JACC: CLINICAL ELECTROPHYSIOLOGY VOL. ■, NO. ■, 2025

PUBLISHED BY ELSEVIER

ORIGINAL RESEARCH

Expert-Level Automated Diagnosis of the Pediatric ECG Using a Deep Neural Network

Joshua Mayourian, MD, PhD, ME,[a] William G. La Cava, PhD,[b] Sarah D. de Ferranti, MD, MPH,[a] Douglas Mah, MD,[a] Mark Alexander, MD,[a] Edward Walsh, MD,[a] John K. Triedman, MD[a]

ABSTRACT

BACKGROUND Disparate access to expert pediatric cardiologist care and interpretation of electrocardiograms (ECGs) persists worldwide. Artificial intelligence–enhanced ECG (AI-ECG) has shown promise for automated diagnosis of ECGs in adults but has yet to be explored in the pediatric setting.

OBJECTIVES This study sought to determine whether an AI-ECG model can accurately perform automated diagnosis of pediatric ECGs.

METHODS This retrospective single-center cohort study included all patients with an ECG at Boston Children's Hospital read by an experienced pediatric cardiologist (≥5,000 reads) between 2000 and 2022. A convolutional neural network was trained (75% of patients) and internally tested (25% of patients) on ECGs to predict ECG diagnoses. The primary outcome was a composite of any ECG abnormality (ie, detecting normal vs abnormal ECG). Secondary outcomes include Wolff-Parkinson-White syndrome (WPW) and prolonged QTc. Model performance was assessed with area under the receiver-operating (AUROC) and precision recall (AUPRC) curves.

RESULTS The main cohort consisted of 201,620 patients (49% male; 11% with known congenital heart disease) and 583,134 ECGs (median age 11.7 years [Q1-Q3: 3.1-16.9 years]; 56% any ECG abnormality, 1.0% WPW, and 5.3% with prolonged QTc). The AI-ECG model outperformed the commercial software interpretations for detecting any abnormality (AUROC 0.94; AUPRC 0.96), WPW (AUROC 0.99; AUPRC 0.88), and prolonged QTc (AUROC 0.96; AUPRC 0.63). During readjudication of ECGs with AI-ECG/original cardiologist read discordance, blinded expert readers were more likely to agree with AI-ECG classification than the original reader to detect any abnormality ($P = 0.001$), WPW ($P = 0.01$), and prolonged QTc ($P = 0.07$).

CONCLUSIONS Our model provides expert-level automated diagnosis of the pediatric 12-lead ECG, which may improve access to care. (JACC Clin Electrophysiol. 2025;■:■-■) © 2025 by the American College of Cardiology Foundation.

From the [a]Department of Cardiology, Boston Children's Hospital, Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA; and the [b]Computational Health Informatics Program, Boston Children's Hospital, Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received January 3, 2025; revised manuscript received January 31, 2025, accepted February 5, 2025.

ISSN 2405-500X/$36.00 https://doi.org/10.1016/j.jacep.2025.02.003

ABBREVIATIONS AND ACRONYMS

AI-ECG = artificial intelligence-enhanced electrocardiogram

AUROC = area under the receiver-operating curve

AUPRC = area under the precision-recall curve

ECG = electrocardiogram

PPV = positive predictive value

WPW = Wolff-Parkinson-White syndrome

The electrocardiogram (ECG) is a ubiquitous and inexpensive diagnostic test used in pediatric clinical practice worldwide for a wide range of indications: known acquired/congenital heart disease,[1] suspicion for arrhythmia or structural heart disease,[2] sports clearance,[3] medication initiation/monitoring,[4] and others. The growing volume of ECGs, along with considerations for universal ECG screening,[5] underscores the need for rapid and reliable ECG interpretations.

Specialized expertise is required for pediatric ECG interpretation to account for the progressive anatomic/physiological changes from the newborn stage to adulthood, which corresponds to significantly different ECG patterns and epidemiology. Although many ECG vendors have implemented rule-based diagnostic algorithms, these are generally focused on adult literature with limited standalone value in the pediatric ECG setting. It would be of considerable value, especially in low-resource settings with limited pediatric cardiology expertise, to have access to a novel automated ECG diagnostic tool designed for pediatric use.

Deep learning-based approaches have proven successful for rapid and reliable automated interpretation of ECGs in the adult population,[6] making it plausible that an artificial intelligence-enhanced ECG (AI-ECG) model may similarly aid in ECG interpretation for the pediatric population. However, given the unique considerations of pediatric ECGs, AI-ECG algorithms for adults are expected to have poor generalizability to pediatric cohorts. There are few available AI-ECG applications to pediatric and congenital cardiology,[7,8] due in part to the paucity of large, well-validated pediatric ECG databases, which hinders similar research. Thus, there are currently, to our knowledge, no AI-ECG algorithm for pediatric ECG diagnosis.

The current article addresses this gap by training and testing a convolutional neural network using >500,000 ECGs on >200,000 patients to reliably identify common and rare ECG findings in the pediatric population. Model performance was benchmarked to commercial software, and discordantly classified tracings were readjudicated to 4 expert pediatric electrophysiologists. Subgroup analysis defined model performance within a range of ages. Finally, model explainability analysis was performed.

METHODS

STUDY POPULATION AND PATIENT ASSIGNMENT. Patient data from Boston Children's Hospital were used. Inclusion criteria consisted of any patient with at least one ECG between 2000 and 2022 without missing metadata (eg, missing medical record number, ECG event number, reading provider, ECG diagnosis). Pediatric and adult patients with congenital heart disease were included, given that: 1) adult patients with congenital heart disease have unique ECG patterns/findings that are related to their underlying congenital heart lesion and their sequela of cardiac interventions; 2) pediatric cardiologists are frequently caring for these patients and interpreting their ECGs; and 3) there is a global shortage of the adult congenital heart disease workforce.

To optimize training label accuracy, we removed ECGs from less experienced readers (ie, providers with ≤5,000 ECGs). ECGs failing to pass quality control (quality control details are discussed in the "Quality Control and Data Preprocessing" section) were also removed. Finally, ECGs with equivocal (eg, "probably normal ECG variant") or outdated (eg, "counterclockwise rotation") diagnoses were removed. The remaining ECGs comprised the main cohort, which was then partitioned at the patient level into training (75%) and testing (25%) cohorts (Supplemental Figure 1).

DATA RETRIEVAL. ECG waveforms (I, II, and V1-V6) were obtained from the MUSE ECG data management system (GE Healthcare), with each lead corresponding to a one-dimensional vector sampled at 250 Hz for a 10 second duration (2,500 samples). Leads III, aVF, aVL, and aVR were reconstructed using Einthoven's law[9] and the Goldberger equation.[10]

ECG diagnoses, age, sex, and congenital heart lesion diagnoses were identified based on the institutional Fyler coding system.[11] During the study period, all ECGs were read using a custom internal software (ECG Reader) that requires physicians to select one or more predetermined ECG diagnostic codes, providing a cleanly labeled data set for training and testing purposes (Supplemental Figure 2).

QUALITY CONTROL AND DATA PREPROCESSING. Our quality control and preprocessing pipeline has been previously published.[7] Briefly, ECGs without 2,500 samples or missing lead information were removed. For each passing ECG, a high-pass filter and trimming

process was then applied to 2,048 samples (approximately 8 seconds) for ease of working with convolutional neural networks.

DEFINITION OF PRIMARY AND SECONDARY OUTCOMES. At a high level, we envision the following clinical workflow for an AI-ECG automated diagnosis algorithm: Is the ECG predicted by the AI-ECG model normal or abnormal? If normal, this may conceivably spare an expert from reviewing the ECG. If abnormal, the AI-ECG algorithm can provide a comprehensive list of pertinent ECG diagnoses.

Guided by this framework, the primary composite outcome was ECG diagnosis of any abnormality (defined as any diagnostic code other than "Normal for age" and "Sinus arrhythmia") by the original cardiologist reader. Secondary outcomes included individual ECG diagnoses that are especially pertinent to pediatric ECG screening, namely Wolff-Parkinson-White syndrome (WPW) and prolonged QTc. A comprehensive list of all ECG diagnoses that were predicted by the AI-ECG model herein is shown in Supplemental Table 1.

As secondary analyses, we also evaluated time-to-diagnosis of ECG abnormality, WPW, and prolonged QTc (discussed in the "Time-to-Event Analysis" section).

MODEL SELECTION, ARCHITECTURE, AND TRAINING. Our model development and architecture mimics our previous work.[7] As previously described, the convolutional neural network involves a residual block architecture that has been adapted for unidimensional signals (diagram shown elsewhere).[7]

Briefly, the model was developed exclusively on the training set, of which 5% was designated for validation and hyperparameter tuning. The input to the convolutional neural network was $12 \times 2{,}048$ ECG samples. The final hyperparameters were obtained via a grid search: kernel size (3, 9, and 17), batch size (8, 32, and 64), and initial learning rate (0.01, 0.001, and 0.0001). The average cross-entropy was minimized by using the Adam optimizer. Maximum 150 epochs were used with early stopping based on validation loss. Final hyperparameters for this model were kernel size of 17, batch size of 32, and learning rate of 0.001.

Given the known role of age and sex on ECG characteristics, we created a second model that incorporated age/sex as inputs along with ECG waveforms (AI-ECG+age+sex). The architecture is shown elsewhere.[7] For the AI-ECG+age+sex model, final hyperparameters were kernel size of 9, batch size of 32, and learning rate of 0.001.

PERFORMANCE EVALUATION AND STATISTICAL ANALYSES. Model performance was evaluated exclusively on the testing cohort. Area under the receiver-operating curve (AUROC) and area under the precision-recall (ie, positive predictive value [PPV]-sensitivity) curve (AUPRC) were computed. Other performance metrics assessed include PPV, negative predictive value, sensitivity, specificity, F1 score, and accuracy. Given the imbalanced data set and our objective to emulate human behavior when interpreting ECGs (ie, balancing precision and recall), these metrics were calculated based on thresholds achieving the optimal F1 score. A similar cutoff strategy was implemented by prior AI-ECG works for automated ECG diagnosis in adults.[6,12] The F1 score is defined as the harmonic mean of the precision and sensitivity, symmetrically representing both within one metric. 95% CIs were obtained via resampling with 1,000 bootstraps.

BENCHMARKING MODEL PERFORMANCE. Model performance was benchmarked to commercial (MUSE, GE Healthcare) ECG interpretations.

TIME-TO-EVENT ANALYSIS. Time-to-event analysis was performed for the following outcomes: diagnosis of any abnormality, WPW, and prolonged QTc. For each outcome, time-to-diagnosis onset analysis was performed by: 1) including only patients with multiple ECGs, whereby the first ECG is negative for the diagnosis of interest; 2) stratifying patients into 2 groups based on the AI-ECG classification of the first ECGs (true negative or false positive); and 3) assessing time-to-diagnosis after ECG within each group.

Cox proportional hazards regression was used to evaluate AI-ECG classification association with time from ECG until the diagnosis of interest. HRs were adjusted for age and sex. Statistical comparison between groups was based on log-rank testing. Patients who never had the ECG diagnosis of interest were censored at the time of last ECG.

READJUDICATION AND EXPERT AGREEMENT. Readjudication was performed on ECGs with discrepancies between the diagnostic classification assigned by the original reader and the AI-ECG classification (using the same cutoff as noted earlier). More specifically, for each outcome (any abnormality, WPW, and prolonged QTc), 50 false-positive findings (ie, deemed positive by AI-ECG but negative by the original reader) and 50 false-negative findings (ie, deemed negative by AI-ECG but positive by the original reader) were re-read by 4 senior pediatric electrophysiologists. These experts were blinded to the

TABLE 1 Baseline Characteristics of Training and Testing Cohorts

	Training Cohort	Testing Cohort
Patients	n = 151,215	n = 50,405
Male	74,015 (49)	24,742 (49)
Known CHD diagnosis	17,193 (11)	5,605 (11)
Cardiomyopathy	2,597 (1.7)	937 (1.9)
ASD	5,052 (3.3)	1,592 (3.2)
CAVC	743 (0.5)	225 (0.4)
CoA	3,076 (2.0)	1,021 (2.0)
DORV	1,054 (0.7)	359 (0.7)
D-loop TGA	1,677 (1.1)	584 (1.2)
Ebstein	467 (0.3)	148 (0.3)
HLHS	1,061 (0.7)	321 (0.6)
L-loop TGA	642 (0.4)	232 (0.5)
Pulmonary atresia	1,094 (0.7)	373 (0.7)
TAPVC	630 (0.4)	209 (0.4)
Tricuspid atresia	443 (0.3)	134 (0.3)
Truncus arteriosus	281 (0.2)	100 (0.2)
VSD	7,671 (5.1)	2,544 (5.0)
Dextrocardia	619 (0.4)	202 (0.4)
ToF	2,486 (1.6)	796 (1.6)
ECGs	n = 437,350	n = 145,784
Age at ECG, y	11.6 (3.1-16.9)	12.0 (3.3-17.0)
No. of diagnoses		
1	292,517 (67)	97,654 (67)
2	86,184 (20)	28,735 (20)
3	36,957 (8.5)	12,122 (8.3)
4	15,333 (3.5)	5,121 (3.5)
5	5,002 (1.1)	1,719 (1.2)
>5	1,357 (0.3)	433 (0.3)
Composite outcome		
Any abnormality	244,522 (56)	81,127 (56)
Individual diagnoses		
RSR'	50,218 (11)	17,115 (12)
NSSTT	46,084 (11)	15,728 (11)
Atrial fibrillation	1,680 (0.4)	680 (0.5)
Atrial flutter	1,618 (0.4)	496 (0.3)
High-grade AV block	1,423 (0.3)	398 (0.3)
Pericardial ST/T-wave changes	1,344 (0.3)	420 (0.3)
Ischemic ST/T-wave changes	1,341 (0.3)	486 (0.3)
Prolonged QTc	23,242 (5.3)	7,679 (5.3)
Sinus bradycardia	14,180 (3.2)	5,137 (3.5)
Sinus tachycardia	21,618 (4.9)	7,363 (5.1)
SVT	1,405 (0.3)	510 (0.3)
WPW	4,441 (1.0)	1,493 (1.0)
Technically inadequate study	4,021 (0.9)	1,415 (1.0)
RVH	29,857 (6.8)	9,092 (6.2)
LVH	10,664 (2.4)	3,303 (2.3)
T-wave inversions	11,051 (2.5)	3,795 (2.6)
CRBBB	34,134 (7.8)	10,773 (7.4)

Values are no. (%) or median (Q1-Q3).

ASD = atrial septal defect; AV = atrioventricular; CAVC = complete atrioventricular canal defect; CHD = congenital heart disease; CoA = coarctation of the aorta; CRBBB = complete right bundle branch block; DORV = double-outlet right ventricle; ECG = electrocardiogram; HLHS = hypoplastic left heart syndrome; LVH = left ventricular hypertrophy; NSSTT = nonspecific ST/T-wave changes; RVH = right ventricular hypertrophy; SVT = supraventricular tachycardia; TAPVC = total anomalous pulmonary venous connection; TGA = transposition of the great arteries; ToF = tetralogy of Fallot; VSD = ventricular septal defect; WPW = Wolff-Parkinson-White syndrome.

diagnoses by AI-ECG, the clinical indication, and the original reader. They were presented with the patient age, sex, and automated measurements of axes/intervals.

Agreement was assessed between the original reader and experts, AI-ECG and experts, and all 4 experts using the Cohen kappa (for agreement between 2 entities) or the Fleiss kappa (for agreement between >2 entities).[13] Values <0 indicate no agreement, with 0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1 as near perfect agreement.

MODEL EXPLAINABILITY. Model behavior was investigated via median waveform analysis and saliency mapping.[7] Briefly, median waveforms provide visual representations of high- and low-risk ECGs. The 100 highest predicted ECGs for WPW and prolonged QTc were used to create high-risk median waveforms. To contrast with normal ECGs, the 100 lowest predicted ECGs for any abnormality were used to create low-risk median waveforms.

Saliency mapping provides insight into important ECG patterns that contribute to model prediction. Using a Shapley Additive Explanations framework,[14] saliency maps highlight ECG regions where a change in ECG voltage input corresponds to a change in output prediction. The 100 ECGs with highest predicted probability for each diagnosis were used to create saliency maps. Details of median waveform analysis and saliency mapping are provided elsewhere.[7]

RESULTS

PATIENT POPULATION BASELINE CHARACTERISTICS. There were 734,800 ECGs (238,072 patients) between 2000 and 2022; after removing ECGs with missing metadata (18,539 ECGs), less experienced readers (77,363 ECGs), failed quality control (10,683 ECGs), and equivocal diagnoses (45,081 ECGs), 583,134 ECGs (201,620 patients) comprised the main cohort (Supplemental Figure 1). Within the main cohort, 11% of patients had known congenital heart disease.

The training and testing cohorts comprised 437,350 ECGs (151,215 patients; 49% male; median age 11.6 years [Q1-Q3 3.1-16.9 years]) and 145,784 ECGs (50,405 patients; 49% male; median age 12.0 years [3.3-17.0 years]), respectively. In both cohorts, 33% had more than one diagnosis; 56% had any abnormality, 1.0% had WPW, and 5.3% had prolonged QTc. The prevalence of congenital heart disease lesions and diagnoses of other ECG findings are highlighted in Table 1.

MODEL PERFORMANCE. Performance of the AI-ECG and AI-ECG+age+sex models to detect any abnormality, WPW, and prolonged QTc is shown in **Figure 1A**. Excellent performance was achieved for the AI-ECG model to detect any abnormality (AUROC 0.94; AUPRC 0.96), WPW (AUROC 0.99; AUPRC 0.88), and prolonged QTc (AUROC 0.96; AUPRC 0.63). Model performance was nearly identical when adding age and sex as inputs (**Figure 1**, Supplemental Figure 3), outperforming the commercial MUSE interpretation benchmark for any abnormality with more pronounced differences for WPW and prolonged QTc (**Figure 1A**). Sensitivity, specificity, negative predictive value, PPV, accuracy, and F1 scores for individual ECG diagnosis are presented in Supplemental Table 1.

Patients with an initial false-positive AI-ECG classification were more likely to have a future abnormal ECG (HR: 2.0; 95% CI: 1.8-2.2; $P < 0.001$), WPW (HR: 88; 95% CI: 35-218; $P < 0.001$), and prolonged QTc (HR: 3.4; 95% CI: 2.9-4.0; $P < 0.001$) (**Figure 1B**).

SUBGROUP ANALYSIS. In a subgroup analysis (**Figure 2**), there was variation in performance according to age, sex, and reading provider. In general, there was lower performance for ages <3 year, most notably for age <1 week for any abnormality, and prolonged QTc. Interestingly, performance increased from <1 week to 1 month and then 6 months, after which it plateaued. Performance for ages ≥18 years were similar to the overall cohort. Performance did not vary according to sex. Finally, there was slightly higher performance for a pediatric electrophysiology specialist, most notably for prolonged QTc. There was no appreciable difference in performance when testing on a provider's first 5,000 ECG reads (any abnormality: AUROC 0.93, AUPRC 0.96; WPW: AUROC 0.98, AUPRC 0.88; and prolonged QTc: AUROC 0.96, AUPRC 0.61) to their last 5,000 ECG reads (any abnormality: AUROC 0.94, AUPRC 0.96; WPW: AUROC 0.99, AUPRC 0.87; and prolonged QTc: AUROC 0.96, AUPRC 0.62).

READJUDICATION. Readjudication was performed by 4 blinded expert electrophysiologists on ECGs with discrepancies between the original reader and the AI-ECG classification. As shown in the **Figure 3A** heatmap, AI-ECG, but not the original reader, clustered with the experts. Interestingly, the inter-expert Fleiss kappa agreement was modest across each outcome of interest, ranging from 0.21 to 0.49 (all P values not significant). The expert readers on average were more likely to agree with the AI-ECG model than with the original reader (**Figure 3B**).

There was a 2-2 tie among experts for 20% of the any abnormality readjudication ECGs, 17% of the WPW readjudication ECGs, and 19% of the prolonged QTc readjudication ECGs. Readjudication results were similar when excluding the ECGs without majority vote (Supplemental Figure 4).

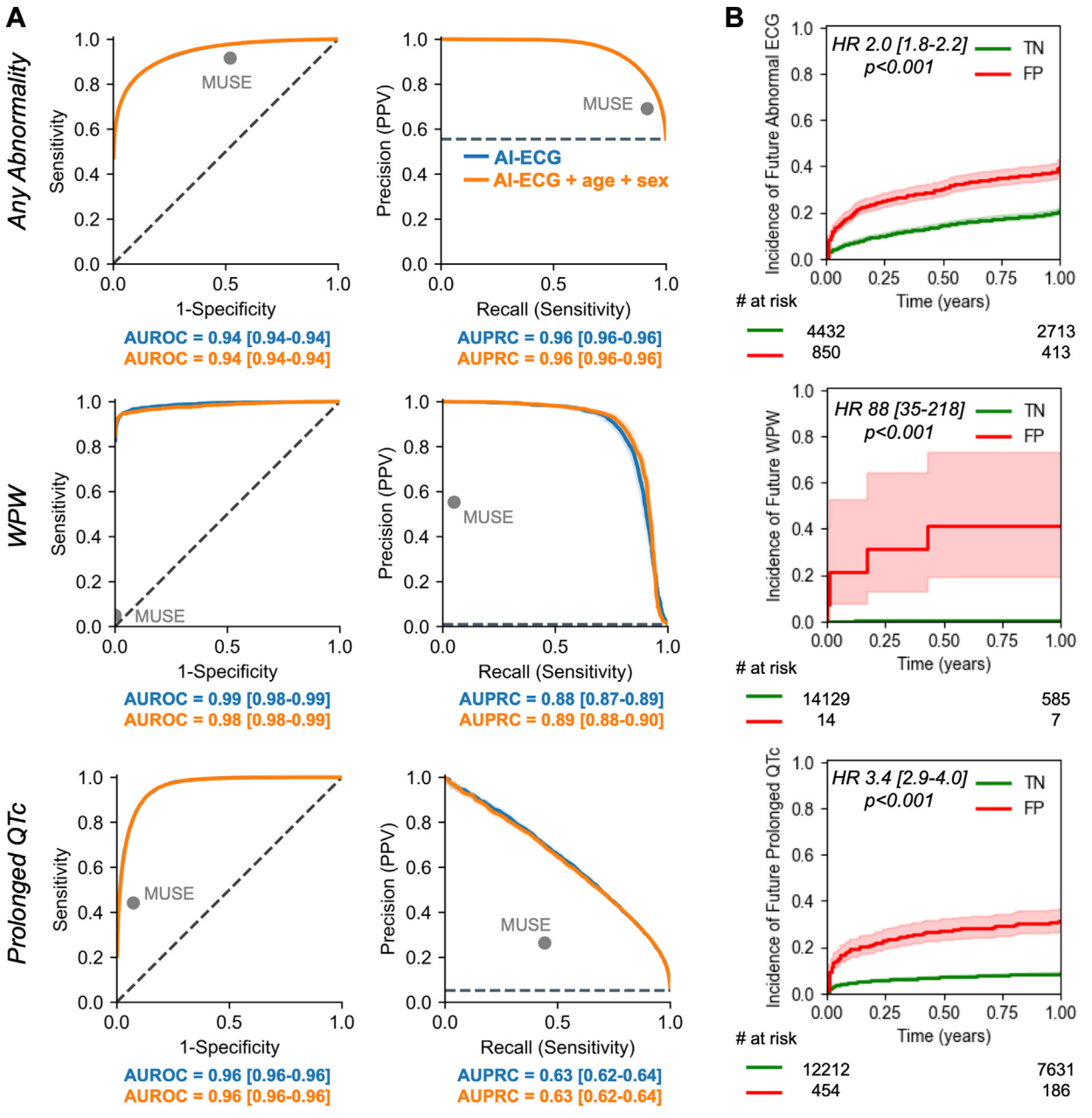
MODEL EXPLAINABILITY. Model behavior analysis was performed to compare salient features noted by the AI-ECG tool compared with conventional rule-based approaches for diagnostics (**Figure 4**). For WPW, the most salient signals were within P waves, QRS complexes, and PR signals (limb leads I-II and precordial leads V1, V5-V6). High-risk waveforms unsurprisingly exhibited preexcitation (a "Delta wave") with short PR intervals. For prolonged QTc, the most salient features were QRS complexes and T waves (precordial leads V1 and V6). High-risk waveforms predictably exhibited a prolonged QTc interval.

DISCUSSION

Given the proliferation of ECGs for screening in children, the burgeoning population of young patients with congenital heart disease, and the unique considerations of the pediatric ECG, it is of great interest to create a pediatric-specific ECG diagnostic model. This work represents the first (to our knowledge) application of ECG-based deep learning to automatically diagnose ECG findings in the pediatric and adult congenital heart disease population (**Central Illustration**). After training on >400,000 ECGs from nearly 150,000 patients with and without cardiac abnormalities, we found that model performance was excellent (AUROC >0.9) for the main outcomes of interest, outperforming the commercially available MUSE software. Model performance remained robust across a range of subgroups, and readjudication of misclassified ECGs showed that 4 blinded senior electrophysiologists were more likely to agree with AI-ECG diagnoses than with an experienced reader in these boundary cases. Finally, saliency mapping findings align with conventional rule-based methodologies implemented by humans, promoting clinician trust. Altogether, these findings show the promise of the AI-ECG model to assist clinicians with the rapid and reliable interpretation of ECGs, which may: 1) reduce missed diagnoses by less experienced clinicians; 2) promote screening programs; 3) facilitate improved access to expert care; and 4) decrease physician workload, which may help reduce burnout.[15]

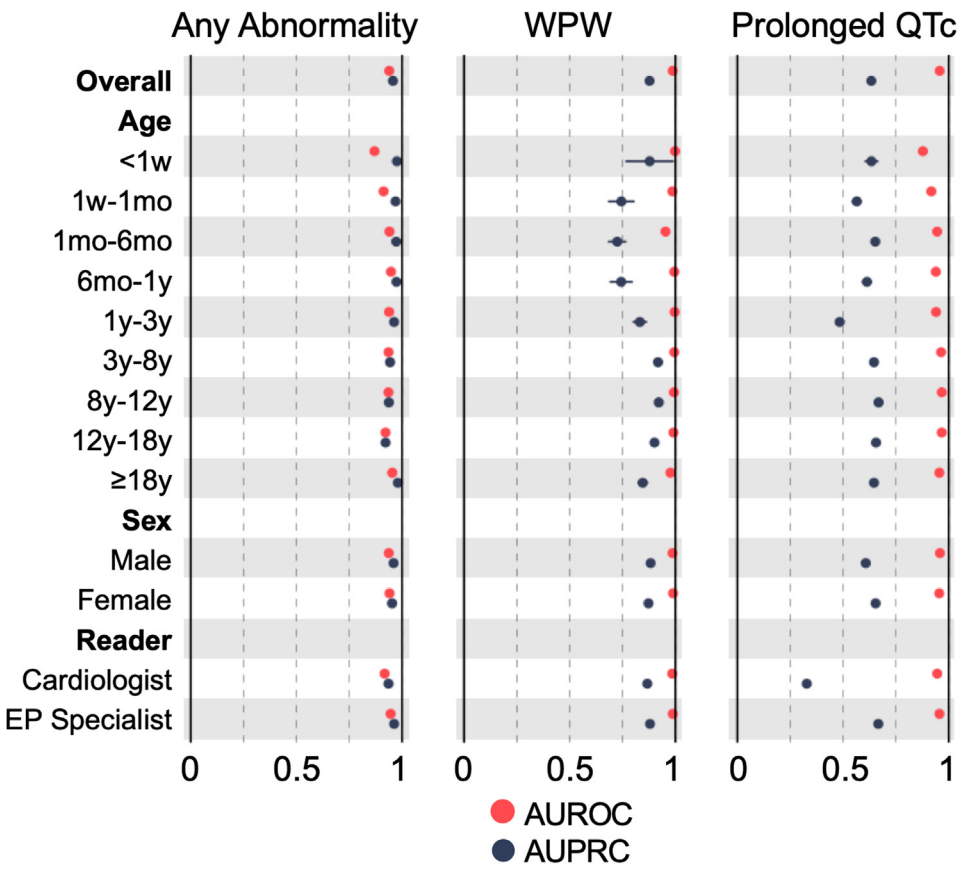
CURRENT STATE OF ECG INTERPRETATION AND AUTOMATED DIAGNOSIS. Deep learning approaches have been leveraged to enhance ECG diagnoses broadly in the general adult population. For example,

FIGURE 1 **AI-ECG Model Performance to Detect ECG Abnormalities**



(A) Performance of the artificial intelligence–enhanced electrocardiogram model alone (AI-ECG; blue) and with demographic data (AI-ECG+age+sex; orange) evaluated in the testing cohorts using receiver-operating (AUROC; left) and precision-recall (AUPRC; right) curves for the following outcomes: any abnormality, Wolff-Parkinson-White syndrome (WPW), and prolonged QTc. Model benchmarked to commercial MUSE interpretations (gray dot). AUROC and AUPRC metric values for each model and outcome are inset below with 95% CIs in brackets. Dotted line represents chance. (B) Incidence of future abnormal ECG (top), WPW (middle), or prolonged QTc (bottom) for the test cohort stratified according to initial network classification (true negative [TN] in green, false positive [FP] in red). Number of patients at risk over the 1-year period inset below. HRs and log-rank *P* values are inset.

FIGURE 2 Model Performance Across Age and Sex Subgroups



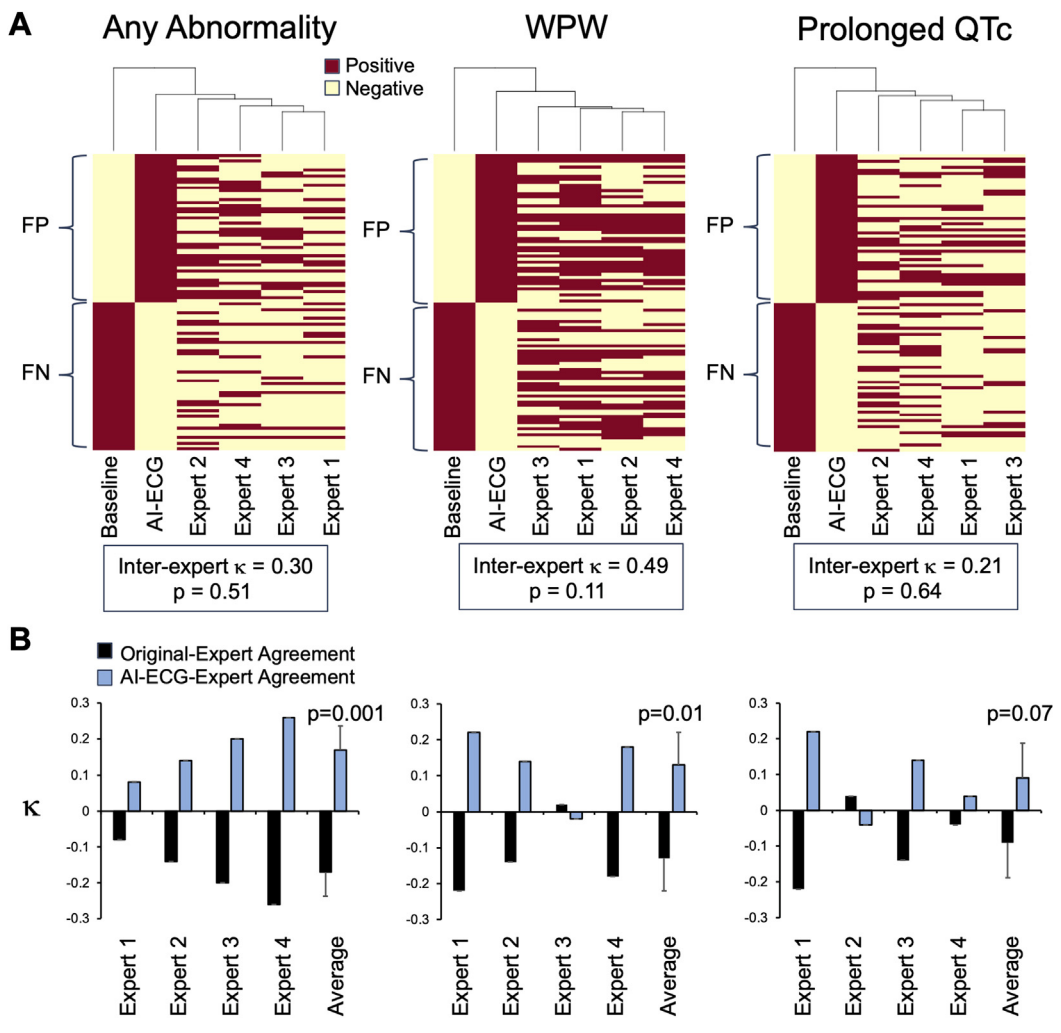
Forest plots showing AUROC (red) and AUPRC (black) curve performance when stratifying according to age, sex, and reader for diagnosing any abnormality (left), WPW (middle), and prolonged QTc (right). Abbreviations in Figure 1.

Ribeiro et al[6] reported that a deep neural network can outperform cardiology trainees in recognizing common ECG abnormalities in adults. Recently, AI-ECG has been used to detect myocardial infarction,[16] with a pragmatic randomized controlled trial showing that an AI-ECG-assisted triage of ST-segment elevation myocardial infarction decreased the door-to-balloon time for patients presenting to the emergency department.[17]

Utilization of adult AI-ECG models for ECG interpretation are unlikely to be well suited for pediatric ECG diagnoses for several reasons. First, the rapid and nonlinear anatomic and physiological changes from newborn to adulthood make adult AI-ECG models not amenable to pediatric generalization. These changes include: 1) a right-axis deviation in newborns that can be normal and reflect in utero physiology; 2) normal heart rate ranges that evolve through infancy and through adolescence; 3) T waves that are upright in the first weeks of life, then invert, and then gradually become upright during adolescence; and 4) axes and intervals that vary according to age and sex.[18] Second, there is a disproportionate prevalence of patients with congenital heart lesions with highly prevalent and specific ECG abnormalities that are rare in adult populations (eg, superior axis deviation). Finally, the clinical issues commonly asked of the adult ECG (eg, the presence of ischemia/infarction) are rarely encountered in pediatrics, with pediatric use cases gravitating toward screening applications for evaluation of nonspecific symptoms and identification of rare ECG findings such as WPW and long QT syndrome.

BENCHMARKING TO COMMERCIAL SOFTWARE. To our knowledge, there are no published studies that validate the accuracy of MUSE interpretation for pediatric patients. This prompted us to include the

FIGURE 3 Model Classification Readjudication



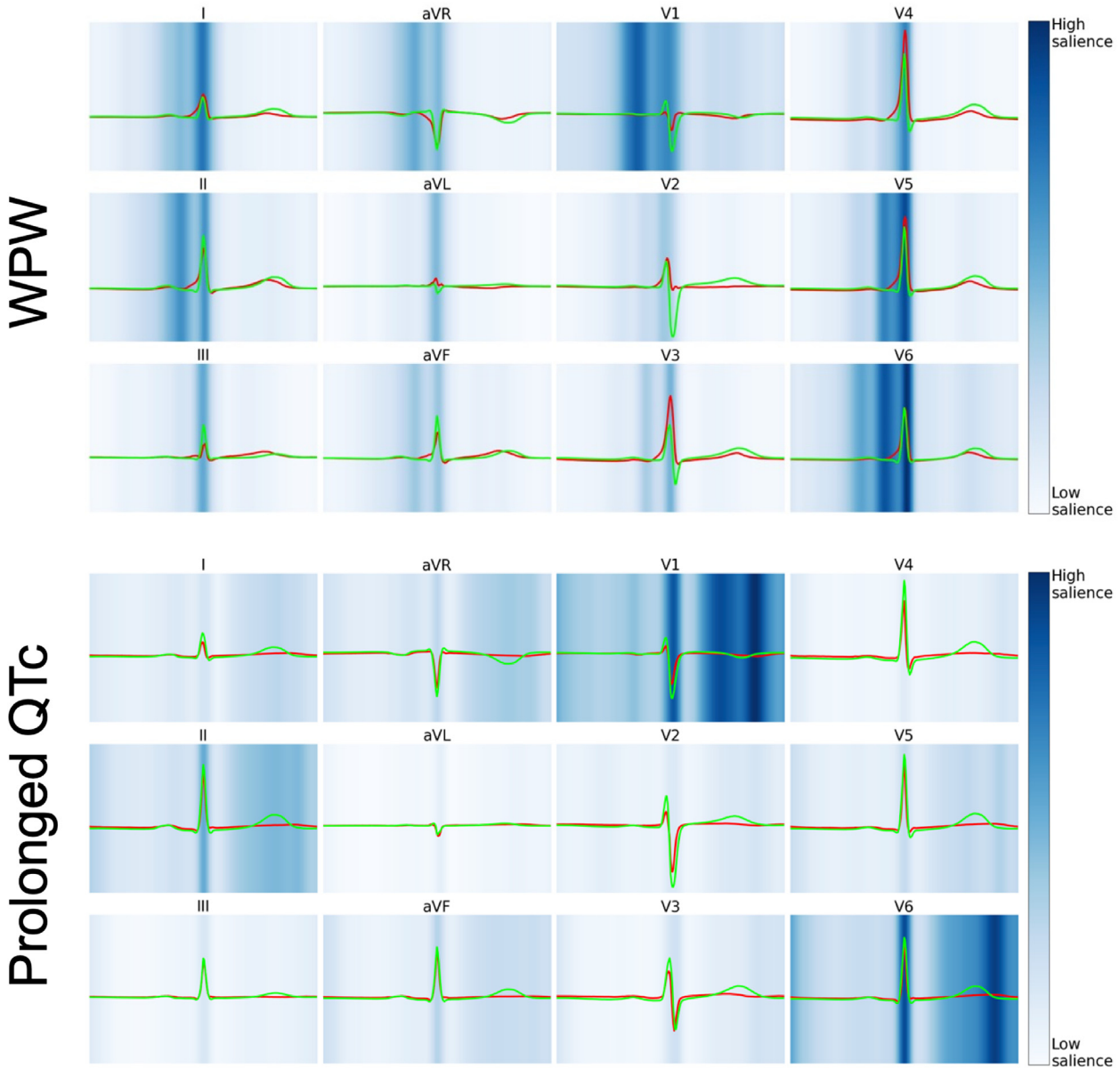
ECGs deemed as FPs (AI-ECG positive but original reader negative) or false negatives (FN; AI-ECG negative but original reader positive) were readjudicated. (A) Heatmap of classifications (positive = dark; negative = light) by original reader (Original), AI-ECG, and 4 expert readers. Hierarchical clustering of AI-ECG and readers shown above. Fleiss kappa (to assess agreement across experts) shown below. (B) Cohen kappa for agreement between the original reader and expert readjudicator (Original-Expert Agreement; black) vs the AI-ECG and expert readjudicator (AI-ECG-Expert Agreement; blue). The average from 4 experts is shown with Student's *t*-test comparisons between Original-Exert Agreement and AI-ECG-Expert Agreement. Abbreviations in Figure 1.

MUSE interpretations as a benchmark. As shown in Figure 1A, our model outperformed MUSE across all outcomes of interest, with more prominent differences for highly relevant pediatric diagnoses (eg, WPW, prolonged QTc).

CLINICAL SIGNIFICANCE AND IMPLICATIONS. From a clinical and translational perspective, we envision that this rapid, reliable, and automated pediatric and congenital AI-ECG model may: 1) facilitate improved access to care; 2) promote screening programs; and 3) enhance physician workflow to reduce workload/burnout.

Although pediatric and congenital heart disease is relatively common, unfortunately >90% of children in low- and middle-income countries do not have access to cardiovascular care.[19] Similarly, there are geographic and socioeconomic disparities in access to care from expert cardiologists in the United States.[20] Delays in expert ECG reads may cause hold-ups in care or missed care opportunities. Given that ECG is a ubiquitous and inexpensive diagnostic test, we

FIGURE 4 **Model Explainability**

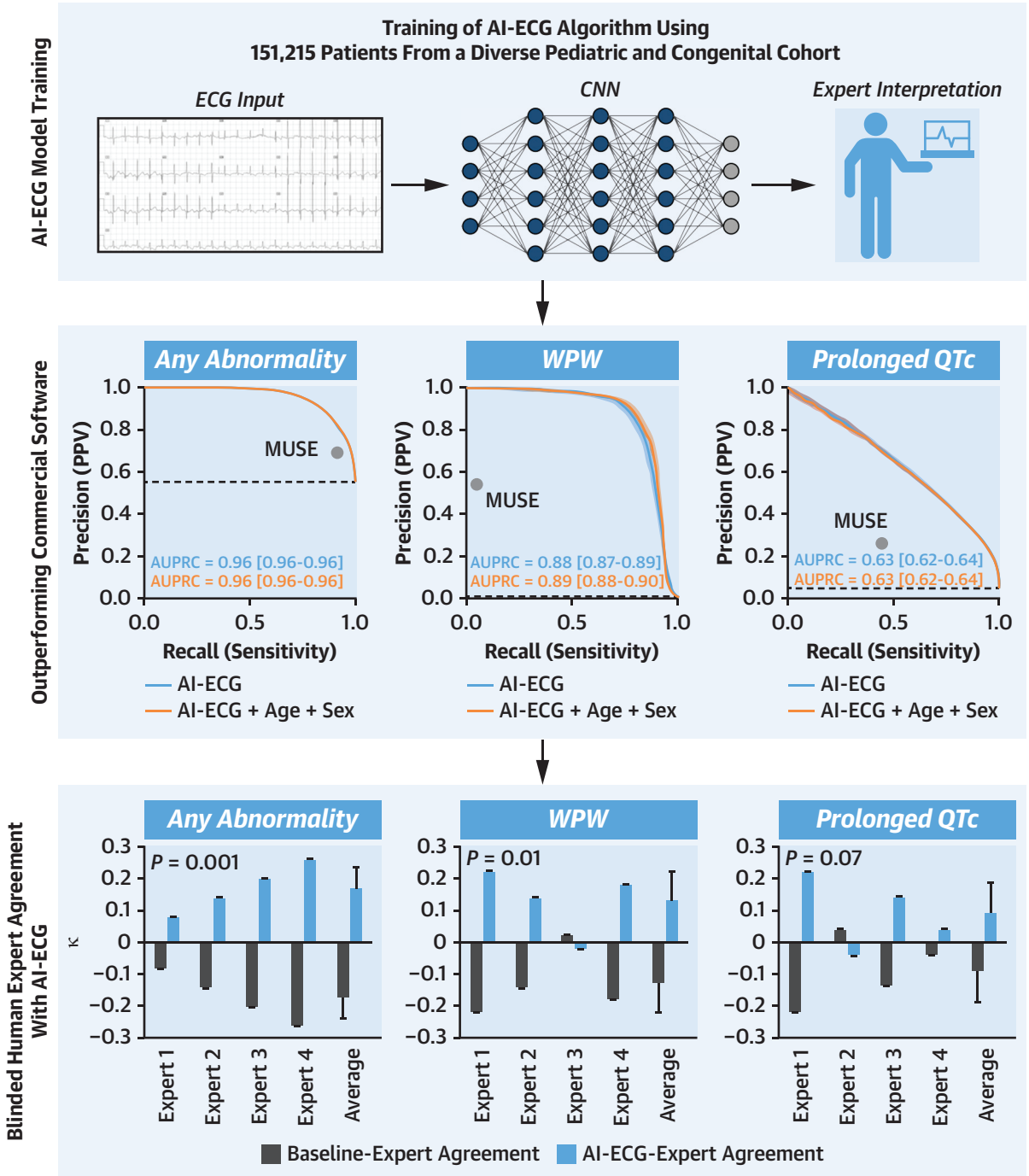


Visualization of high-risk (red) and low-risk (green) median waveforms for WPW (top) and prolonged QTc (bottom). Saliency mapping demarcates ECG regions with greatest (dark blue) and least (light blue) influence on each diagnosis. Abbreviations in Figure 1.

envision this tool may help democratize pediatric cardiology care by providing ECG interpretation for all and aid in prioritizing patients for referral, particularly for screening of WPW and prolonged QTc in areas with limited access to care. In addition, an AI-ECG would inform the provider if the ECG is normal or abnormal, and for ECGs deemed abnormal, a list of predicted ECG diagnoses would be available.

Large-scale ECG screening has been considered for decades within the pediatric cardiology community for sudden cardiac arrest and young athlete sports clearance.[3,18,21,22] However, the lack of an experienced workforce and lack of inter-reader consensus/reproducibility have undermined these efforts. It is conceivable that an AI-ECG may help address these limitations to enable such screening efforts, which

CENTRAL ILLUSTRATION **Artificial Intelligence-Enhanced Electrocardiogram for Expert-Level Interpretation of the Pediatric Electrocardiogram**



Mayourian J, et al. JACC Clin Electrophysiol. 2025;■(■):■-■.

An artificial intelligence-enhanced electrocardiogram (AI-ECG) convolutional neural network (CNN) algorithm trained on a diverse pediatric cohort at Boston Children's Hospital outperformed commercial software to predict any ECG abnormality, Wolff-Parkinson-White syndrome (WPW), and prolonged QTc. Blinded human experts were more likely to agree with the AI-ECG than with the original reader. AUROC = area under the receiver-operating curve; AUPRC = area under the precision-recall curve.

may also enhance physician workflow to reduce workload/burnout. For example, because approximately 50% of ECGs are normal, and our AI-ECG tool performs at an expert level to detect normal ECGs, then conceivably this tool may reduce the total number of ECGs reviewed by an expert by one half. Of note, such promise to reduce workload has been shown for AI-automated quantification of left ventricular ejection fraction from echocardiography.[23]

MODEL INSIGHTS INTO ECG INTERPRETATION. Several insights were developed from model performance and behavior analysis. First, we note that the AI-ECG model performed similarly to the AI-ECG+age+sex model (**Figure 1A**) for primary/secondary outcomes and sinus tachycardia/bradycardia (Supplemental Figure 3), suggesting the model may also learn this demographic data that are known to relate to changes in ECG axis and intervals.[2,18,24,25] Second, we observed that false-positive predictions by the AI-ECG were more likely to be positive on ECGs soon thereafter (with an HR of 88 for WPW), suggesting either initial misdiagnoses by the reader or the AI-ECG detecting subtle findings that become more pronounced on follow-up ECGs (eg, intermittent preexcitation; borderline prolonged QTc). This is supported by the low inter-expert kappa achieved for each outcome (**Figure 3A**), which suggests that the discrepant reads between the original reader and AI-ECG are borderline or difficult boundary cases with significant disagreement even among expert readers. Finally, we note the lower performance of any abnormality and prolonged QTc in newborns (aged <1 week), which is consistent with the literature[26] and may be attributed to the rapid evolution of the normal ECG in the first weeks of life. Indeed, performance increased from 1 week to 1 month and then 6 months, after which it plateaued. In contrast, performance for adults was comparable to the overall cohort, suggestive the model works across the lifespan.

STUDY LIMITATIONS. First, we acknowledge the lack of external validation. Unfortunately, the sizeable network of outside pediatric institutions contacted do not have similar infrastructures available for coding ECG diagnoses, with most using free text rather than coded diagnoses. Although there are public ECG waveforms and diagnoses available,[6] these data sets exclusively represent the general adult population, which is outside the scope of the current study. Future work therefore includes multicenter collaboration (via federated learning[27] to compile a larger and more heterogeneous training set) and external validation. Second, our model architecture requires access to digital waveform data. This limits translation to low-resource settings where digitized data are less accessible and motivates future efforts to generate a model that uses ECG image inputs.[28] Third, our threshold to optimize F1 was used to represent a human reader's behavior when interpreting an ECG; however, if considering a screening tool, other thresholds may be considered (eg, prioritizing negative predictive value or sensitivity). Further consideration is required to weigh the impact of resultant false-positive/false-negative findings. Fourth, other model architectures (eg, foundation model) must be considered for improving performance. Fifth, we acknowledge that a physician's access to history of presenting illness may further enhance ECG interpretation. Finally, the limitations of saliency mapping and model explainability are acknowledged.[29]

FUTURE DIRECTIONS. Translation of the AI-ECG model to clinical practice will require several additional steps, including: 1) multicenter collaboration to improve the power and generalizability of this model; 2) pragmatic randomized clinical trials[30] to study the accuracy and safety of the AI-ECG model and guide clinical implementation; 3) cost-benefit analyses to determine optimal application of pediatric and congenital AI-ECG models to various carefully crafted clinical use cases for screening and disease management; and 4) generation of AI-ECG models that take ECG photo inputs (rather than digital waveform inputs) to broaden global accessibility.[28]

CONCLUSIONS

These findings show the promise of AI-ECG to interpret ECGs rapidly and reliably. This tool may facilitate larger screening program efforts, improved access to care, decreased physician workload, and potentially even improved accuracy of ECG reading.

DATA AVAILABILITY AND SOFTWARE

Requests for Boston Children's Hospital data and related materials will be internally reviewed to clarify if the request is subject to intellectual property or confidentiality constraints. Shareable data and materials will be released under a material transfer agreement for non-commercial research purposes. Use of Boston Children's Hospital data was approved by their respective Institutional Review Boards.

Programming code used to perform the analyses are available upon reasonable request. The convolutional neural network used the Keras framework with a TensorFlow (Google) backend using Python 3.9 (Python Software Foundation). Deep learning was executed on institutional graphics processing units. All other preprocessing and postprocessing code was written in Python 3.9 and R 4.0 (R Foundation for Statistical Computing), which was executed locally.

ACKNOWLEDGMENTS The authors acknowledge Boston Children's Hospital's High-Performance Computing Resources Clusters Enkefalos 2 (E2) made available for conducting the research reported in this publication.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Funding support was received from the Thrasher Research Fund Early Career Award (Dr Mayourian), Boston Children's Hospital Electrophysiology Research Education Fund and Kostin Innovation Fund (Drs Mayourian and Triedman), and National Institutes of Health grant R00-LM012926 from the National Library of Medicine (Dr La Cava). The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr John K. Triedman, Department of Cardiology, Boston Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. E-mail: john.triedman@cardio.chboston.org.

PERSPECTIVES

COMPETENCY IN PATIENT CARE: AI-ECG analysis shows promise to provide expert-level automated diagnosis of the pediatric 12-lead ECG. This model outperformed commercial software, and blinded experts were more likely to agree with the model than the original reader to detect any ECG abnormality, WPW, and prolonged QTc.

TRANSLATIONAL OUTLOOK: Our model provides expert-level automated diagnosis of the pediatric 12-lead ECG, which may promote screening programs, facilitate improved access to expert care, and decrease physician workload.

REFERENCES

1. Khairy P, Marelli AJ. Clinical use of electrocardiography in adults with congenital heart disease. *Circulation*. 2007;116(23):2734–2746.

2. Saarel EV, Granger S, Kaltman JR, et al. Electrocardiograms in healthy North American children in the digital age. *Circ Arrhythm Electrophysiol*. 2018;11(7):e005808. https://doi.org/10.1161/CIRCEP.117.005808

3. Myerburg RJ, Vetter VL. Electrocardiograms should be included in preparticipation screening of athletes. *Circulation*. 2007;116(22):2616–2626 [discussion 2626].

4. Vetter VL, Elia J, Erickson C, et al. Cardiovascular monitoring of children and adolescents with heart disease receiving medications for attention deficit/hyperactivity disorder [corrected]: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young Congenital Cardiac Defects Committee and the Council on Cardiovascular Nursing. *Circulation*. 2008;117(18):2407–2423.

5. Vetter VL. Electrocardiographic screening of all infants, children, and teenagers should be performed. *Circulation*. 2014;130(8):688–697 [discussion 697].

6. Ribeiro AH, Ribeiro MH, Paixao GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1760. https://doi.org/10.1038/s41467-020-15432-4

7. Mayourian J, La Cava WG, Vaid A, et al. Pediatric ECG-based deep learning to predict left ventricular dysfunction and remodeling. *Circulation*. 2024;149(12):917–931.

8. Mayourian J, Gearhart A, La Cava WG, et al. Deep learning-based electrocardiogram analysis predicts biventricular dysfunction and dilation in congenital heart disease. *J Am Coll Cardiol*. 2024;84(9):815–828.

9. Einthoven W. Weiteres über das Elektrokardiogramm. *Archiv für die gesamte Physiologie des Menschen und der Tiere*. 1908;122(12):517–584.

10. Goldberger E. A simple, indifferent, electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar, extremity leads. *Am Heart J*. 1942;23(4):483–492.

11. Colan SD. Early database initiatives: the Fyler codes. In: Barach PR, Jacobs JP, Lipshultz SE, Laussen PC, eds. *Pediatric and Congenital Cardiac Care: Volume 1: Outcomes Analysis*. Springer; 2015:163–169.

12. Sangha V, Mortazavi BJ, Haimovich AD, et al. Automated multilabel diagnosis on electrocardiographic images and signals. *Nat Commun*. 2022;13(1):1583. https://doi.org/10.1038/s41467-022-29153-3

13. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276–282.

14. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.

15. McCormick AD, Lim HM, Strohacker CM, et al. Paediatric cardiology training: burnout, fulfilment, and fears. *Cardiol Young*. 2023;33(11):2274–2281.

16. Gustafsson S, Gedon D, Lampa E, et al. Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients. *Sci Rep*. 2022;12(1):19615. https://doi.org/10.1038/s41598-022-24254-x

17. Lin C, Liu W-T, Chang C-H, et al. Artificial Intelligence–Powered Rapid Identification of ST-Elevation Myocardial Infarction via Electrocardiogram (ARISE)—a pragmatic randomized controlled trial. *NEJM AI*. 2024;1(7):AIoa2400190. https://doi.org/10.1056/AIoa2400190

18. Bratincsak A, Kimata C, Limm-Chan BN, Vincent KP, Williams MR, Perry JC. Electrocardiogram standards for children and young adults using z-scores. *Circ Arrhythm Electrophysiol*. 2020;13(8):e008253. https://doi.org/10.1161/CIRCEP.119.008253

19. Zheleva B, Atwood JB. The invisible child: childhood heart disease in global health. *Lancet*. 2017;389(10064):16–18.

20. Kim JH, Cisneros T, Nguyen A, van Meijgaard J, Warraich HJ. Geographic disparities in access to cardiologists in the United States. *J Am Coll Cardiol*. 2024;84(3):315–316.

21. Chaitman BR. An electrocardiogram should not be included in routine preparticipation screening

of young athletes. *Circulation*. 2007;116(22):2610–2614 [discussion 2615].

22. Halkin A, Steinvil A, Rosso R, Adler A, Rozovski U, Viskin S. Preventing sudden death of athletes with electrocardiographic screening: what is the absolute benefit and how much will it cost? *J Am Coll Cardiol*. 2012;60(22):2271–2276.

23. He B, Kwan AC, Cho JH, et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature*. 2023;616(7957):520–524.

24. O'Sullivan D, Anjewierden S, Greason G, et al. Pediatric sex estimation using AI-enabled ECG analysis: influence of pubertal development. *NPJ Digit Med*. 2024;7(1):176. https://doi.org/10.1038/s41746-024-01165-x

25. Dickinson DF. The normal ECG in childhood and adolescence. *Heart*. 2005;91(12):1626–1630.

26. Stramba-Badiale M, Karnad DR, Goulene KM, et al. For neonatal ECG screening there is no reason to relinquish old Bazett's correction. *Eur Heart J*. 2018;39(31):2888–2895.

27. Goto S, Solanki D, John JE, et al. Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection. *Circulation*. 2022;146(10):755–769.

28. Sangha V, Nargesi AA, Dhingra LS, et al. Detection of left ventricular systolic dysfunction from electrocardiographic images. *Circulation*. 2023;148(9):765–777.

29. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745–e750.

30. 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.

KEY WORDS artificial intelligence, pediatric screening, electrocardiogram, long QTc, WPW

APPENDIX For supplemental tables and a figure, please see the online version of this article.