances in AI will allow the earlier detection of disease, more efficient triage of health care resources, and additional complex decision support for clinicians. Electrocardiograms have particularly high potential for implementation of AI systems. Given the availability of standardized infrastructure, use in many clinical scenarios, and interpretations by clinicians with a wide range of expertise in the interpretation, ECGs are a key example of underused nontabular clinical data for which AI models may be revolutionary.

## ARTICLE INFORMATION

**Author Affiliations:** Department of Cardiology, Cedars-Sinai Medical Center, Los Angeles, California (Ouyang, Albert); Division of Artificial Intelligence in Medicine, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California (Ouyang).

**Corresponding Author:** David Ouyang MD, Department of Cardiology, Cedars-Sinai Medical Center, 127 S San Vincente Blvd, A3100, Los Angeles, CA 90048 (david.ouyang@cshs.org).

**Published Online:** August 4, 2021. doi:10.1001/jamacardio.2021.2878

**Conflict of Disclosures:** None reported.

## REFERENCES

**1**. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*. 2020;580(7802):252-256. doi:10.1038/s41586-020-2145-8

**2**. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6

**3**. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med*. 2018;1:6. doi:10.1038/s41746-017-0013-1

**4**. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat Med*. 2021;27(5):815-819. doi:10.1038/s41591-021-01335-4

**5**. 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(10201):861-867. doi:10.1016/S0140-6736(19)31721-0

**6**. Hughes JW, Olgin JE, Avram R, et al. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiol*. Published August 4, 2021. doi:10.1001/jamacardio.2021.2746

**7**. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65-69. doi:10.1038/s41591-018-0268-3

**8**. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900-908. doi:10.1038/s41591-020-0842-3

**9**. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Published August 9, 2016. Accessed June 25, 2021. https://arxiv.org/abs/1602.04938

**10**. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. 2021;27(4):582-584. doi:10.1038/s41591-021-01312-x