Editorial

Can Left Atrial Enlargement Be Estimated From the Electrocardiogram? An Artificial Intelligence Approach to Screening Risk of Atrial Fibrillation and Heart Disease

Vijay S. Chauhan, MD
Division of Cardiology, Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada

See article by Chou et al., pages 585-594 of this issue.

Screening for Left Atrial Enlargement: Rationale and Challenges

Assessing the risk of cardiovascular disease is challenging, but essential to appropriate clinical work-up and management. Among the plethora of cardiac risk markers available to primary care physicians, internists, and cardiologists, left atrial enlargement (LAE) is rarely considered, but it has been shown to independently predict atrial fibrillation (AF), heart failure, coronary artery disease, and all-cause mortality. In fact, severe LAE (left atrial diameter [LAD] ≥ 50 mm) is associated with a 4-fold increased risk of new-onset AF. As such, there may be value in screening for LAE even before it is apparent on echocardiography. The 12-lead electrocardiogram (ECG) is a readily available diagnostic tool, but standard criteria for LAE are insensitive because only a few P-wave dimensions in leads II and V1 are considered. In the past decade, artificial intelligence using deep learning has been applied to a variety of complex biological signals, including the ECG, to enhance the diagnosis and prognosis of cardiovascular disease. Given the nuances in abnormal P-wave morphology and its company with the QRS-T wave, analysis of the 12-lead ECG waveform with the use of deep learning may improve ECG-based detection of LAE.

Detecting Severe LAE With Deep Learning—Enabled ECG Models

In this issue of the Canadian Journal of Cardiology, Chou et al. develop a deep learning—enabled ECG model to detect LAE ≥ 50 mm and its long-term cardiovascular risk. Their single-centre, retrospective study included 382,593 consecutive adults each with a 12-lead ECG and echocardiogram performed within 2 weeks of one another. The echocardiogram provided the gold standard for classifying LAE (ie, LAD in the parasternal long-axis view). A convolutional neural network (CNN) was trained using 1 of 4 different 10-second ECG data sets—12-lead P-QRS-T waves, 12-lead segmented P-waves, and single-lead I or V2 P-QRS-T waves—to discriminate LAD ranging in size from 38 to 50 mm (ECGs for training 35%, internal validation 15%, testing 50%). External validation was performed using a separate cohort of 91,425 patient ECGs from another hospital. Cardiovascular events associated with LAE ≥ 50 mm were assessed from administrative International Classification of Diseases codes.

The main study finding was that all 4 deep learning ECG models performed similarly in classifying LAE, with performance improving as LAE cutoff points increased from 38 mm (receiver operating characteristic (ROC) area under curve (AUC) 0.74-0.79) to 50 mm (ROC AUCs 0.86-0.88). The models performed equally well with external validation (ROC AUCs 0.85-0.88). When all 4 deep learning ECG models were developed separately on the 338,264 sinus rhythm ECGs vs 60,766 non-sinus rhythm ECGs (eg, AF, flutter, frequent premature atrial beats), model performance was better for the sinus rhythm ECGs in detecting LAE ≥ 50 mm (ROC AUCs 0.85-0.73). With deep learning ECG—detected LAE ≥ 50 mm, the hazard ratios (HRs) for predicting new-onset AF (HR 9.7-10.1), heart failure (HR 4.6-5.3), stroke (HR 1.5-1.6) and acute myocardial infarction (HR 1.7-2.1) over 1-year follow-up was similar with all 4 ECG models. These HRs were higher than those for echocardiography-detected LAE ≥ 50 mm, due in part to a lower predicted event rate among patients classified as LAE < 50 mm using the CNN ECG models.

Strengths of the Study

Overall, this is a well designed, rigorous study with novel findings that build on the literature of deep learning—enabled ECG analysis for diagnostic screening and prognostication. The selection of the database used for training significantly affects the quality and generalisability of machine learning.
models. Here, a large, diverse real-world ECG database was used that included sinus, nonsinus and even noisy ECGs unique to each patient. In addition, the model was externally validated to improve generalisability. The optimal ECG input format for deep neural network training remains an open question, but the signal-based ECG used in this study permitted single lead training without ECG grid interference, and this may offer higher fidelity than 2-dimensional ECG images. The CNN architecture itself is well described and necessary for the multidimensional ECG waveform analysis that would not be possible with supervised machine learning. With CNN-assisted diagnosis comes the concern of whether the output can be trusted and interpreted by the clinician and end-user. Chou et al.\(^6\) demonstrate better performance with sinus rhythm ECGs than non—sinus rhythm ECGs, suggesting that the P-wave is relevant in their CNN-based classification of LAE \(\geq 50\) mm. Among all ECGs, the 12-lead P-QRS-T waves only marginally improved model performance compared with the 12-lead segmented P waves, further supporting the relevance of the P-wave in model classification.

**Limitations of the Study**

Before considering the clinical relevance and application of this study, it is important to understand its shortcomings. Foremost, although a large number of ECGs were included with random allocation to training, validation, and testing, significant class imbalance exists in these cohorts because only 5% of all ECGs were from patients with LAE \(\geq 50\) mm. Standard classification metrics such as ROC AUC do not adequately quantify model performance when data sets are imbalanced. Instead, the precision-recall AUC and F1 score (ie, harmonic mean of precision and recall) are more accurate, and in this study both were low (F1 score 0.27, precision-recall AUC 0.31) indicating poor performance in the internal testing and external validation cohorts. Data augmentation strategies\(^8\) can help to improve class imbalance, but these were not considered. Data augmentation increases the number of cases by adding slightly modified copies of already existing data or newly created synthetic data from existing data. This approach was used by Liao et al.\(^10\) when developing a CNN model to classify the type 1 Brugada pattern from 12-lead Holters because only a limited number of cases were available for model training. Another important consideration is selecting the optimal deep learning architecture for training and inference. Although CNNs have dominated the field until now, recent advancements in attention-based generative pretrained transformers and vision transformers offer promising avenues for improving machine learning—enhanced ECG classification.\(^9,12\) Pretraining transformer models on large unlabelled data sets may enable the creation of specialised models when there are limited cases or rare diseases. Using a similar approach with transfer learning, Liu et al.\(^11\) trained a CNN to accurately classify the type 1 Brugada pattern from 12-lead ECGs. Their model parameters from a source model trained on 12-lead ECGs of right bundle branch block were transferred to a target model trained on type 1 Brugada ECGs to enhance its performance.

Besides model development, the study has limitations in how clinical metrics and outcomes were evaluated. Although a large real-world ECG database was collected, there may still be referral bias among patients undergoing both ECG and echocardiography within 2 weeks. As such, the performance of the model in an unselected population for LAE screening may be quite different, and class imbalance would become even more significant. Second, detecting LAE based on LAD is far less sensitive than using LA volume, which is now more commonly reported. Third, LAE often does not exist in isolation, but rather coexists with structural heart disease (eg, mitral regurgitation, cardiomyopathy, left ventricular hypertrophy) and clinical conditions (eg, congestive heart failure, AF, hypertension). Without adjusting for these confounders, it is unknown whether LAE independently predicts cardiovascular events or is simply a marker of underlying disease that is driving adverse outcomes. Finally, the interpretation of the higher hazard ratio for cardiovascular events with the CNN ECG model—detected LAE \(\geq 50\) mm compared with echocardiography-detected LAE \(\geq 50\) mm is complex and needs clarification. The CNN ECG models classified 5- to 6-fold more patients with LAE \(>50\) mm compared with echocardiography, indicating many false positives. However, the cardiovascular event rate with CNN ECG—detected LAE \(<50\) mm was lower (high negative predictive value of 0.96) compared with echocardiography. This difference contributed to the higher hazard ratio with the CNN ECG model—detected LAE \(\geq 50\) mm, especially for AF and heart failure.

**Future Considerations for LAE Screening**

Notwithstanding these limitations, the study findings provide a framework for future development in this field that may ultimately pave the way for ECG-enabled LAE screening and prognostication. Interestingly, the performance of the single leads I and V2 ECG models (accuracy \(\sim 77\%)\), ROC AUC \(\sim 0.85\) in discriminating LAE \(\geq 50\) mm was similar to that of the 12-lead ECG model (accuracy 78%, ROC AUC 0.88). However, single-lead recordings, for example, from wearables, may not be relevant in detecting LAE, unlike arrhythmias, and may compromise ECG signal quality. Model performance also did not improve significantly with inclusion of the QRS-T complex. In contrast, the machine learning—enabled single-lead II ECG model of Hsu et al.\(^14\) identified LAE \(>40\) mm based on features from the QRS complex rather than P-wave. It is unexpected that a full 12-lead P-QRS-T disclosure was not more discriminatory for LAE compared with the single-lead P-wave. The QRS-T complex would be influenced by ventricular pathology (eg, left ventricular hypertrophy, cardiomyopathy), which is an important cause of LAE. Furthermore, P-wave morphology would be affected by LA fibrosis burden independently from LA volume, leading to changes in P-wave morphology in multiple leads. As such, deep learning—enabled 12-lead P-QRS-T ECG models may hold the greatest promise in discriminating LAE.

The premise of this study is that deep learning ECG models will detect severe LAE in a relatively asymptomatic population without atrial arrhythmias or cardiovascular disease to inform early work-up and treatment. Although this is a rational premise, it may not be clinically relevant because the prevalence of this population is very low (\(\sim 3\%)\), and it is uncertain whether early intervention will actually change
cardiovascular outcomes. Identifying a cohort with a higher pretest probability of severe LAE, such as those with more cardiovascular risk factors, may improve diagnostic yield. Ultimately, echocardiography will still be required to confirm deep learning—ECG detection of severe LAE. Therefore, minimising false positives with appropriate patient selection and a robust deep learning model are essential for cost-effective screening.

**Conclusion**

Chou et al. are to be commended for their novel application of deep learning—enabled ECG analysis with inclusion of the P-wave to detect severe LAE. Future studies in diverse populations will be needed to confirm deep learning model performance and the patient populations best served by this screening tool. Although in its infancy, this point of care approach to characterising abnormal atrial electroanatomic substrate has the potential to enhance the prediction of heart failure, AF, stroke, and even response to AF catheter ablation. Whether this prompts early intervention and changes clinical outcomes remains to be determined.

**Funding Sources**

The authors have no funding sources to declare.

**Disclosures**

The authors have no conflicts of interest to disclose.

**References**


