Expanding Composite Disease Labels Improves ECG Deep Learning Model Performance for Structural Heart Disease Detection

Pierre Elias, MD[1,2,3], Linyuan Jing, PhD[3], Joshua Finer, MS[3], Dustin Hartzel[3], Christopher Kelsey[3], Daniel Rocha[3], Jeffrey Ruhl, MS,[3] Christopher M. Haggerty, PhD[2,3], and Timothy J. Poterucha, MD[1]

[1] Seymour, Paul, and Gloria Milstein Division of Cardiology, Department of Medicine, Columbia University Irving Medical Center [2] Department of Biomedical Informatics, Columbia University [3] NewYork-Presbyterian Hospital

Background. Deep learning models trained on electrocardiogram (ECG) voltage data have demonstrated remarkable performance in detecting heart diseases. These models can also learn features that extend beyond a single detection target, as shown by recent studies that trained models based on composite disease labels. For example, the 'ValveNet' model can detect three left-sided valvular heart diseases (aortic stenosis, aortic regurgitation, or mitral regurgitation) and achieved superior performance compared to a model trained to detect individual valve diseases. This finding raises the question of how binary label classification of a larger composite label compares in performance to binary classification of individual labels.

Methods. We trained a new model called 'EchoNext' using 601,771 unique ECG-echo pairs by extending the ValveNet model to include a broader composite label for structural heart disease. In addition to left-sided valve disease, this expanded label comprised left and right ventricular systolic dysfunction, increased left ventricular wall thickness, pericardial effusion, increased pulmonary artery pressure, and right-sided valvular regurgitation. We evaluated the performance of both models on a test set of 1,997 unique patients with an ECG and matched echocardiogram. All studies were clinically acquired at Columbia University Medical Center between January and February 2023, and thus temporally distinct from data used to train both models.

Results. Within this test set, 44% of patients satisfied the definition for structural heart disease, while 11% met criteria for left-sided valve disease. EchoNext (area under the receiver operating characteristic curve (AUROC) = 0.846, area under the precision-recall curve (AUPRC) = 0.827) outperformed ValveNet (AUROC = 0.792, AUPRC = 0.344) with respect to the labels used in training each model (structural heart disease vs. left-sided valve disease, respectively). Notably, this performance difference persisted when the ValveNet model was used to classify the broader structural heart disease label (AUROC = 0.773, AUPRC = 0.716). These findings provide evidence for additional performance gains with expansion of the composite disease label for training. We also compared the sensitivities of both models with respect to the original left-sided valve disease label and found that EchoNext (0.768) was again comparable to ValveNet (0.693) for that specific label.

Conclusion. Some patterns that are detectable on ECG waveforms by deep neural networks may be shared across various structural heart diseases. Thus, combining these endpoints in model training can yield improved performance, as shown by the superior overall performance of EchoNext compared to ValveNet. These models have clinical potential to improve timely recognition and treatment of structural heart diseases.