HHS Public Access

Author manuscript

NEJM AI. Author manuscript; available in PMC 2025 August 06.

Published in final edited form as:

NEJM AI. 2025 July ; 2(7): . doi:10.1056/aioa2401033.

An Electrocardiogram Foundation Model Built on over 10 Million Recordings

Jun Li, B.S.[1,2,3], **Aaron D. Aguirre, M.D., Ph.D.**[4,5], **Valdery Moura Junior, Ph.D.**[5,6], **Jiarui Jin, B.S.**[1,2,3], **Che Liu, M.S.**[7], **Lanhai Zhong, B.S.**[8], **Chenxi Sun, Ph.D.**[5,9], **Gari Clifford, Ph.D.**[10,11], **M. Brandon Westover, M.D., Ph.D.**[5,9], **Shenda Hong, Ph.D.**[1,2,3]

[1]National Institute of Health Data Science, Peking University, Beijing

[2]Institute of Medical Technology, Peking University Health Science Center, Beijing

[3]Institute for Artificial Intelligence, Peking University, Beijing

[4]Cardiology Division, Massachusetts General Hospital, Boston

[5]Harvard Medical School, Boston

[6]Department of Medicine, Massachusetts General Hospital, Boston

[7]Department of Computing, Data Science Institute, Imperial College London

[8]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

[9]Department of Neurology, Beth Israel Deaconess Medical Center, Boston

[10]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta

[11]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta

Abstract

BACKGROUND: Artificial intelligence (AI) has demonstrated significant potential in electrocardiogram (ECG) analysis and cardiovascular disease assessment. Recently, foundation models have played a remarkable role in advancing medical AI, bringing benefits such as efficient disease diagnosis and cross-domain knowledge transfer. The development of an ECG foundation model holds the promise of elevating AI-ECG research to new heights. However, building such a model poses several challenges, including insufficient database sample sizes and inadequate generalization across multiple domains. In addition, there is a notable performance gap between single-lead and multilead ECG analysis.

METHODS: We propose a general-purpose ECG foundation model (ECGFounder), which leverages real-world ECG annotations from cardiologists to broaden the diagnostic capabilities of ECG analysis. ECGFounder was built on 10,771,552 ECGs from 1,818,247 unique subjects with 150 label categories from the Harvard–Emory ECG Database, enabling comprehensive cardiovascular disease diagnosis. The model is designed to be both an effective out-of-the-box solution and easily fine-tunable for downstream tasks, maximizing usability. Importantly, we

Dr. Hong can be contacted at hongshenda@pku.edu.cn or at 38 Xueyuan Rd, Haidian District, Beijing, 100191, China.
Drs. Clifford and Westover contributed equally to this article.

extended its application to reduced-lead ECGs, particularly single-lead ECGs. ECGFounder is therefore applicable to various downstream tasks in mobile and remote monitoring scenarios.

RESULTS: Experimental results demonstrate that ECGFounder achieves expert-level performance on internal validation sets, with area under the receiver operating characteristic curve (AUROC) exceeding 0.95 for 80 diagnoses. It also shows strong classification performance and generalization across various diagnoses on external validation sets. When fine-tuned, ECGFounder outperforms baseline models in demographic analysis, clinical event detection, and cross-modality cardiac rhythm diagnosis, surpassing baseline methods by 3 to 5 points in the AUROC.

CONCLUSIONS: The ECG foundation model offers an effective solution, allowing it to generalize across a wide range of tasks. By enhancing existing cardiovascular diagnostics and facilitating integration with cloud-based systems, which analyze ECG data uploaded from wearable devices, it significantly contributes to the advancement of the cardiovascular AI community and enables management of cardiac conditions. (Funded by the National Science Foundation and others.)

Introduction

The electrocardiogram (ECG) is a crucial diagnostic and monitoring tool that measures and records cardiac electrical activity using electrodes placed on the skin.[1] However, comprehensive interpretation of ECGs is complex and time-consuming, requiring extensive training. A typical ECG expert undergoes nearly 10 years of training, including medical school, residency, and specialized ECG training.[2] Recent advances in deep learning, together with efforts to assemble relatively large databases of ECGs, have led to significant progress in the field, extending ECG analysis beyond traditional medical domains.[3-6] However, owing to the lack of large-scale publicly available ECG databases with diverse diagnostic information, developing general-purpose artificial intelligence (AI)–ECG models remains a challenging task. Existing models are typically limited to specific diagnostic tasks and datasets — limitations that make it difficult to extend the model to real-world ECG analysis.

Foundation models, with their strong generalization capabilities, provide a promising approach for enhancing the performance of AI-ECG models through transfer learning. Recently, these models have seen significant advancements in the field of medical AI. In retinal disease diagnosis, the RETFound model, through pretraining on a large number of retinal images, has achieved excellent performance across various clinical diagnostic tasks.[7] In computational pathology, the UNI model was trained on a vast number of whole-slide images, reaching expert-level performance in multiple cancer diagnostic tasks.[8] These studies defined foundation models as large-scale AI models trained on extensive datasets, capable of adapting to a wide range of downstream tasks. Specifically, they meet the following criteria: pretraining on a large-scale dataset, an enormous number of parameters, and the ability to generalize across a wide range of downstream tasks.[9]

There have been several claims made in the literature on developing foundation models for the ECG.[10,11] However, owing to limitations of existing ECG databases in terms of sample size, patient numbers, the variety of diagnoses, and, importantly, patient demographics, these models have not addressed the challenges of diversity across regions, ethnic groups, and

diagnostic variations.[12] Moreover, to qualify as foundational, the model must be capable of generalizing to multiple domains outside the initial training paradigm.[13] In addition, current ECG models have significant performance degradation when applied to single-lead ECGs versus multilead ECGs.[14]

We present ECGFounder — a foundation model trained on over 10 million clinically annotated ECGs across all known classifications, capable of diagnosing 150 cardiac abnormalities. As the most comprehensive ECG foundation model to date, it performs effectively across a wide range of tasks in various domains. ECGFounder provides the medical AI community with an accessible, high-performing model that supports fine-tuning. Unlike traditional AI-ECG models trained from scratch, it achieves superior results through fine-tuning, advancing AI-ECG development.

To address the challenges of incomplete annotations in real-world data, we introduced a novel method for preprocessing and training on these annotations, ensuring robust performance even under suboptimal conditions (Fig. 1). Moreover, by training the single-lead ECG model based on lead augmentation, we can maintain high diagnostic performance on single-lead ECGs as well. We validate the model's performance on internal and external test sets, where it consistently matches expert-level diagnoses. In downstream task fine-tuning, we demonstrate ECGFounder's versatility in addressing various tasks, including demographics detection, clinical event detection, and cross-modality diagnosis (Fig. 1B). Specifically, we evaluate ECGFounder on various downstream clinical tasks, such as ECG age regression and classification, sex detection, chronic kidney disease (CKD) detection, chronic heart disease (CHD) detection, left ventricular ejection fraction (LVEF) regression and abnormality classification, N-terminal pro–B-type natriuretic peptide (NT-proBNP, a biomarker indicating heart failure and cardiac stress) regression and abnormality classification, and atrial fibrillation (AF) detection based on photoplethysmography (PPG). The results highlight ECGFounder's potential as a foundational model for the further development of AI-ECG models.

Methods

DATASETS AND PREPROCESSING

Our dataset, the Harvard–Emory ECG Database (HEEDB), is the largest open-access ECG dataset, containing 10,771,552 expert-annotated ECGs from 1,818,247 subjects.[15] These predominantly 10-second, 12-lead clinical ECGs are paired with cardiologist and technician annotations, describing morphology, rhythm, and diagnostic information. Cardiologists used the Marquette 12SL ECG analysis program (GE Healthcare), version 4, for annotations, which provides waveform parameters for doctors' reference.[16] We used regular expressions to parse annotations and tally independent labels, removing irrelevant phrases after reviewing with doctors and defining 150 meaningful labels, including diagnostic, rhythm, and morphological information (see Table S4 in the Supplementary Appendix for details).

Our external validation data comprised three large ECG databases: the Clinical Outcomes in Digital Electrocardiology (CODE)-test database,[17] the PTB-XL database,[18] and the PhysioNet Challenge 2017 database.[19] The CODE-test database contains ECG records from

827 patients across 811 municipalities in Brazil, collected by the Telehealth Network of Minas Gerais. There are six common arrhythmia labels in this ECG database, annotated by experienced ECG experts.

The PTB-XL dataset contains 21,837 clinical ECGs from 18,885 patients in Germany. Each ECG is a 10-second, 12-lead recording. Labels were reviewed and verified by two cardiologists. This dataset is currently one of the best publicly accessible ECG collections, both in terms of the number of samples and the quality of labels.[18]

The PhysioNet Challenge 2017 is a large single-lead ECG dataset. The ECG recordings were collected using AliveCor devices.[19] The training set includes 8528 single-lead ECG recordings, with durations ranging from 9 to 60 seconds, while the test set contains 3658 ECG recordings of similar lengths. The dataset requires classification of ECG recordings into normal rhythm, AF rhythm, other rhythms, and noise.

In addition, we utilized the Medical Information Mart for Intensive Care (MIMIC) IV ECG database to finetune ECGFounder for various downstream tasks. The MIMIC-IV-ECG dataset is part of the MIMIC series, focusing on the collection and analysis of ECG data.[20,21] MIMIC-IV-ECG originates from real clinical settings at Beth Israel Deaconess Medical Center in Boston, and contains 800,035 clinical ECGs from 161,352 patients treated in the intensive care unit. Moreover, ECG recordings in the dataset can be linked to the electronic health records of patients from the MIMIC emergency department database, allowing the association with clinical conditions. We used this linkage to explore downstream tasks, including detection of age, sex, CKD, CHD, LVEF, and NT-proBNP to evaluate the performance improvement of the fine-tuned ECGFounder model in detecting diseases other than arrhythmias and clinical events. More details about the split and labels of the MIMIC-IV-ECG database can be found in Table S5. Proportions of positive and negative cases naturally reflect clinical prevalence but were not deliberately controlled or balanced.

To explore ECGFounder's generalization capabilities on other similar physiological signals, we fine-tuned and evaluated it using the DeepBeat dataset, a PPG-based AF detection dataset.[22] The dataset includes over 500,000 labeled signals from more than 100 individuals.

For data preprocessing, unreadable files, missing data, and unmatched data were excluded. Our final development dataset includes 7,519,035 ECGs from 1,319,128 patients, and the test dataset includes 834,926 ECGs from 146,570 patients. We applied linear interpolation to resample ECG frequencies to 500 Hz. We used a high-pass filter with a cutoff frequency of 0.5 Hz to suppress residual baseline drift and a second-order 50 Hz Butterworth low-pass filter to reduce high-frequency noise. A 50/60-Hz notch filter was utilized to eliminate electrical interference. For ECG records longer than 10 seconds, we extracted 10-second windows in sequence. If a sequence was less than 10 seconds, we applied zero padding. We normalized all signals using the mean and standard deviation of each individual signal segment before inputting them into the model.

MODEL ARCHITECTURE

To establish the model, we used an architecture tailored for ECG, capable of learning generalizable representations from large-scale ECG datasets. The increase of ECG data and the number of leads meant that the model must not only consider temporal information, but also spatial relationships (i.e., interactions between different leads and the overall pattern of cardiac electrical activity). This is crucial for ensuring that the ECG foundation model is applicable to real-world clinical ECG scenarios, as it mitigates the impact of nonuniform ECG durations and missing leads in the training dataset.

Considering these factors, we built our model architecture based on our previously proposed Net1D.[23] It is built on top of the simple regular networks (RegNet) architecture.[24] This structure begins with a stagewise network design in which each stage consists of a set number of blocks and channels that scale with network depth, allowing us to expand and design the blocks and channels of the ECG foundation model. Unlike traditional uniform scaling across layers, RegNet employs a quantized linear model to predict optimal widths and depths, ensuring efficient performance across a range of model sizes. The initial layers focus on capturing low-level features with fewer channels, which gradually increase as the network deepens, thus optimizing computational efficiency and model capacity. Then the model utilizes a series of bottleneck blocks that combine group convolutions and channel-wise attention mechanisms, enhancing the richness of representation in both temporal and spatial dimensions while controlling model complexity. This tailored configuration makes the model suitable for ECG data. More details about the model architecture can be found in section S3.1.

TRAINING WITH REAL-WORLD ANNOTATIONS: NOISY, IMBALANCED, POSITIVE UNLABELED

Unlike conventional ECG diagnostic models that typically use single-label classification methods, we employed multilabel classification during the training phase of the ECG foundation model. This approach aligns more closely with clinical practice, where an ECG diagnosis often consists of multiple diagnostic labels. For example, an abnormal ECG diagnosis might be something like: “normal sinus rhythm | premature ventricular complexes | premature ectopic complexes.” During training, utilizing a multilabel classification approach provides our model with rich semantic value and facilitates generalization to different annotation systems.

However, the nature of multilabel annotations presents significant challenges in ECG diagnostics. Owing to the high cost and complexity of annotation, cardiologists typically annotate only a limited subset of potential diagnoses for each ECG. Although a skilled cardiologist might annotate three classes simultaneously, the actual number of ECG diagnostic categories far exceeds this number.[25] This results in a dataset with numerous unlabeled instances that are not necessarily negative samples but rather potentially positive samples that cardiologists have not explicitly labeled.

To address this, we introduced a positive unlabeled (PU) learning method. PU learning is defined as a scenario where labeled positive samples are correctly identified, but unlabeled

samples are not necessarily negative examples.[26] Conventional multilabel classification methods typically treat unlabeled categories as negative by default, but this assumption does not hold true for ECG data.

Positive unlabeled learning in the ECG context leads to a severe imbalance in the predicted probability distribution: an ECG usually contains a few positive diagnostic labels, with the remaining labels being treated as negative. This creates a severe positive–negative imbalance, where negative samples far outnumber positive samples for each label. When using conventional loss functions, such as the binary cross-entropy loss function, the model tends to learn from simpler samples, namely the true negative samples, while the more challenging samples, namely the false negative samples (which may in fact be missed positives), are harder to fit. As a result, the model's predicted probabilities tend to skew toward 0 rather than 1, diminishing its ability to detect clinically important abnormalities.

To enhance the model's ability to identify potentially missed positive samples, we improved the multilabel classification loss function by dynamically adjusting the weights of positive and negative samples. Our key insight is that for missed positive labels, a well-trained multilabel model's predicted probability should be close to 1 rather than 0. Therefore, we reduced the loss weights for negative labels predicted by the model to have a high probability (close to 1), acknowledging that these may be positive labels that simply were not annotated. The loss function is given by:

$$\mathscr{L} = -\left(\gamma - p\right)p^2$$

Here, γ is a hyperparameter of the model and p is the predicted probability of the model. In this case, it is set to $\gamma = 1.5$, which we find optimally balances the weights, allowing the model to learn good representations of both positive and negative samples. Our model training used AdamW to minimize the loss function, with an initial learning rate set to 0.001. The learning rate decayed by a factor of 10 every 5 epochs. The trainable temperature parameter was initialized to 0. Training spanned a maximum of 20 epochs, with early stopping based on validation loss. We used a batch size of 1024.

TRAINING A SINGLE-LEAD ECG MODEL BASED ON LEADS AUGMENTATION

In recent years, portable and wearable ECG-device development has revolutionized continuous cardiac monitoring, offering a noninvasive method for real-time assessment. Beyond the conventional diagnosis of cardiac arrhythmias, another critical challenge in this field is accurately detecting and interpreting ECG axis deviation on single-lead ECGs from wearable devices (typically lead I), which can significantly impact the diagnosis of various cardiac abnormalities. For instance, left axis deviation can provide additional insights into diagnosis, such as left ventricular hypertrophy, left bundle-branch block, left anterior fascicular block, preexcitation syndromes, and inferior myocardial infarction (MI).[27]

To address this issue, we developed a novel training method utilizing lead-augmented wearable ECG models. By systematically enhancing standard 12-lead ECG data, we

simulated various clinical scenarios of axis inversion, thereby enhancing the model's robustness and versatility. Understanding the relationship between ECG vectors and leads is crucial for this method. The cardiac electrical activity generates a vector representation of cardiac signals, captured by different ECG leads placed on the body. Each lead provides a unique perspective of the cardiac electrical axis, offering a comprehensive view when combined. The standard 12-lead ECG system includes limb leads (I, II, III, augmented voltage right [aVR], augmented voltage left [aVL], augmented voltage foot [aVF]) and precordial leads (V1 to V6), with each lead representing a specific projection of the cardiac electrical vector.[28] Specifically, we primarily utilize ECG signals from limb leads. Based on the angular position of each limb lead's axis relative to the heart, we consider lead I as the center of the semicircle (i.e., 0°) and calculate the signals from all leads around the semicircle (i.e., from −90° to 90°), thereby obtaining six additional leads (aVL, –aVR, II, –III, aVF, –aVF). We then trained a model for wearable ECG devices, extracting the lead I ECG from the HEEDB 12-lead data and randomly incorporating one of the remaining six augmented leads into the model for training with a 50% probability. This ensures that the model can learn arrhythmia features from lead I data and axis deviation from the additional six leads. Furthermore, we have scaled down the model's parameter size to optimize for wearable devices with limited computational resources.

FINE-TUNING

When adapting to specific ECG downstream tasks, we needed to retain the parameters of the base model and discard the initial classification linear layer. The number of classes in the downstream task determines the number of neurons needed in the final layer of the new linear layer. The training objective is to generate classification outputs that match the labels. We adopted two different methods of fine-tuning: linear probing and full fine-tuning. During the linear probing experiments, we only fine-tuned the parameters of the linear classification head on top of the pretrained model, keeping all other pretrained model weights frozen. During full fine-tuning, we allowed all pretrained model weights to be updated and adapted to the downstream classification tasks.

The total training period is 30 epochs, with a learning-rate adjustment strategy that utilizes the scheduler. After every epoch, the scheduler monitors the specified metric, and if the performance does not improve for 10 consecutive epochs, the learning rate is reduced by a factor of 0.1. The learning-rate reduction is triggered based on the maximization of the monitored metric. This approach ensures a dynamic adjustment of the learning rate depending on the training progress. After training in each epoch, the model is evaluated on the validation set.[8] The model weights with the highest area under the receiver operating characteristic curve (AUROC) on the validation set are saved as a checkpoint for future evaluation.

CLINICAL VALIDATION

To validate and compare the performance of our model, we followed the committee experimental design of Hannun et al.[29] We established the committee consisting of three experienced ECG cardiologists to annotate a subset of the internal test set, which includes the 523 most recent ECGs from 523 unique patients. We developed an ECG

annotation system for cardiologists and 20 label types, including cardiac-rate abnormalities, conduction blocks, myocardial dilation, MI, and ECG morphologies, with sublabels under each category. Table 1 displays the complete list of label types. After independent annotation by the committee, a consensus determination was made; labels that did not reach consensus were removed, providing an expert standard for model evaluation. Labels that the committee could not interpret or agree on were eliminated from our test dataset.

In addition, to compare the diagnostic accuracy between the model and cardiologists, four additional ECG cardiologists were involved and provided specific instructions on how to use the system. Table S6 shows the cardiologists' ages, levels of experience, and education. Each cardiologist was required to annotate each ECG from the previous internal test set. The annotations from these cardiologists were then compared with the model's results.

The evaluation of the model was conducted by calculating accuracy, AUROC, sensitivity, and specificity, each with bilateral 95% confidence intervals. To compute these confidence intervals, we used different statistical methods tailored to the properties of each metric:

AUROC confidence interval: We employed the DeLong method, a nonparametric approach specifically designed to evaluate the variability of the AUROC. This method calculates the variance of the AUROC estimate based on the rankings of the predicted scores for the positive and negative classes, then derives the confidence interval using a normal approximation.

Sensitivity and specificity confidence interval: As sensitivity and specificity are proportions, we calculated their 95% confidence intervals using a binomial proportion method based on the Wilson score interval. This approach provides an accurate interval estimate even for small sample sizes.

Other metrics: For additional metrics (such as F1 score and others for which direct analytical confidence interval estimation methods are not available), we used the bootstrap method. This involves repeatedly resampling the data with replacement, calculating the metric for each resampled dataset, and then determining the percentile range that covers 95% of the computed metric values.

Results

Experimental results demonstrate that ECGFounder achieves superior performance on internal validation sets of 12-lead ECGs, with AUROC exceeding 0.95 for 80 diagnoses (Fig. S3). We further validated ECGFounder for 12-lead ECGs using the committee's internal test set. The algorithm's average AUROC score for diagnosing all 20 classifications was 0.968 (95% confidence interval [CI], 0.955 to 0.982), sensitivity was 0.971 (95% CI, 0.639 to 0.988), and specificity was 0.937 (95% CI, 0.912 to 0.953). When comparing the model with cardiologists, our model achieved an overall average F1 score of 0.677 (95% CI, 0.480 to 0.802), outperforming the cardiologists' average F1 score of 0.640. The model's performance was compared with cardiologists' performance across 20 diagnostic categories (Table 1). When comparing the model's receiver operating characteristic curve with the

true-positive rate and false-positive rate of cardiologists, the model outperformed the average performance of cardiologists for most labels (Fig. S2).

In the external test set experiments for 12-lead ECGs, we evaluated the performance of our model and other models on the CODE-test and PTB-XL datasets. On CODE-test, our model achieved an average AUROC of 0.981 (95% CI, 0.979 to 0.984), outperforming other baseline models such as S12L-ECG, which had an average AUROC of 0.980 (95% CI, 0.978 to 0.982); CTN, which had an average AUROC of 0.963 (95% CI, 0.960 to 0.967); and ECG Squeeze-and-Excitation Residual Neural Network (ECG-SE-ResNet), which had an average AUROC of 0.963 (95% CI, 0.961 to 0.967) (Table 2). On PTB-XL, as the other two models can only diagnose arrhythmias and cannot complete other diagnostic classifications, we validated this dataset only on the class that the models have. Our model achieved an average AUROC of 0.924 (95% CI, 0.917 to 0.931) (Table 3).[17,30,31] These results indicate that our model generalizes to different regions, hospitals, and patients.

Internal test set experiments focused on single-lead ECGs, with the model demonstrating excellent performance in rhythm-type diagnosis (Fig. S3). It achieved an AUROC above 0.95 for common heart-rate abnormalities such as normal sinus rhythm, sinus bradycardia, sinus tachycardia, marked sinus bradycardia, sinus arrhythmia, marked sinus arrhythmia, and AF. These diagnoses can be reliably identified from single-lead ECGs. In addition, the model achieved an AUROC above 0.8 for diagnosis, including premature ventricular complexes, premature supraventricular complexes, pacemaker, first-degree atrioventricular block, branch block, fascicular block, and chamber enlargement. The performance observed for some of these diagnoses was exceptional for the use of single-lead ECGs and is worthy of note. In MI diagnosis, the model also showed good diagnostic performance for lateral infarct, anterolateral infarct, acute MI, and ST-segment elevation MI. In ECG research, these diagnoses are associated with shifts in the heart's electrical axis. This suggests that data-augmentation methods based on the electrical axis are effective for training single-lead ECG models.

In the external test set experiments on single-lead ECG devices, our model accurately classified sinus rhythm and AF. The model's performance, as shown in Table S17, achieved an AUROC of 0.975 (95% CI, 0.972 to 0.977) and 0.957 (95% CI, 0.955 to 0.959) for normal sinus rhythm and AF, respectively. These results demonstrate excellent performance in analyzing ECG signals collected from portable and wearable ECG devices under real-world conditions. It should be emphasized that ECGFounder is executed on a cloud-based system, analyzing data uploaded from wearable devices, rather than operating directly on the wearable devices themselves.

We next validated the performance of ECGFounder in transfer learning. Our model was fine-tuned and evaluated on six downstream tasks using supervised learning on the MIMIC-IV-ECG dataset, resulting in both a 12-lead ECG model and a single-lead ECG model. The fine-tuning results are shown in Figure 2. As shown, ECGFounder outperforms the baseline methods in every downstream task. Specifically, it achieves 2 to 3 percentage points higher performance than ECG-SimCLR and 4 to 6 percentage points higher than ECG- in age, sex, NT-proBNP, LVEF, CKD, and CHD detection. In addition, for comparison with

the previously published CKD study,[32] we created an independent test set consisting of individual patients from the MIMIC-IV-ECG dataset for validation. The results are shown in Table 4. It should be noted that ECGFounder was the internal validation, while the compared method was the external validation.

Discussion

We have developed and demonstrated the generalizability and robust diagnostic capabilities of ECGFounder, a universal foundation model for ECG analysis, which consistently outperforms other ECG models and represents substantial clinical diagnostic value. Furthermore, we enhanced the model's performance on single-lead ECGs through novel data augmentation based on the cardiac axis. The internal validation results for arrhythmia diagnosis using single-lead ECGs demonstrated exceptional performance, broadening the prospects for the model's application in mobile health. Moreover, by leveraging a fine-tuned pretrained model, ECGFounder effectively adapts to a wide range of downstream tasks, significantly enhancing the diagnosis of other diseases such as CKD and CHD.

ECGFounder enhances diagnostic performance by learning to identify ECG features associated with cardiovascular diseases, which are typically diagnosed based on specific waveform patterns and rhythm characteristics, such as the elevated ST segments of MI and the irregular fluctuations of AF. These features involve anomalies in cardiac electrical activity, appearing significantly different from normal ECG waveforms. On training, ECGFounder can recognize these disease-related waveform patterns and rhythms, accurately diagnosing corresponding cardiovascular conditions. As observed in Table 1 and Figure S2, ECGFounder matches or even exceeds the performance of cardiologists on the internal review test sets. Furthermore, ECGFounder's sensitivity exceeds that of the average cardiology expert, indicating its ability to more accurately capture subtle signs of cardiovascular diseases that may be overlooked by human experts.

Previous research has demonstrated that deep-learning models can support clinical ECG analysis and achieve good performance, but lack a universal clinical diagnostic capability for ECGs.[6,29,33] First, the training datasets used by existing models are not large or diverse enough, which can lead to overfitting and poor performance on new data, limiting their generalizability.[34] Second, limited demographic diversity in training data can reduce model performance for some groups, leading to biased, less fair, and inaccurate outcomes. Third, the labels for most model training datasets are derived from time-consuming, labor-intensive, manual annotations by cardiologists. This limits the number of ECGs available for training. Moreover, as cardiologists typically use a unified annotation system, the richness of the dataset labels is not very high, often only including common ECG abnormalities and omitting many important but rare diagnostic labels.[35] Fourth, existing models do not support both 12-lead and single-lead ECGs, which this proposed foundation model does. As shown in Tables 2 and 3, we observed that ECGFounder ranks first in average performance across various external tests. The other baseline models, including S12L-ECG, CTN, and ECG-SE-ResNet, have previously achieved the best performance in AI-ECG models.[17,30,31] S12L-ECG was trained using supervised learning on the CODE-ECG dataset, which includes 2 million ECGs with six common types of arrhythmia diagnosis.[17] CTN and

ECG-SE-ResNet were trained on the PhysioNet Challenge 2020 dataset, which includes 60,000 ECGs covering 27 common ECG diagnoses.[30,31] We demonstrate that by training on a larger, more diverse ECG dataset, a scalable foundation model can further improve the diagnosis of cardiovascular diseases and surpass previous baseline models.

Despite the effectiveness of ECGFounder, challenges remain. First, as ECGFounder was developed using mostly U.S. cohorts, its data diversity is limited. Different parts of the world may exhibit unique ECG patterns — such as race-related and region-specific rhythm variations — requiring the model to handle diverse populations. Second, while ECG foundation models offer high diagnostic accuracy, their black-box nature can hinder trust in clinical use. Developing explainable AI is essential for enabling doctors to understand and adopt these models. Third, systematic bias is a limitation of our work, observed in regression tasks involving ECG-derived variables such as LVEF. Although the discrimination capability of our model remains strong, predictions for continuous heart-function metrics tend to show systematic biases, either overestimating or underestimating true values. Such biases may originate from intrinsic characteristics of the training dataset, annotation inconsistencies, or inherent challenges in modeling continuous physiological parameters from ECG signals. Addressing and mitigating this systematic bias in ECG regression tasks represents an important direction for future research.[36] Finally, some clinically relevant data (e.g., medical histories) that could serve as effective covariates in cardiovascular disease research are not yet included in the model. Future work should include more ECG data from diverse regions and ethnicities, incorporate patient demographics as model inputs, and develop explainable AI models to help doctors better understand model processes and outcomes. Recent natural language processing methods also show promise in processing cardiologist ECG annotations, enabling better use of embedded clinical knowledge.[37]

Conclusion

We have validated the efficacy of ECGFounder in adapting to a wide range of cardiovascular diagnostic applications, demonstrating its high performance and versatility across various downstream tasks as a foundational ECG model. By overcoming the limitations related to ECG data and labeling quality and diversity as well as training methods, our ECG foundation model confirms its potential to transform the standard of care in cardiology and to provide real-time, accurate cardiac assessments in diverse clinical settings. By providing open access to the model, code, and training data, we invite the research community to build on our model and help advance the state of the art.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

This document is the result of the research project funded by the National Science Foundation and Emory University via an unrestricted gift.

Retrospective analysis of data for this project was conducted with waiver of informed consent under approved institutional review board protocols (BIDMC: 2022P000417; MGH: 2013P001024).

Dr. Westover was supported by grants from the National Institutes of Health (NIH) (Grant Nos. RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119), and the NSF (2014431). Dr. Shenda Hong was supported by the National Natural Science Foundation of China (grant number 62102008), Clinical Medicine Plus X — Young Scholars Project of Peking University, the Fundamental Research Funds for the Central Universities (Grant Nos. PKU2024LCXQ030), CCF-Zhipu Large Model Innovation Fund (Grant Nos. CCF-Zhipu202414). Dr. Clifford was partially supported by the National Institute of Biomedical Imaging and Bioengineering under NIH award number R01EB030362.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the authors' current and past employers and funding bodies.

The data are publicly available at the Brain Data Science Platform (https://bdsp.io/content/heedb); the model weights at HuggingFace (https://huggingface.co/PKUDigitalHealth/ECGFounder); and the code at GitHub (https://github.com/PKUDigitalHealth/ECGFounder).

We thank Dr. Xinxin Di, Dr. Qinghao Zhao, Dr. Kun Lu, Dr. Zhengkai Xue, Dr. Wenbo Dai, Dr. Jing Zhao, and Dr. Hongqian Zhou for annotating the internal test set. We also thank Dr. Chenglong Li for statistical advice, and Ms. Shijia Geng for improving the manuscript.

References

1. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021;18:465–478. DOI: 10.1038/s41569-020-00503-2. [PubMed: 33526938]
2. Berkaya SK, Uysal AK, Gunal ES, Ergin S, Gunal S, Gulmezoglu MB. A survey on ECG analysis. Biomed Signal Process Control 2018;43:216–235. DOI: 10.1016/j.bspc.2018.03.003.
3. 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. DOI: 10.1038/s41591-018-0240-2. [PubMed: 30617318]
4. 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;394:861–867. DOI: 10.1016/S0140-6736(19)31721-0. [PubMed: 31378392]
5. Zhu H, Jiang Y, Cheng C, et al. Four-channel ECG as a single source for early diagnosis of cardiac hypertrophy and dilation — a deep learning approach. NEJM AI 2024;1(10). DOI: 10.1056/AIoa2300297.
6. Zhu H, Cheng C, Yin H, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. Lancet Digit Health 2020;2:e348–e357. DOI: 10.1016/S2589-7500(20)30107-2. [PubMed: 33328094]
7. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature 2023;622:156–163. DOI: 10.1038/s41586-023-06555-x. [PubMed: 37704728]
8. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nat Med 2024;30:850–862. DOI: 10.1038/s41591-024-02857-3. [PubMed: 38504018]
9. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. 2021 (10.48550/arXiv.2108.07258). Preprint.
10. Yu H, Guo P, Sano A. ECG semantic integrator (ESI): a foundation ECG model pretrained with LLM-enhanced cardiological text. 2024 (10.48550/arXiv.2405.19366). Preprint.
11. McKeen K, Oliva L, Masood S, Toma A, Rubin B, Wang B. ECG-FM: an open electrocardiogram foundation model. 2024 (10.48550/arXiv.2408.05178). Preprint.
12. Merdjanovska E, Rashkovska A. Comprehensive survey of computational ECG analysis: databases, methods and applications. Expert Syst Appl 2022;203:117206. DOI: 10.1016/j.eswa.2022.117206.

13. Clifford GD. Past, present and future challenges in sharing science: from PhysioNet to foundation models. Poster presented at 51st Computing in Cardiology, Karlsruhe, Germany, September 8–11, 2024 (https://cinc.org/archives/2024/pdf/CinC2024-039.pdf).
14. Reyna MA, Sadr N, Alday EAP, et al. Will two do? Varying dimensions in electrocardiography: the PhysioNet/computing in cardiology challenge 2021. In: Proceedings of 2021 Computing in Cardiology (CinC). Brno, Czechia: IEEE, 2021:1–4. (https://ieeexplore.ieee.org/document/9662687).
15. Koscova Z, Li Q, Robichaux C, et al. The Harvard-Emory ECG database. 2024 (10.1101/2024.09.27.24314503). Preprint.
16. GE Healthcare. MarquetteTM 12SLTM ECG analysis program: statement of validation and accuracy, revision B. 2007 (https://www.numed.co.uk/documents/download/216).
17. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020;11:1760. DOI: 10.1038/s41467-020-15432-4. [PubMed: 32273514]
18. Wagner P, Strodthoff N, Bousseljot R-D, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020;7:154. DOI: 10.1038/s41597-020-0495-6. [PubMed: 32451379]
19. Clifford GD, Liu C, Moody B, et al. AF classification from a short single lead ECG recording: the PhysioNet/computing in cardiology challenge 2017. In: Proceedings of 2017 Computing in Cardiology (CinC). Rennes, France. IEEE, 2017:1–4. (https://www.cinc.org/archives/2017/pdf/065-469.pdf).
20. Gow B, Pollard T, Nathanson LA, et al. MIMIC-IV-ECG: diagnostic electrocardiogram matched subset (version 1.0). PhysioNet 2023. DOI: 10.13026/4nqg-sb35.
21. Goldberger AL, Amaral LAN, Glass L, et al. Physiobank, PhysioToolkit, and PhysioNet. Circulation 2000;101:e215–e220. DOI: 10.1161/01.cir.101.23.e215. [PubMed: 10851218]
22. Torres-Soto J, Ashley EA. Multi-task deep learning for cardiac rhythm detection in wearable devices. NPJ Digit Med 2020;3:116. DOI: 10.1038/s41746-020-00320-4. [PubMed: 32964139]
23. Hong S, Xu Y, Khare A, et al. Holmes: health online model ensemble serving for deep learning models in intensive care units. In: Proceedings of 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA: Association for Computing Machinery, 2020:1614–1624. (https://arxiv.org/abs/2008.04063).
24. Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P. Designing network design spaces. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA: IEEE, 2020:10428–10436. (https://arxiv.org/abs/2003.13678).
25. Kashou AH, Noseworthy PA, Beckman TJ, et al. ECG interpretation proficiency of healthcare professionals. Curr Probl Cardiol 2023;48:101924. DOI: 10.1016/j.cpcardiol.2023.101924. [PubMed: 37394202]
26. Zhao Y, Xu Q, Jiang Y, Wen P, Huang Q. Dist-pu: positive-unlabeled learning from a label distribution perspective. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA: IEEE, 2022:14461–14470 (https://openaccess.thecvf.com/content/CVPR2022/papers/Zhao_Dist-PU_Positive-Unlabeled_Learning_From_a_Label_Distribution_Perspective_CVPR_2022_paper.pdf).
27. Kamga P, Mostafa R, Zafar S. The use of wearable ECG devices in the clinical setting: a review. Curr Emerg Hosp Med Rep 2022;10:67–72. DOI: 10.1007/s40138-022-00248-x. [PubMed: 35789964]
28. Meek S, Morris F. ABC of clinical electrocardiography: introduction. I-leads, rate, rhythm, and cardiac axis. BMJ 2002;324:415. DOI: 10.1136/bmj.324.7334.415. [PubMed: 11850377]
29. 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;25:65–69. DOI: 10.1038/s41591-018-0268-3. [PubMed: 30617320]
30. Natarajan A, Chang Y, Mariani S, et al. A wide and deep transformer neural network for 12-lead ECG classification. In: Proceedings of 2020 Computing in Cardiology. Rimini, Italy: IEEE, 2020:1–4. (https://ieeexplore.ieee.org/document/9344053).

31. Zhu Z, Wang H, Zhao T, et al. Classification of cardiac abnormalities from ECG signals using SE-ResNet. In: Proceedings of 2020 Computing in Cardiology. Rimini, Italy: IEEE, 2020:1–4. (https://ieeexplore.ieee.org/document/9344255).

32. Holmstrom L, Christensen M, Yuan N, et al. Deep learning-based electrocardiographic screening for chronic kidney disease. Commun Med 2023;3:73. DOI: 10.1038/s43856-023-00278-w. [PubMed: 37237055]

33. Hughes JW, Olgin JE, Avram R, et al. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. JAMA Cardiol 2021;6:1285. DOI: 10.1001/jamacardio.2021.2746. [PubMed: 34347007]

34. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. J Am Coll Cardiol 2017;70:1183–1192. DOI: 10.1016/j.jacc.2017.07.723. [PubMed: 28838369]

35. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE J Biomed Health Inform 2021;25:1519–1528. DOI: 10.1109/JBHI.2020.3022989. [PubMed: 32903191]

36. Nie G, Tang G, Hong S. Dist loss: enhancing regression in few-shot region through distribution distance constraint. Poster presented at 13th International Conference on Learning Representations (ICLR), Singapore, April 24–28, 2025 (https://openreview.net/pdf?id=YeSxbRrDRl).

37. Li J, Liu C, Cheng S, Arcucci R, Hong S. Frozen language model helps ECG zero-shot learning. In: Proceedings of 2024 Medical Imaging with Deep Learning. Paris, France: PMLR, 2024:402–415 (https://proceedings.mlr.press/v227/li24a/li24a.pdf).

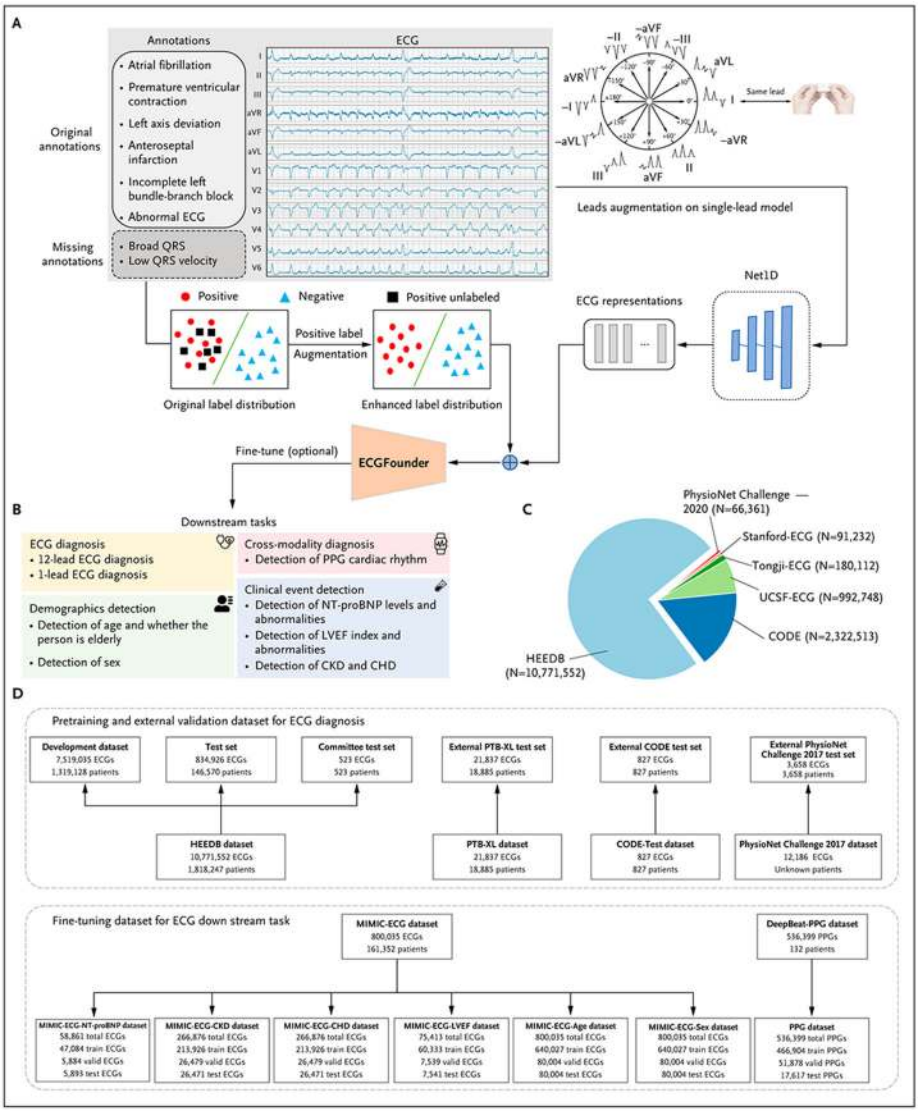

Figure 1. ECGFounder Is a General ECG Encoder Based on the RegNet Architecture.

Panel A shows how ECGFounder was trained on 10,771,552 ECGs with 150 types of ECG diagnostic labels. Owing to missing diagnoses in real-world expert annotations, we implemented positive label augmentation by modifying the loss function of the pretraining method. For more details, see the Methods section. Panel B shows how ECGFounder was applied to different clinical downstream tasks, covering ECG diagnosis, demographics detection, clinical event detection, and cross-modality diagnosis. In comparison with baseline methods, ECGFounder achieved state-of-the-art performance across all tasks. Panel C shows a comparison of the HEEDB dataset used by ECGFounder with other large ECG datasets. Panel D shows data used in model development and validation for the ECG diagnosis and downstream tasks. aVF denotes augmented voltage foot; aVL, augmented voltage left; aVR, augmented voltage right; CHD, chronic heart disease; CKD, chronic kidney disease; CODE, Clinical Outcomes in Digital Electrocardiology; ECG, electrocardiogram; HEEDB, Harvard–Emory ECG Database; LVEF, left ventricular

ejection fraction; MIMIC, Medical Information Mart for Intensive Care; NT-ProBNP, N-terminal pro–B-type natriuretic peptide; PPG, photoplethysmography; RegNet, Self-Regulated Network for Image Classification; and UCSF, University of California, San Francisco.

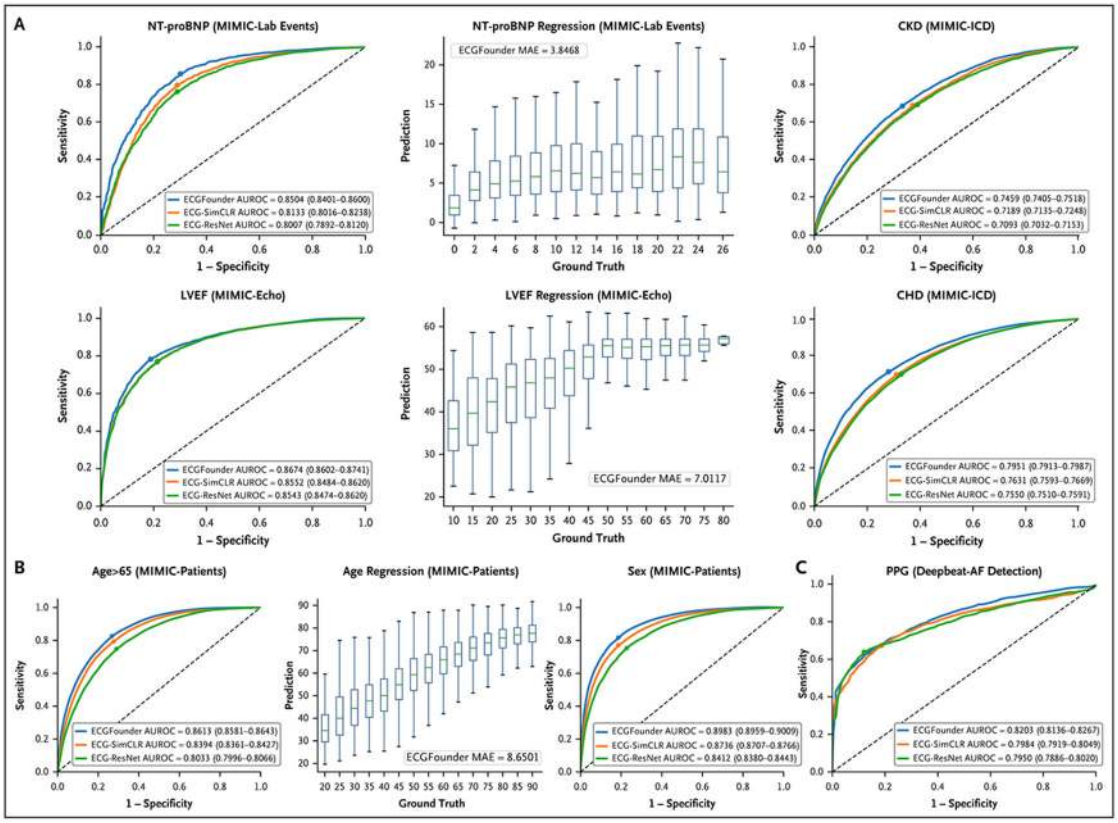

Figure 2. Results of Different Fine-Tuning Downstream Tasks.

Panel A shows the results of NT-proBNP classification, NT-proBNP regression, CKD classification, LVEF classification, LVEF regression, and CHD classification tasks of ECGFounder and other baseline models for clinical event detection tasks. Panel B shows the results of age classification over 65 years of age, age regression, and sex classification tasks of ECGFounder and other baseline models for demographic detection tasks. Panel C shows the results of the PPG atrial fibrillation detection tasks. Figures in parentheses denote 95% confidence intervals. AUROC denotes area under the receiver operating characteristic curve; CHD, chronic heart disease; CKD, chronic kidney disease; ECG, electrocardiogram; ICD, International Classification of Diseases; LVEF, left ventricular ejection fraction; MAE, mean absolute error; MIMIC, Medical Information Mart for Intensive Care; PPG, photoplethysmography; ResNet, Residual Neural Network; and SimCLR, Simple Framework for Contrastive Learning of Visual Representations.

Table 1. Performance of ECGFounder and Cardiologists on the Committee Test Set.*

		ECGFounder				Cardiologists		
Class	**Count**	**AUROC**	**Sens**	**Spec**	**F1**	**Sens**	**Spec**	**F1**
Anterior infarct	2	0.908 (0.868–0.941)	1.000 (0.207–1.000)	0.906 (0.877–0.928)	0.408 (0.000–0.503)	0.750	0.999	0.625
Atrial fibrillation	59	0.996 (0.992–0.998)	1.000 (0.916–1.000)	0.972 (0.952–0.983)	0.866 (0.795–0.921)	0.798	0.981	0.796
Atrial flutter	19	0.993 (0.984–0.999)	1.000 (0.741–1.000)	0.986 (0.971–0.993)	0.759 (0.622–0.898)	0.568	0.990	0.514
Atrial-paced rhythm	3	0.991 (0.986–0.994)	1.000 (0.439–1.000)	0.988 (0.974–0.994)	0.500 (0.000–0.739)	0.833	0.995	0.657
Incomplete right bundle-branch block	4	0.893 (0.853–0.951)	1.000 (0.510–1.000)	0.698 (0.656–0.736)	0.511 (0.295–0.867)	0.563	0.977	0.299
Inferior infarct	1	0.998 (0.950–0.999)	1.000 (0.207–1.000)	0.998 (0.989–1.000)	0.667 (0.000–1.000)	1.000	0.997	0.725
Lateral infarct	2	0.998 (0.970–0.999)	1.000 (0.207–1.000)	0.998 (0.989–1.000)	0.667 (0.000–0.925)	0.250	0.999	0.250
Left bundle-branch block	11	1.000 (1.000–1.000)	1.000 (0.566–1.000)	0.998 (0.989–1.000)	0.909 (0.697–1.000)	0.600	0.932	0.371
Normal sinus rhythm	364	0.969 (0.952–0.976)	0.934 (0.904–0.955)	0.926 (0.870–0.960)	0.952 (0.934–0.963)	0.916	0.862	0.930
Premature atrial complexes	19	0.997 (0.994–0.999)	1.000 (0.832–1.000)	0.963 (0.942–0.976)	0.679 (0.577–0.786)	0.579	0.989	0.613
Premature ventricular complexes	40	0.981 (0.962–0.989)	0.975 (0.871–0.996)	0.889 (0.857–0.915)	0.600 (0.522–0.647)	0.850	0.996	0.898
Right axis deviation	12	0.996 (0.993–1.000)	1.000 (0.676–1.000)	0.992 (0.979–0.997)	0.800 (0.593–0.927)	0.594	0.979	0.489
Right bundle-branch block	50	0.984 (0.971–0.988)	0.940 (0.838–0.979)	0.969 (0.948–0.981)	0.847 (0.779–0.941)	0.775	0.970	0.762
Sinus bradycardia	41	0.995 (0.990–0.998)	1.000 (0.908–1.000)	0.989 (0.975–0.995)	0.938 (0.895–0.958)	0.717	0.994	0.791
Sinus rhythm	191	0.970 (0.949–0.982)	0.970 (0.947–0.983)	0.926 (0.870–0.960)	0.971 (0.965–0.979)	0.916	0.862	0.930
Sinus tachycardia	58	0.996 (0.990–0.999)	1.000 (0.879–1.000)	0.945 (0.921–0.962)	0.683 (0.624–0.741)	0.821	0.990	0.833
Ventricular tachycardia	10	0.903 (0.883–0.932)	1.000 (0.342–1.000)	0.882 (0.850–0.907)	0.635 (0.076–0.820)	0.875	0.996	0.678
Ventricular-paced rhythm	20	0.988 (0.975–0.992)	0.950 (0.764–0.991)	0.940 (0.915–0.958)	0.559 (0.457–0.717)	0.725	0.997	0.756
With first-degree AV block	15	0.864 (0.802–0.913)	0.667 (0.417–0.848)	0.831 (0.795–0.862)	0.187 (0.070–0.320)	0.633	0.962	0.560
With sinus arrhythmia	6	0.976 (0.960–0.986)	1.000 (0.610–1.000)	0.945 (0.922–0.962)	0.308 (0.082–0.506)	0.542	0.979	0.319

*Figures in parentheses denote 95% confidence intervals. AUROC denotes area under the receiver operating characteristic curve; AV, atrioventricular; and ECG, electrocardiogram.

Table 2. Performance of ECGFounder and Other ECG Deep-Learning Models on External Test Set CODE-Test.*

Model		ECGFounder			S12L-ECG (Internal)[17]			CTN[30]			ECG-SE-ResNet[31]		
Class	**Count**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**
Sinus bradycardia	16	0.967 (0.961–0.973)	0.995 (0.992–0.998)	0.955 (0.954–0.956)	0.955 (0.953–0.957)	0.938 (0.936–0.939)	0.996 (0.994–0.998)	0.965 (0.964–0.966)	0.987 (0.985–0.989)	0.942 (0.941–0.943)	0.932 (0.931–0.934)	0.995 (0.994–0.996)	0.937 (0.934–0.940)
Atrial fibrillation	13	0.999 (0.998–0.999)	0.997 (0.994–1.000)	0.996 (0.994–0.998)	0.963 (0.962–0.966)	0.769 (0.754–0.784)	1.000 (0.998–1.000)	0.966 (0.959–0.973)	0.972 (0.968–0.976)	0.969 (0.967–0.971)	0.976 (0.974–0.978)	0.980 (0.978–0.982)	0.970 (0.968–0.972)
Sinus tachycardia	37	0.989 (0.986–0.992)	0.974 (0.972–0.976)	0.970 (0.964–0.976)	0.977 (0.974–0.980)	0.973 (0.968–0.978)	0.997 (0.994–0.999)	0.972 (0.969–0.975)	0.958 (0.956–0.960)	0.943 (0.942–0.945)	0.976 (0.974–0.978)	0.950 (0.978–0.953)	0.957 (0.956–0.958)
Right bundle-branch block	34	0.989 (0.986–0.992)	0.971 (0.967–0.975)	0.971 (0.965–0.977)	0.995 (0.994–0.997)	1.000 (0.998–1.000)	0.995 (0.992–0.997)	0.989 (0.985–0.993)	0.946 (0.943–0.949)	0.986 (0.985–0.987)	0.965 (0.961–0.978)	0.949 (0.946–0.952)	0.947 (0.945–0.949)
Left bundle-branch block	30	0.998 (0.997–0.999)	1.000 (0.999–1.000)	0.996 (0.994–0.998)	1.000 (0.999–1.000)	1.000 (0.998–1.000)	1.000 (0.998–1.000)	0.961 (0.959–0.964)	0.988 (0.985–0.990)	0.942 (0.941–0.944)	0.971 (0.969–0.973)	0.949 (0.947–0.951)	0.976 (0.975–0.977)
With first-degree AV block	28	0.949 (0.945–0.953)	0.864 (0.862–0.866)	0.957 (0.955–0.959)	0.989 (0.987–0.991)	0.929 (0.928–0.930)	0.995 (0.993–0.997)	0.930 (0.926–0.934)	0.862 (0.860–0.864)	0.925 (0.924–0.926)	0.958 (0.956–0.960)	0.856 (0.855–0.857)	0.922 (0.921–0.923)

*For the baseline method S12L-ECG, the CODE-test is the internal validation. For the ECGFounder method, the CODE-test is the external validation. Figures in parentheses denote 95% confidence intervals. AUROC denotes area under the receiver operating characteristic curve; AV, atrioventricular; CODE, Clinical Outcomes in Digital Electrocardiology; ECG, electrocardiogram; and SE-ResNet, Squeeze-and-Excitation Residual Neural Network.

Table 3. Performance of ECGFounder and Other ECG Deep-Learning Models on External Test Set PTB-XL.*

Model		ECG Founder			S12L-ECG[17]			CTN[30]			ECG-SE-ResNet[31]		
Class	**Count**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**
Anterior infarct	360	0.635 (0.618–0.644)	0.816 (0.781–0.845)	0.447 (0.445–0.451)									
Anterolateral infarct	297	0.945 (0.938–0.960)	0.837 (0.826–0.890)	0.918 (0.915–0.921)									
Anteroseptal infarct	2415	0.945 (0.942–0.947)	0.891 (0.879–0.900)	0.882 (0.878–0.884)									
Atrial fibrillation	1514	0.993 (0.992–0.994)	0.969 (0.968–0.975)	0.982 (0.980–0.984)	0.925 (0.916–0.934)	0.776 (0.765–0.778)	0.989 (0.985–0.993)	0.953 (0.945–0.961)	0.935 (0.927–0.942)	0.915 (0.913–0.917)	0.950 (0.944–0.956)	0.946 (0.942–0.949)	0.907 (0.905–0.909)
Atrial flutter	73	0.993 (0.988–0.997)	0.973 (0.954–0.997)	0.969 (0.966–0.970)				0.948 (0.946–0.95)	0.962 (0.956–0.968)	0.937 (0.936–0.937)	0.941 (0.936–0.946)	0.966 (0.964–0.67)	0.956 (0.955–0.957)
Electronic atrial pacemaker	298	0.909 (0.904–0.916)	0.935 (0.921–0.954)	0.745 (0.738–0.747)									
Inferior infarct	2726	0.851 (0.846–0.855)	0.825 (0.819–0.837)	0.732 (0.726–0.740)									
Lateral infarct	1066	0.916 (0.917–0.920)	0.819 (0.808–0.835)	0.859 (0.854–0.861)									
Left anterior fascicular block	1657	0.972 (0.970–0.974)	0.922 (0.918–0.932)	0.919 (0.916–0.920)				0.965 (0.964–0.967)	0.882 (0.876–0.887)	0.914 (0.912–0.917)	0.968 (0.966–0.97)	0.819 (0.816–0.821)	0.901 (0.899–0.903)
Left atrial enlargement	427	0.799 (0.788–0.814)	0.742 (0.696–0.757)	0.704 (0.700–0.709)									
Left bundle-branch block	539	0.989 (0.987–0.993)	0.959 (0.949–0.970)	0.965 (0.964–0.966)	0.983 (0.981–0.985)	0.963 (0.96–0.966)	0.991 (0.989–0.993)	0.912 (0.91–0.915)	0.932 (0.931–0.933)	0.933 (0.93–0.937)	0.931 (0.928–0.935)	0.917 (0.916–0.917)	0.930 (0.926–0.934)
Left posterior fascicular block	187	0.840 (0.82–0.862)	0.605 (0.531–0.674)	0.904 (0.902–0.908)									
Left ventricular hypertrophy	2419	0.900 (0.895–0.908)	0.860 (0.843–0.866)	0.788 (0.782–0.792)									
Low-voltage QRS	182	0.581 (0.572–0.587)	0.690 (0.682–0.697)	0.423 (0.418–0.430)				0.605 (0.592–0.618)	0.604 (0.598–0.610)	0.814 (0.798–0.830)	0.593 (0.591–0.595)	0.655 (0.653–0.657)	0.741 (0.732–0.749)
N-specific intraventricular block	789	0.766 (0.757–0.775)	0.339 (0.313–0.356)	0.932 (0.929–0.934)				0.788 (0.782–0.794)	0.487 (0.482–0.492)	0.906 (0.903–0.909)	0.776 (0.769–0.784)	0.699 (0.698–0.700)	0.793 (0.792–0.794)
Normal ECG	9857	0.887 (0.884–0.889)	0.952 (0.947–0.955)	0.639 (0.632–0.643)									
Premature ventricular complexes	1146	0.987 (0.985–0.989)	0.965 (0.960–0.974)	0.959 (0.957–0.962)				0.963 (0.961–0.965)	0.955 (0.949–0.961)	0.921 (0.920–0.922)	0.937 (0.931–0.943)	0.939 (0.936–0.942)	0.880 (0.879–0.881)
QT has lengthened	119	0.931 (0.922–0.941)	0.897 (0.850–0.909)	0.827 (0.824–0.831)				0.930 (0.929–0.931)	0.881 (0.879–0.883)	0.957 (0.949–0.964)	0.930 (0.929–0.932)	0.901 (0.899–0.913)	0.930 (0.929–0.932)

Model		ECG Founder			S12L-ECG[17]			CTN[30]			ECG-SE-ResNet[31]		
Class	**Count**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**	**AUROC**	**Sens**	**Spec**
Right atrial enlargement	99	0.959 (0.955–0.970)	0.869 (0.827–0.927)	0.915 (0.912–0.919)									
Right bundle-branch block	1155	0.976 (0.971–0.98)	0.925 (0.92–0.931)	0.925 (0.923–0.928)	0.967 (0.966–0.968)	0.911 (0.910–0.912)	0.976 (0.975–0.977)	0.932 (0.930–0.935)	0.922 (0.921–0.923)	0.931 (0.926–0.936)	0.946 (0.944–0.948)	0.842 (0.829–0.854)	0.912 (0.909–0.915)
Right ventricular hypertrophy	136	0.913 (0.890–0.938)	0.833 (0.807–0.891)	0.859 (0.857–0.861)									
Septal infarct	1423	0.947 (0.942–0.949)	0.894 (0.885–0.896)	0.880 (0.877–0.881)									
Sinus bradycardia	1159	0.950 (0.948–0.957)	0.972 (0.963–0.980)	0.855 (0.852–0.857)	0.967 (0.965–0.969)	0.938 (0.936–0.939)	0.989 (0.987–0.991)	0.996 (0.995–0.997)	0.966 (0.964–0.968)	0.946 (0.944–0.968)	0.983 (0.981–0.985)	0.935 (0.934–0.935)	0.963 (0.959–0.967)
Sinus rhythm	16,785	0.923 (0.917–0.926)	0.936 (0.932–0.94)	0.751 (0.742–0.762)				0.977 (0.975–0.979)	0.942 (0.939–0.944)	0.920 (0.919–0.921)	0.963 (0.961–0.965)	0.926 (0.924–0.928)	0.913 (0.912–0.914)
Sinus tachycardia	826	0.994 (0.993–0.995)	0.976 (0.967–0.988)	0.987 (0.985–0.988)	0.988 (0.986–0.990)	0.946 (0.945–0.946)	0.996 (0.995–0.997)	0.986 (0.984–0.988)	0.957 (0.956–0.958)	0.968 (0.966–0.970)	0.979 (0.977–0.981)	0.932 (0.931–0.933)	0.944 (0.941–0.947)
Supraventricular tachycardia	27	0.995 (0.992–0.998)	0.976 (0.926–1.000)	0.958 (0.957–0.961)									
Ventricular tachycardia	24	0.987 (0.979–0.995)	0.905 (0.818–0.952)	0.993 (0.992–0.993)									
With first-degree AV block	802	0.911 (0.903–0.916)	0.661 (0.631–0.679)	0.921 (0.919–0.925)	0.962 (0.959–0.964)	0.925 (0.923–0.927)	0.992 (0.989–0.995)	0.943 (0.939–0.947)	0.783 (0.779–0.786)	0.911 (0.910–0.912)	0.930 (0.929–0.931)	0.735 (0.731–0.738)	0.893 (0.890–0.896)
With QRS widening	45	0.618 (0.608–0.626)	0.414 (0.398–0.427)	0.742 (0.738–0.747)									
Wolff–Parkinson–White syndrome	18	0.919 (0.889–0.947)	0.747 (0.648–0.791)	0.982 (0.982–0.984)									

*For the ECGFounder method, PTB-XL is the external validation. Figures in parentheses denote 95% confidence intervals. AUROC denotes area under the receiver operating characteristic curve; AV, atrioventricular; ECG, electrocardiogram; and SE-ResNet, Squeeze-and-Excitation Residual Neural Network.

Table 4. Performance of the Other ECG Deep-Learning Model and ECGFounder on the Chronic Kidney Disease ECG Test Set.*

Models		ECGFounder		Holmstrom et al.	
Class	**Count**	**12-Lead AUROC**	**l-Lead AUROC**	**12-Lead AUROC**	**l-Lead AUROC**
Any-stage chronic kidney disease	13,990	0.746 (0.741–0.752)	0.707 (0.698–0.715)	0.639 (0.628–0.650)	0.622 (0.615–0.629)
Mild chronic kidney disease	514	0.590 (0.587–0.593)	0.590 (0.586–0.595)	0.576 (0.572–0.579)	0.549 (0.547–0.551)
Moderate to severe chronic kidney disease	7075	0.713 (0.711–0.716)	0.695 (0.692–0.698)	0.661 (0.657–0.665)	0.632 (0.630–0.634)
End-stage renal disease	6401	0.795 (0.789–0.801)	0.725 (0.721–0.729)	0.653 (0.648–0.659)	0.634 (0.632–0.636)

*For ECGFounder, this is internal validation; for Holmstrom et al.'s model,[32] this is external validation. Figures in parentheses denote 95% confidence intervals. AUROC denotes area under the receiver operating characteristic curve; and ECG, electrocardiogram.