JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2020 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER
VOL. 75, NO. 7, 2020

Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram

Wei-Yin Ko, MS, MEng,[a,*] Konstantinos C. Siontis, MD,[a,*] Zachi I. Attia, MSEE,[a] Rickey E. Carter, PhD,[b] Suraj Kapa, MD,[a] Steve R. Ommen, MD,[a] Steven J. Demuth, BA,[c] Michael J. Ackerman, MD, PhD,[a] Bernard J. Gersh, MB, ChB DPhil,[a] Adelaide M. Arruda-Olson, MD, PhD,[a] Jeffrey B. Geske, MD,[a] Samuel J. Asirvatham, MD,[a] Francisco Lopez-Jimenez, MD,[a] Rick A. Nishimura, MD,[a] Paul A. Friedman, MD,[a] Peter A. Noseworthy, MD[a]

ABSTRACT

BACKGROUND Hypertrophic cardiomyopathy (HCM) is an uncommon but important cause of sudden cardiac death.

OBJECTIVES This study sought to develop an artificial intelligence approach for the detection of HCM based on 12-lead electrocardiography (ECG).

METHODS A convolutional neural network (CNN) was trained and validated using digital 12-lead ECG from 2,448 patients with a verified HCM diagnosis and 51,153 non-HCM age- and sex-matched control subjects. The ability of the CNN to detect HCM was then tested on a different dataset of 612 HCM and 12,788 control subjects.

RESULTS In the combined datasets, mean age was 54.8 ± 15.9 years for the HCM group and 57.5 ± 15.5 years for the control group. After training and validation, the area under the curve (AUC) of the CNN in the validation dataset was 0.95 (95% confidence interval [CI]: 0.94 to 0.97) at the optimal probability threshold of 11% for having HCM. When applying this probability threshold to the testing dataset, the CNN's AUC was 0.96 (95% CI: 0.95 to 0.96) with sensitivity 87% and specificity 90%. In subgroup analyses, the AUC was 0.95 (95% CI: 0.94 to 0.97) among patients with left ventricular hypertrophy by ECG criteria and 0.95 (95% CI: 0.90 to 1.00) among patients with a normal ECG. The model performed particularly well in younger patients (sensitivity 95%, specificity 92%). In patients with HCM with and without sarcomeric mutations, the model-derived median probabilities for having HCM were 97% and 96%, respectively.

CONCLUSIONS ECG-based detection of HCM by an artificial intelligence algorithm can be achieved with high diagnostic performance, particularly in younger patients. This model requires further refinement and external validation, but it may hold promise for HCM screening. (J Am Coll Cardiol 2020;75:722–33) © 2020 by the American College of Cardiology Foundation.

Listen to this manuscript's audio summary by Editor-in-Chief Dr. Valentin Fuster on JACC.org.

From the [a]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota; [b]Health Sciences Research, Mayo Clinic College of Medicine, Jacksonville, Florida; and [c]Information Technology, Mayo Clinic, Rochester, Minnesota. *W-Y Ko and Dr. Siontis are joint first authors and contributed equally to this work. Mr. Attia has equity in Eko, Inc.; and has served as an advisor to AliveCor. Dr. Kapa owns equity in Eko, Inc.; has served on the advisory boards of Myant, Inc. and Boston Scientific; and has received research grants form Abbott, Inc. and Toray, Inc. The funding source (The Louis V. Gerstner, Jr. Fund at Vanguard Charitable) had no role in the design study of the study, the collection, analysis, or interpretation of data, in the writing of the report or in the decision to submit the paper for publication. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Manuscript received December 1, 2019; accepted December 4, 2019.

ISSN 0735-1097/$36.00
https://doi.org/10.1016/j.jacc.2019.12.030

Hypertrophic cardiomyopathy (HCM) is among the leading causes of sudden cardiac death among adolescents and young adults and is associated with significant morbidity in all age groups (1). The implications of an HCM diagnosis are important for sudden cardiac death risk stratification, genetic counseling, family screening, and longitudinal clinical follow-up. However, the condition is uncommon, affecting approximately 1 in 200 to 500 individuals (2,3). It can also be difficult to distinguish HCM from left ventricular hypertrophy (LVH) due to other causes. Therefore, even though echocardiography is the mainstay modality for the diagnosis and initial evaluation of HCM, the optimal approach to HCM detection in asymptomatic individuals is unknown.

SEE PAGE 734

More than 90% of patients with HCM have electrocardiographic abnormalities (4) and 12-lead electrocardiography (ECG) may offer an attractive noninvasive, low-cost, and rapid means of screening for the condition. However, the associated ECG characteristics are nonspecific so ECG screening is limited by high false-positive rates (5). Generally, ECG screening has relied on manual or automated detection of particular features, such as LVH, left axis deviation, prominent Q waves, and T-wave inversions. Such approaches have insufficient diagnostic performance to justify routine ECG screening (6).

Nevertheless, it may be possible to refine the ECG detection of HCM by leveraging the power of state-of-the-art computing technology, large datasets, nonlinear models, and automated features extraction using convolution layers that allow the artificial intelligence (AI) network to "see" features that are not obvious to even an expert ECG interpreter. Indeed, we have used this approach successfully to identify LV dysfunction with a model area under the curve (AUC) of 0.93 (7). Accordingly, the aim of the current study was to train, validate, and test an AI-based deep learning approach for the detection of HCM based solely on the 12-lead ECG in a large cohort of HCM cases and non-HCM control subjects.

METHODS

DATA SOURCES AND STUDY POPULATION. Following institutional review board approval, we obtained data from the Mayo Clinic digital data vault. We identified 3,060 adult patients (18 years or older) with a diagnosis of HCM based on standard diagnostic criteria (8) who had at least 1 digital, standard, 10-s, 12-lead ECG acquired in the supine position between July 1, 1987, and November 30, 2017. All HCM patients had been evaluated in the HCM clinic at our institution. The HCM diagnoses have been validated previously by detailed manual review of records and the cohort has been used in other publications from our institution (9,10). Among patients with >1 ECG, the first ECG per patient was selected for network training, validation, or testing. ECGs with presence of ventricular pacing or left bundle branch block were not included. Patients had comprehensive 2-dimensional, Doppler, and/or 3-dimensional echocardiography. Quantitative echocardiographic data were documented at the time of the acquisition in a Mayo Clinic developed database (Echo Image Management System; EIMS, Rochester, Minnesota).

To derive a control group, we screened a different population of 87,715 patients who had an ECG and echocardiogram as part of routine clinical practice during the same time period. This control population had been identified previously for a project assessing the value of an algorithm for LV systolic dysfunction detection (7). After applying the same exclusion criteria as in the HCM group (presence of pacing or left bundle branch block), we identified 76,397 patients. Because this sample may have included patients with HCM who had not been evaluated at the HCM clinic in our institution (thus not included in the HCM group), we screened all patients for presence of diagnostic codes for HCM (International Classification of Diseases-9th Revision [ICD-9] 425.1, 425.11, 425.18 and ICD-10 I42.1, I42.2). A total of 2,854 patients with at least 1 HCM diagnostic code were identified and excluded because these cases had not been individually verified by the HCM clinic and to avoid contamination of the control sample with possible HCM cases. The ensuing 73,543 control subjects were then included in a match based on sex and age (neighborhood caliper of 5 years) to the patients with HCM. Each patient with HCM was matched to as many non-HCM control subjects as possible while maintaining a stable case to control match ratio for all patients. This resulted in 63,941 control patients being matched to the 3,060 patients with HCM. The sex and age distributions of the resulting training, validation, and testing case groups are the same as in the control groups. The remaining 9,602 control patients could not be matched because there were no patients with HCM that satisfied the sex- and age-match conditions.

ECGs in both the HCM and the control groups were digital, standard, 10-s, 12-lead ECGs acquired in the supine position at a sampling rate of 500 Hz using a GE-Marquette ECG machine (Marquette, Wisconsin). ECGs were stored using the MUSE data management

ABBREVIATIONS AND ACRONYMS

AI = artificial intelligence

AUC = area under the curve

CI = confidence interval

CNN = convolutional neural network

ECG = electrocardiography

HCM = hypertrophic cardiomyopathy

ICD-9 = International Classification of Diseases-9th Revision

LV = left ventricular

LVH = left ventricular hypertrophy

TABLE 1 Glossary of Terms

- **Adam optimizer:** Adam, short for adaptive moment estimation, is an optimizer used to refine network performance. All neural networks are trained with the use of an optimizer that controls the network parameters which change in each training step until an optimal solution is reached. In practice, Adam is one of many algorithms that assist in the optimization by updating the weights of the neurons in a neural network.
- **Binary cross entropy:** Binary cross entropy is a common measure of a binary classification model performance that acts as the overall error score that the optimizer is minimizing.
- **Cochrane-Armitage trend test:** The Cochrane-Armitage trend test is a statistical test used in categorical data analysis in order to assess for the presence of an ordered association between a variable with 2 categories and an ordinal variable with k categories.
- **Epoch:** An epoch is 1 complete presentation of the dataset to the developing neural network during the training process. Typically, networks are trained in an iterative process with many presentations of the dataset (or epochs) during their learning phase.
- **Hyperparameters:** In machine learning, a hyperparameter is a parameter whose value must be set before the learning process can begin. The hyperparameters dictate the behavior of the deep learning model, such as the number of layers in the network, the size of the convolutional kernels, etc. and the parameters that control the learning stage as batch size, learning rate and etc. By contrast, the values of other parameters (model weights) are derived via training.
- **Keras Framework:** Keras is an open-source neural-network library written in the software language Python. It is designed to enable fast experimentation with deep neural networks. It focuses on being user-friendly, modular, and extensible.
- **Loss function:** Loss function measures how well the observed data fits the assumed data structure for the outcome (e.g., a binomial distribution for a classification problem with 2 classes). Specifically, it is an objective function used in the model estimation process. An optimal combination of hyperparameters and parameter estimates that minimize this function can be considered the best model-based solution.
- **Network architecture:** Network architecture is the design/structure of a computer network and serves as a framework for the specification of a network's components and their functional organization and configuration. More generally, the network architecture defines the series of mathematical operations that translate the input data to an estimated classification.
- **Python:** Python is a programming language used for deep learning research, among other purposes. Python is designed to be easy to read and simple to implement. It is open source, and free to use, even for commercial applications.
- **TensorFlow:** TensorFlow is a machine learning platform, created by Google (Mountain View, California) that enables the usage of deep learning models and low-level mathematical applications based on tensors (multi-dimensional numeric arrays of data). TensorFlow is an open source software library designed to make it easier for developers to design, build, and train deep learning models.
- **Wilson score intervals:** Wilson score interval is a formula for the binomial proportion confidence interval.
- **Youden index:** The Youden index, the sum of the sensitivity and specificity minus 1, is used to identify an ideal model 'cut point' to optimize both sensitivity and specificity.

This list of terms can be used to understand the methods used in designing the convolutional neural network.

system (GE Healthcare, Chicago, Illinois) for later retrieval, which is routine practice at our institution.

Baseline characteristics in the HCM and control groups were defined using inpatient or outpatient ICD-9 and ICD-10 diagnostic codes preceding the index ECG. Categorical variables are reported as absolute numbers and percentages, and continuous variables are reported as mean ± SD. Categorical variables were compared with chi-square and continuous variables were compared using Student's *t*-test.

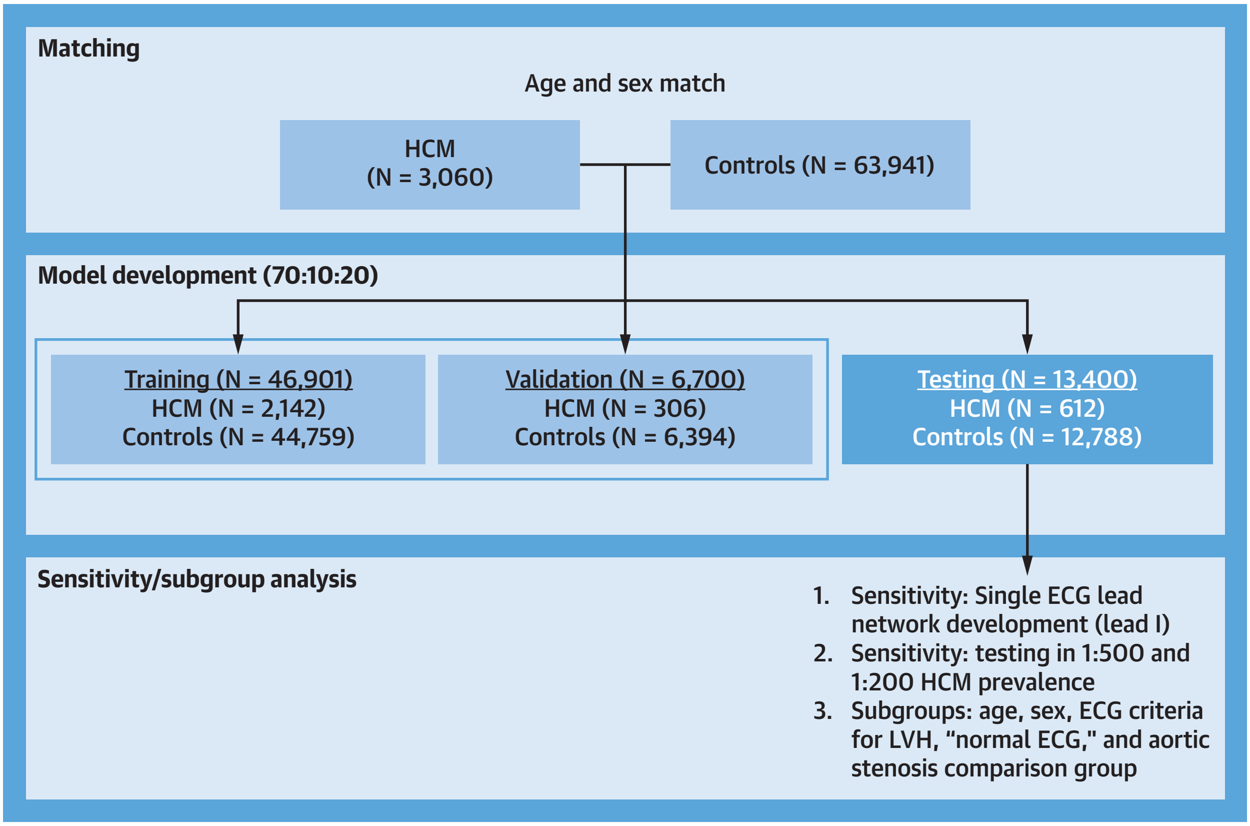
MODEL DEVELOPMENT. Each ECG was converted to a 12 × 5,000 matrix (i.e., 500 Hz sampling over each of 12 separate leads; the first dimension is spatial and the second dimension is temporal). We then applied a convolutional neural network (CNN) using the Keras Framework with a TensorFlow backend (Google, Mountain View, California) and Python (Python Software Foundation, Beaverton, Oregon) (see Table 1 for a glossary of terms). The internal validation dataset was used to optimize the network architecture (identify hyperparameters, batch, and step size). We tested multiple networks and selected the simplest (i.e., the one with fewer parameters or layers) that resulted in the highest AUC of the receiver-operating characteristic curve.

We used a 70%-10%-20% split of both the HCM and the control cohorts for training, validating, and testing of the CNN. This resulted in 46,901 patients with HCM and control patients for training our model, 6,700 for validating the results after each epoch, and 13,400 for testing (Figure 1). We used 1 ECG per patient, and each ECG represents a unique patient. None of the patients overlap among the 3 groups.

For training, ECG were fed to the network and the network weights were updated using the Adam optimizer with binary cross entropy as the loss function. After each epoch, the network was tested using the internal validation dataset. The network hyperparameters were also tuned during this process, and the network with the lowest binary cross entropy loss value was selected once the loss value on the validation set stops decreasing for 5 epochs.

The model that provided the optimal AUC is similar in concept and structure to a previously developed model identifying the ejection fraction based on the 12-lead ECG (7). However, instead of single-dimensional convolutions within the leads, the current model convoluted across the leads. This resulted in a slight AUC score improvement. The model also added an additional convolution before each max pooling layer. Finally, we replaced the multiple fully connected layers at the end with a global average pooling layer, which made the model more efficient computationally and also resulted in a slight

FIGURE 1 **Model Development**


Matching
Age and sex match
HCM (N = 3,060)
Controls (N = 63,941)
Model development (70:10:20)
Training (N = 46,901)
HCM (N = 2,142)
Controls (N = 44,759)
Validation (N = 6,700)
HCM (N = 306)
Controls (N = 6,394)
Testing (N = 13,400)
HCM (N = 612)
Controls (N = 12,788)
Sensitivity/subgroup analysis
1. Sensitivity: Single ECG lead network development (lead I)
2. Sensitivity: testing in 1:500 and 1:200 HCM prevalence
3. Subgroups: age, sex, ECG criteria for LVH, "normal ECG," and aortic stenosis comparison group

A flow diagram indicating the selection of patients for the training, validation, and testing cohorts. ECG = electrocardiography; HCM = hypertrophic cardiomyopathy; LVH = left ventricular hypertrophy.

AUC-receiver-operating characteristic curve score improvement. A sigmoid function for binary classification was used as the final activation.

AI-ENABLED ECG TO DETECT HCM. The main outcome was the ability of the network to identify patients with an HCM diagnosis using the 12-lead ECG alone. After selecting the optimal network using the training and validation datasets, we used the validation dataset receiver-operating characteristic curve to select the optimal probability threshold for binary classification of a test result based on the algorithm's predicted probability for HCM (best combination of sensitivity and specificity, or Youden index). The optimal HCM probability threshold was 11% (i.e., the probability value that an ECG belongs to a patient with HCM) with corresponding test sensitivity and specificity of 90% and 90%, respectively. The CNN model was then applied in the testing dataset to assess its ability to detect HCM and the test was considered positive for HCM if the resulting probability value was >11%. Diagnostic performance parameters were calculated based on this threshold probability value. In sensitivity analyses, we repeated the assessment of the validated CNN in the testing dataset using alternative probability thresholds for HCM of 5%, 25%, 50%, and 75%. In another sensitivity analysis, we repeated training, validation, and testing of the model using only lead I of the ECG, rather than the 12-lead ECG.

To describe the sampling variations for measures of diagnostic performance, 95% confidence intervals (CIs) were estimated using Wilson score intervals. The diagnostic odds ratio, which is the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease, and its associated large sample CI were used to quantify the

TABLE 2 Baseline Characteristics

	Overall			Training			Validation			Testing		
	HCM (n = 3,060)	Control (n = 63,941)	p Value	HCM (n = 2,142)	Control (n = 44,759)	p Value	HCM (n = 306)	Control (n = 6,394)	p Value	HCM (n = 612)	Control (n = 12,788)	p Value
Female	1,357 (44)	27,556 (43)	0.18	971 (45)	19,306 (43)	0.05	120 (39)	2,724 (43)	0.27	266 (44)	5,526 (43)	0.94
Age, yrs	54.8 ± 15.9	57.5 ± 15.5	<0.001	54.6 ± 15.9	57.5 ± 15.5	<0.001	55.7 ± 15.8	57.5 ± 15.5	0.05	54.8 ± 15.8	57.5 ± 15.5	<0.001
CAD	583 (19)	16,958 (27)	<0.001	393 (18)	11,819 (26)	<0.001	67 (22)	1,698 (27)	0.08	123 (20)	3,441 (27)	<0.001
AF	602 (20)	9,986 (16)	<0.001	426 (20)	6,990 (16)	<0.001	67 (22)	961 (15)	0.002	109 (18)	2,035 (16)	0.23
CVA	95 (3)	2,964 (5)	<0.001	69 (3)	2,049 (5)	0.004	8 (3)	318 (5)	0.08	18 (3)	597 (5)	0.06
Diabetes	185 (6)	6,821 (11)	<0.001	118 (6)	4,803 (11)	<0.001	21 (7)	697 (11)	0.03	46 (8)	1,321 (10)	0.03
Hypertension	857 (28)	21,174 (33)	<0.001	593 (28)	14,812 (33)	<0.001	91 (30)	2,150 (34)	0.18	173 (28)	4,212 (33)	0.02
PAD	235 (8)	10,612 (17)	<0.001	154 (8)	7,382 (16)	<0.001	34 (11)	1,108 (17)	0.006	47 (8)	2,122 (17)	<0.001
EF, %	65.8 ± 8.5	57.6 ± 10	<0.001	65.9 ± 8.6	57.6 ± 10	<0.001	66 ± 7.5	57.5 ± 9.8	<0.001	65.5 ± 8.7	57.6 ± 9.9	<0.001
Septum, mm	18.2 ± 5.6	10.6 ± 6.8	<0.001	18.1 ± 5.5	10.6 ± 8.1	<0.001	18 ± 5.6	10.5 ± 2	<0.001	18.3 ± 5.7	10.6 ± 2.1	<0.001
Posterior wall, mm	13.1 ± 3	10.3 ± 2.5	<0.001	13.2 ± 3.1	10.3 ± 2.7	<0.001	13.1 ± 2.8	10.2 ± 1.9	<0.001	13 ± 2.9	10.2 ± 1.9	<0.001

Values are n (%) or mean ± SD.

AF = atrial fibrillation; CAD = coronary artery disease; CVA = cerebrovascular accident; EF = ejection fraction; HCM = hypertrophic cardiomyopathy; PAD = peripheral arterial disease.

degree of discrimination for the algorithm across subgroups of interest. For testing trends in diagnostic performance across age groups, the Cochrane-Armitage trend test was used.

SUBGROUP ANALYSES. In subgroup analyses, we assessed the performance of the trained and validated model in subsets of patients in the testing set defined by age group (<40, 40 to 49, 50 to 59, 60 to 69, 70 to 79, 80+ years), in the subset of patients satisfying Sokolow-Lyon ECG criteria for LVH, and those with a "normal ECG" interpretation. We also assessed the performance of the model in distinguishing between the patients with HCM and the patients in the control group with a diagnosis of aortic stenosis (based on diagnostic codes).

We also identified 860 patients with HCM who had undergone genotyping for sarcomeric mutations. We assessed the performance of the CNN expressed as the rate of false-negative results based on the 11% optimal probability threshold, and the model-derived HCM probability among HCM cases with and without sarcomeric mutations.

PREVALENCE-SPECIFIC ANALYSIS. We assessed model performance in a study sample with an HCM prevalence of 1:500 (mirroring the general population). This was performed by running multiple experiments, with each experiment randomly selecting 50% of the 12,788 control patients along with 500 random patients with HCM in the test group. Repeating the experiment 100× allowed us to get an estimate of the AUC score for a possible screening program considering the prevalence of HCM in the general population. A similar simulation was performed assuming HCM population prevalence of 1:200 (2).

RESULTS

STUDY POPULATION. In the combined training, validation, and testing datasets, the age and sex distributions between patients with HCM and control patients were similar (HCM: mean age 54.8 ± 15.9 years, 44% female; control: mean age 57.5 ± 15.5 years, 43% female). In the HCM and control groups, 540 (17%) and 9,262 patients (14%) were <40 years old, whereas 40 (1%) and 858 patients (1%) were <20 years old, respectively. Table 2 shows patient characteristics for the overall, training, validation, and testing datasets in the HCM and control groups. Prevalence of LVH by ECG, determined by the Sokolow criteria, was 1,443 of 3,060 (47%) and 5,913 of 63,941 (9%) in the 2 respective groups.

MODEL PERFORMANCE. Following training and validation, the AUC of the CNN in the validation dataset was 0.95 (95% CI: 0.94 to 0.97) with the optimal probability threshold of 11%, which results in sensitivity and specificity of 90% and 90%, respectively, for having HCM. When applying this probability threshold in the test dataset, the CNN's AUC was 0.96 (95% CI: 0.95 to 0.96) with sensitivity 87% and specificity 90% (Figure 2). Positive predictive value was 31% and negative predictive value was 99%. The distribution of model-derived HCM probabilities in the HCM cases and control subjects is shown in Figure 3.

Among the control subjects with false-positive detections, 31% satisfied the LVH ECG criteria. A comparison of the clinical characteristics of patients with false and true positives, and those with true and false negatives is shown in Table 3. In a model developed and tested based only on a single ECG lead

(lead I), the AUC was 0.91 (95% CI: 0.89 to 0.92), sensitivity was 83%, and specificity was 81%.

When higher HCM probability thresholds were applied, the performance characteristics changed to favor specificity and to reduce the false-positive rate (Table 4). For example, with a probability threshold of 75%, specificity was 99% and false-positive rate was 1%. The false-positive rate was low and the negative predictive value was high across all tested thresholds, ranging from 1% to 15% and from 98% to 99%, respectively.

Figure 4 demonstrates a case example of the network's performance in a patient with HCM before and after septal myectomy. This patient was not included in either the model training, validation, or testing datasets because they presented to our institution after the completion of the initial phase of the study.

SUBGROUP ANALYSES. Results of test characteristics in subgroups defined by age and sex are shown in Figure 5. The model performance was similar among men and women, but it performed particularly well in younger (<40 years old) patients (sensitivity 95%; specificity 92%; diagnostic odds ratio: 195.0; 95% CI: 84.3 to 451.2).

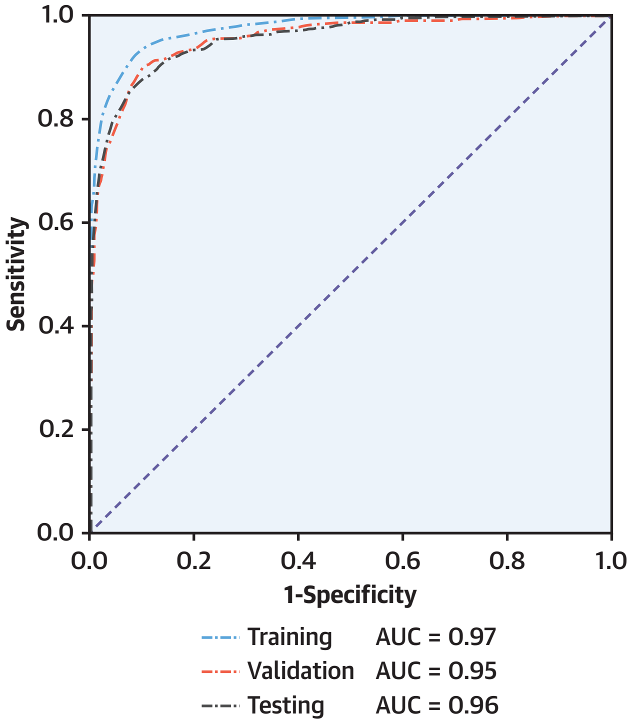
The AUC for predicting HCM when the population of interest was restricted only to patients with LVH by ECG criteria was 0.95 (95% CI: 0.94 to 0.97). Sensitivity, specificity, and positive and negative predictive values for HCM were 97%, 68%, 41%, and 99%, respectively, at an HCM probability threshold of 11%. When restricted to patients <40 years old with LVH, model performance was 97%, 74%, 55%, and 99% for sensitivity, specificity, and positive and negative predictive values, respectively.

The AUC for predicting HCM when the population of interest was restricted only to patients with a "normal ECG" interpretation was 0.95 (95% CI: 0.90 to 1.00). Sensitivity, specificity, and positive and negative predictive values for HCM were 93%, 87%, 31%, and 99%, respectively, at an HCM probability threshold of 11%.

When assessing model performance for HCM cases versus patients in the control group with a diagnosis of aortic stenosis (n = 1,419), the AUC was 0.94 (95% CI: 0.93 to 0.95), with sensitivity and specificity of 87% and 86%, respectively.

Among HCM patients with (n = 286) and without (n = 574) sarcomeric mutations, the model-derived probabilities for an ECG diagnosis of HCM were median 97% (interquartile range: 80% to 99%) and 96% (interquartile range: 70% to 99%), respectively. False-negative rates for HCM detection among the HCM cases were 3.5% and 8% in patients with and without a sarcomeric mutation, respectively.

FIGURE 2 Model Performance



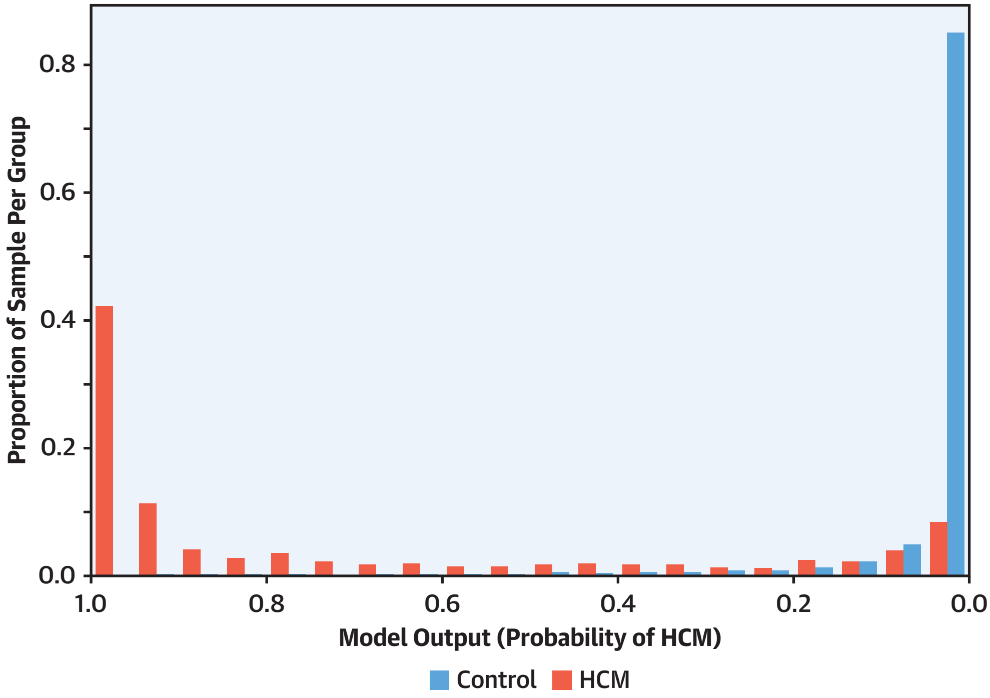
Receiver-operating characteristic curve for the performance of the artificial intelligence-based convolutional neural network to identify patients with hypertrophic cardiomyopathy diagnosis in the training, validation, and testing cohorts. AUC = area under the curve.

PREVALENCE-SPECIFIC ANALYSIS. In a sensitivity analysis applying the network to a testing subset with HCM prevalence that more closely mirrors the general population (1:500 HCM-control), the averaged model performance was similar to the main analysis with a mean AUC of 0.95 ± 0.023. The mean AUC was also similar when we simulated an HCM prevalence of 1:200 in the testing dataset (AUC: 0.96 ± 0.013).

DISCUSSION

In this study, we report the first AI-based CNN for detection of HCM based on the 12-lead ECG (Central Illustration). The network demonstrates high discriminatory ability in distinguishing HCM cases from non-HCM control cases with an AUC of 0.96. The network is associated with low false-positive rates and very high negative predictive value across a range of HCM probability thresholds. The model performance was best in our youngest age

FIGURE 3 Distribution of Model Outputs



Distribution of model-derived hypertrophic cardiomyopathy (HCM) probabilities in the HCM cases and control cases in the testing dataset.

cohort (<40 years) in this study, however, the study did not include any patients <18 years and young adults were under-represented. Furthermore, in the subgroup of subjects with ECG criteria of LVH, model performance is nearly identical to the overall population with AUC of 0.95 (95% CI: 0.94 to 0.97), suggesting that the model is valuable in distinguishing HCM from other causes of LVH. Similar performance was demonstrated among subjects with ECG that had been considered completely normal by standard interpretation. Model performance was high both in groups of patients with and without sarcomeric mutations, even though the rate of false-negative test results was lower among those with sarcomeric mutations.

TABLE 3 Comparison of Patient Characteristics Between True- Versus False-Positive Results and True- Versus False-Negative Results in the Testing Dataset

	True Positive (n = 534)	False Positive (n = 1,226)	p Value	True Negative (n = 11,562)	False Negative (n = 78)	p Value
Female	233 (44)	482 (39)	0.10	5,044 (44)	33 (42)	0.90
Age, yrs	53.8 ± 15.7	59.4 ± 16.3	<0.001	57.3 ± 15.4	61.3 ± 15.2	0.023
EF, %	65.8 ± 8.3	55.6 ± 11.7	<0.001	57.7 ± 9.7	62.8 ± 11.1	0.0019
Septum, mm	18.8 ± 5.7	11 ± 2.4	<0.001	10.5 ± 2.1	14.7 ± 3.8	<0.001
Posterior wall, mm	13.1 ± 2.9	10.6 ± 2	<0.001	10.2 ± 1.9	12.3 ± 2.5	<0.001
LVH by ECG criteria	263 (49)	379 (31)	<0.001	805 (7)	8 (10)	0.36

Values are n (%) or mean ± SD, unless otherwise indicated.
Abbreviations as in Table 2.

Previous research has focused on automated algorithms for ECG-based HCM detection and phenotypic characterization in rather small samples of patients with HCM (11-14). These algorithms relied on certain ECG features of HCM such as high QRS voltage, and abnormal Q or T waves, among others, with performance comparable to that of ECG interpretation by trained electrophysiologists. However, the abnormalities forming the basis of such algorithms are limited, not specific for HCM, and overlap with changes seen in athletic heart conditioning or even normal ECG variants. Furthermore, the ECG can be seemingly normal in about 10% of patients with HCM (4,15). For these reasons, the diagnostic performance of such approaches may be insufficient to justify routine ECG screening. The large-scale model creation and testing reported herein represents the first application of a multilayer CNN that was truly agnostic to any specific ECG features during its development.

Several sets of ECG criteria for HCM in young athletes have been proposed with an emphasis toward distinguishing athletic adaptation and normal ECG variants from HCM, including the European (16), Stanford (17), and Seattle criteria (18), yet follow-up studies and attempts for external validation of these criteria have revealed inconsistencies in their diagnostic performance (19–22). However, the use of an AI-based deep learning network has the potential to overcome these challenges as it is not confined to any “classic” ECG criteria and it is trained to detect even the most subtle ECG patterns associated with structural changes in HCM that may be undetectable by the human eye or by traditional automated algorithms. Even though sudden cardiac death events in athletes are infrequent, such events are tragic, widely publicized, and rekindle the debate regarding the value of population-based screening of competitive athletes. European guidelines have endorsed the use of the ECG as a component of pre-participation screening for athletes, though this practice remains subject to debate in other parts of the world (23). Some data have supported the cost-effectiveness of pre-participation ECG screening (24) and further refinement of the diagnostic performance could tip the scales in favor of screening (22). In this study, we did not specifically study a population of athletes or young adults, and the derived network at its current stage may not have the features of an effective screening tool for widespread application. External validation in other groups is critical, including the evaluation in populations with a higher proportion of younger adults, higher racial admixture, and even higher variation in geographic origin, which has emerged as a determinant of the ECG manifestations of athletic heart adaptation (25).

Screening for uncommon conditions is inherently limited by high false-positive rates and low positive predictive values. When the probability threshold for HCM detection was set at 11% in this study, the false-positive rate was 9% and positive predictive value was 31%. However, it is important to interpret this finding in the appropriate context. For example, when considering application of the network for screening, use of a higher probability threshold for HCM, such as 50% or even 75%, may be appropriate to achieve a higher positive predictive value. Using a probability threshold of 75%, the positive predictive value was 71% and false-positive rate was only 1%. Given the excellent overall performance of the model with an AUC of 0.96 (95% CI: 0.95 to 0.96), the higher probability thresholds should result in only modest reductions in test sensitivity while maintaining a high negative predictive value. Of note, even with the least sensitive (most specific) probability threshold of 75%, the negative predictive value was very high (98%). Thus, a negative test with the developed network practically rules out HCM and can provide reassurance to patients and providers. This very high negative predictive value is important because application of AI methods to the ECG may allow the development of a low-cost screening test that can be used to exclude HCM, provide reassurance, and potentially avoid further costly diagnostic testing. On the other hand, a lower threshold (sensitive test) could even be considered as an alternative to periodic echocardiography for surveillance of yet unaffected relatives of patients with known HCM; however, this scenario would require further study.

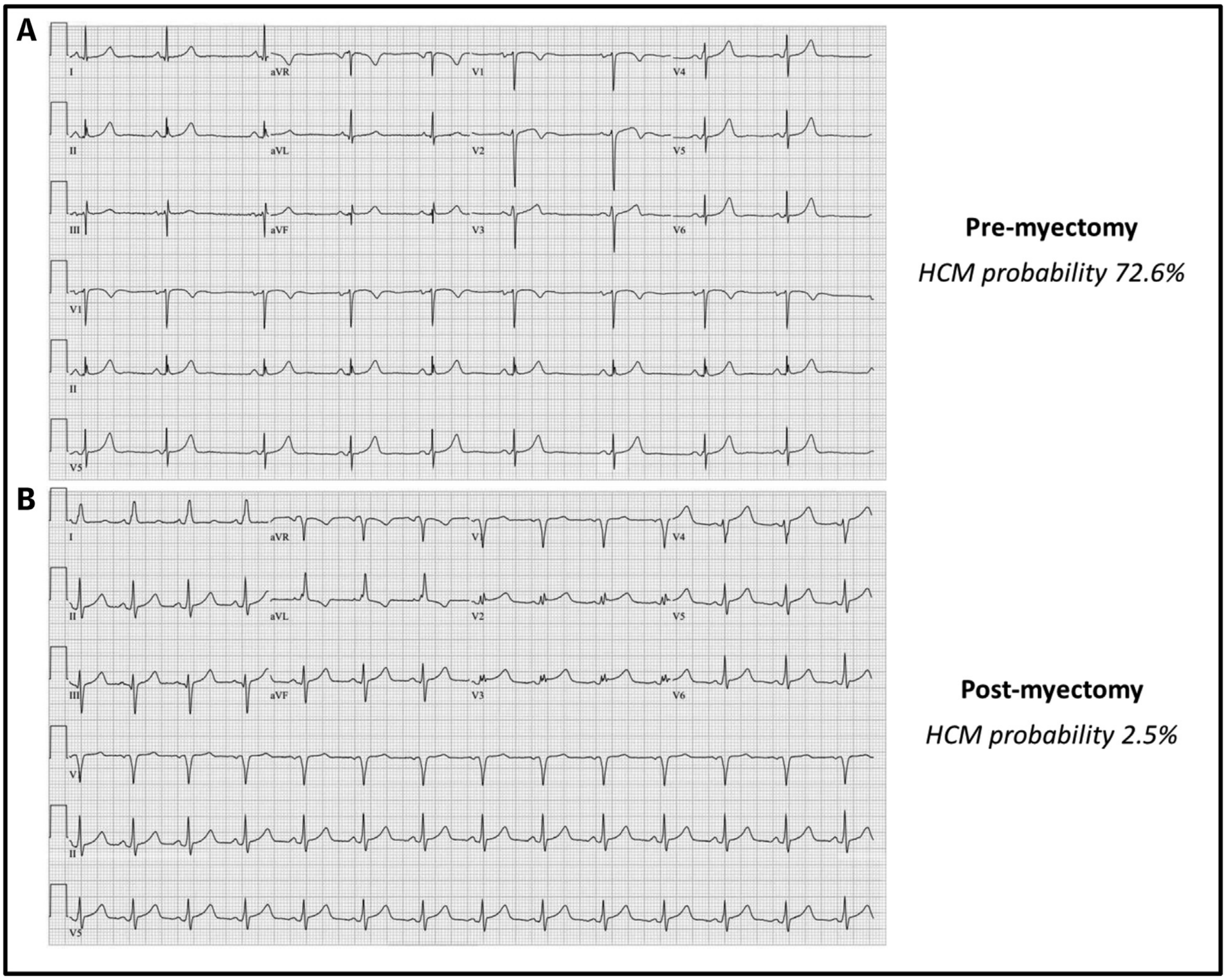
TABLE 4 Performance Characteristics Using Various Thresholds of Network-Derived Probability of HCM

Probability Threshold	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	False Positive	False Negative
5	92	85	23	99	15	9
11	87	90	31	99	9	13
25	82	94	41	99	6	18
50	73	97	57	99	3	27
75	64	99	71	98	1	36

Values are %.

In subgroup analyses, we noted that the diagnostic performance of the network was optimal in younger subjects, while it was comparatively lower in older subjects. This may be attributed to the fact that a larger number of potential comorbidities, resulting in ECG changes, may be present as confounders in older populations, such as LVH, ischemic heart disease, atrial fibrillation, and kyphoscoliosis, among others. Regardless, the practical use and value of the ECG-based AI network reported herein may be more pertinent for younger patients where a distinction between HCM and normal ECG variants is important. Among all tested subgroups, model performance was superior in young (age <40 years) female patients. Women with HCM are diagnosed and referred for specialty care with delay, likely leading to worse outcomes than for men (10,26). Application of the AI-enabled ECG as an HCM detection tool in this subgroup may help combat this effect. Furthermore, when restricting the analyses to the patients with LVH by ECG criteria, the model performance was unchanged (AUC: 0.95; 95% CI: 0.94 to 0.97), suggesting that the network has sufficient discriminatory

FIGURE 4 Clinical Example of Model Performance

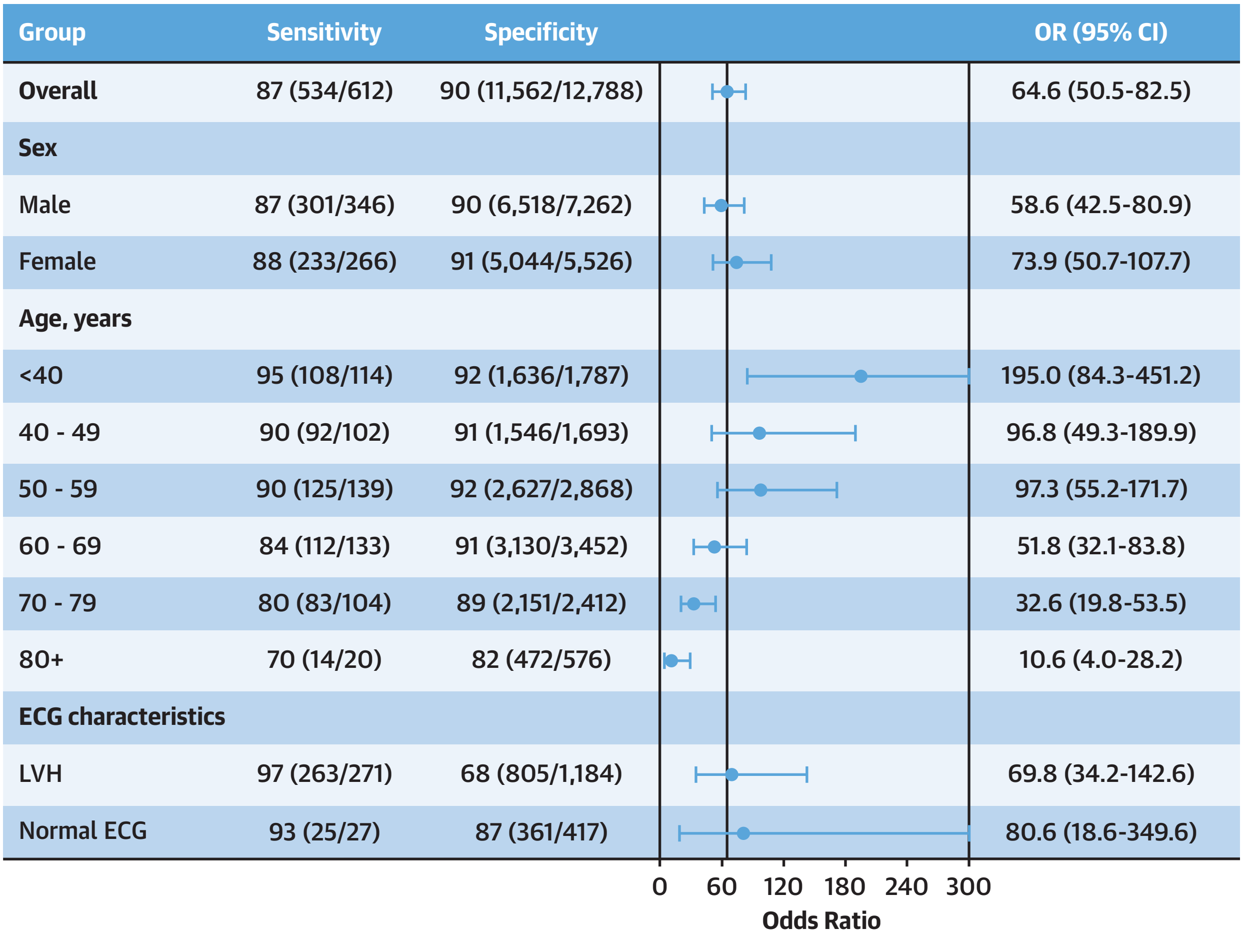


Artificial intelligence model performance in a 21-year-old woman with obstructive hypertrophic cardiomyopathy (HCM) before **(A)** and after **(B)** septal myectomy. Prior to myectomy, the patient had massive septal hypertrophy (30 mm).

ability to distinguish between true HCM and non-HCM LVH—a very important clinical dilemma. The network also demonstrated high diagnostic performance for HCM among patients with a seemingly entirely normal ECG, a subset that can comprise up to 10% of patients with HCM. This observation may be particularly important because no manually applied, rule-based approach to ECG-based HCM screening would be valuable in this group of patients. It also underscores the fact that the model is “seeing” patterns through the convolutional process that are not readily apparent to the naked eye.

The network-derived probabilities for HCM diagnosis were very high in genotyped HCM patients regardless of their sarcomeric mutation status. We did not perform a formal comparison of the false-negative rates in the groups of patients with and without sarcomere mutations because mutation status was not a balanced variable across the training, validation, and testing datasets. However, in absolute terms, the rate of false negatives was higher in mutation-negative than in mutation-positive patients. This could be because ECG manifestations may be less evident in patients with HCM without a pathogenic sarcomeric mutation. Another explanation is that some of the HCM cases without a sarcomeric mutation may represent non-HCM causes of LVH. Even though it is difficult to distinguish these

FIGURE 5 **Subgroup Analyses**



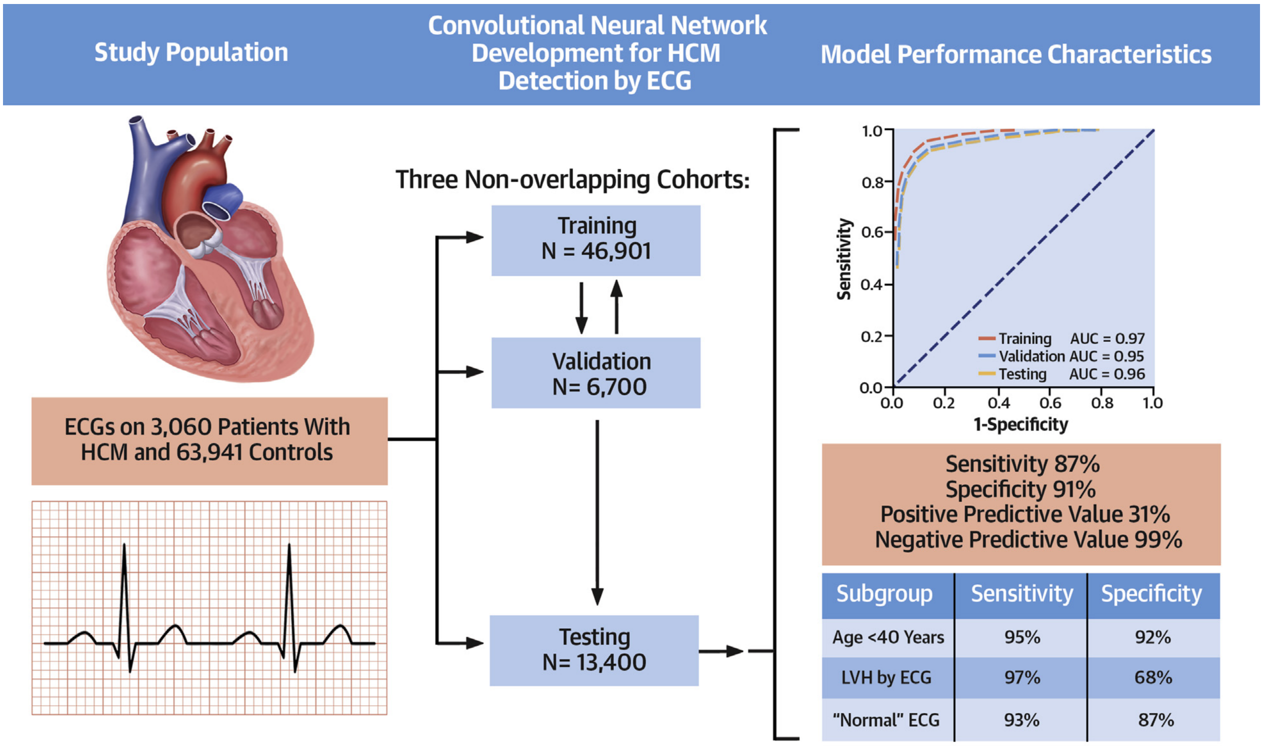
Group	Sensitivity	Specificity	OR (95% CI)
Overall	87 (534/612)	90 (11,562/12,788)	64.6 (50.5-82.5)
Sex			
Male	87 (301/346)	90 (6,518/7,262)	58.6 (42.5-80.9)
Female	88 (233/266)	91 (5,044/5,526)	73.9 (50.7-107.7)
Age, years			
<40	95 (108/114)	92 (1,636/1,787)	195.0 (84.3-451.2)
40 - 49	90 (92/102)	91 (1,546/1,693)	96.8 (49.3-189.9)
50 - 59	90 (125/139)	92 (2,627/2,868)	97.3 (55.2-171.7)
60 - 69	84 (112/133)	91 (3,130/3,452)	51.8 (32.1-83.8)
70 - 79	80 (83/104)	89 (2,151/2,412)	32.6 (19.8-53.5)
80+	70 (14/20)	82 (472/576)	10.6 (4.0-28.2)
ECG characteristics			
LVH	97 (263/271)	68 (805/1,184)	69.8 (34.2-142.6)
Normal ECG	93 (25/27)	87 (361/417)	80.6 (18.6-349.6)

Forest plot demonstrating the subgroup results of network performance by age, sex, and presence of LVH by ECG criteria, and the presence of a normal ECG. CI = confidence interval; OR = odd ratio; other abbreviations as in Figure 1.

cases retrospectively, the model-derived “not HCM” diagnoses can be considered as true negatives rather than false negatives in these subjects. With a degree of subjectivity frequently involved in establishing an HCM diagnosis, the use of an AI-developed ECG-based tool may help improve diagnostic certainty in cases that are otherwise considered borderline based on standard clinical and imaging criteria.

STUDY LIMITATIONS. First, this is a single-center study with an HCM population comprised of a large proportion of referral patients, thus applicability in other settings remains to be determined. Second, even though all HCM diagnoses have been validated as part of standard clinical evaluation at our institution, we cannot entirely rule out that a small number of patients with athletic heart conditioning or non-HCM LVH were classified as HCM based on information available at the time of the evaluation. Third, it should be emphasized that subjects’ race, geographic origin, and athlete status were not available (25). We also note that some patients considered as “false positives” in our study may in fact have undiagnosed HCM. This bias could result in underestimation of the accuracy of the test. Finally, even though this engineered, validated, and tested AI model performed best in the younger adults in this study and provides encouragement for the prospects of a more effective HCM screening program for adolescents, no subjects in this current study were <18 years old. As such, the

CENTRAL ILLUSTRATION Training, Validation, and Testing of an AI-Based Electrocardiography Screen for Hypertrophic Cardiomyopathy



Study Population

Convolutional Neural Network Development for HCM Detection by ECG

Model Performance Characteristics

ECGs on 3,060 Patients With HCM and 63,941 Controls

Three Non-overlapping Cohorts:

Training
N = 46,901

Validation
N= 6,700

Testing
N= 13,400

Sensitivity 87%
Specificity 91%
Positive Predictive Value 31%
Negative Predictive Value 99%

Subgroup	Sensitivity	Specificity
Age <40 Years	95%	92%
LVH by ECG	97%	68%
"Normal" ECG	93%	87%

Ko, W.-Y. et al. J Am Coll Cardiol. 2020;75(7):722-33.

The study population included 3,060 patients with hypertrophic cardiomyopathy (HCM) and 63,941 control patients that contributed to mutually exclusive training, validation, and testing cohorts. Model performance was assessed in the "untouched" testing cohort and demonstrated an area under the curve (AUC) of 0.96, with 87% sensitivity and 91% specificity. The model performance was consistent in younger populations (age <40 years old), those with left ventricular hypertrophy (LVH) by ECG, and those with a "normal" ECG. AI = artificial intelligence; ECG = electrocardiography.

performance characteristics of this novel AI network in children and adolescents remains to be determined. Lastly, one of the key limitations in existing neural networks is lack of "explainability." Through the convolutional process, the precise features the network sees are obscured; however, determining what features contribute to model performance is an area of active ongoing investigation by our team.

CONCLUSIONS

A fully automated, AI-based network to detect HCM based on the standard, 12-lead ECG is feasible and can be performed with high diagnostic performance. After refinement and external validation in less selected and more heterogeneous populations, our model may possibly help improve the yield of current approaches to HCM diagnosis and of population-based ECG screening for those at risk of HCM-related adverse events.

ADDRESS FOR CORRESPONDENCE: Dr. Peter A. Noseworthy, Department of Cardiovascular Medicine, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905. E-mail: Noseworthy.Peter@mayo.edu. Twitter: @noseworthypeter.

PERSPECTIVES

COMPETENCY IN PATIENT CARE AND PROCEDURAL SKILLS: A deep learning algorithm that incorporates ECG data can accurately identify patients with HCM across a range of subgroups including those with normal ECG or LVH patterns, particularly in young patients.

TRANSLATIONAL OUTLOOK: The model may improve ECG screening for HCM in future studies.

REFERENCES

1. Maron BJ, Haas TS, Murphy CJ, Ahluwalia A, Rutten-Ramos S. Incidence and causes of sudden death in U.S. college athletes. J Am Coll Cardiol 2014;63:1636–43.

2. Semsarian C, Ingles J, Maron MS, Maron BJ. New perspectives on the prevalence of hypertrophic cardiomyopathy. J Am Coll Cardiol 2015;65: 1249–54.

3. Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults: echocardiographic analysis of 4111 subjects in the CARDIA study: Coronary Artery Risk Development in (Young) Adults. Circulation 1995;92: 785–9.

4. McLeod CJ, Ackerman MJ, Nishimura RA, Tajik AJ, Gersh BJ, Ommen SR. Outcome of patients with hypertrophic cardiomyopathy and a normal electrocardiogram. J Am Coll Cardiol 2009;54:229–33.

5. Pelliccia A, Maron BJ, Culasso F, et al. Clinical significance of abnormal electrocardiographic patterns in trained athletes. Circulation 2000;102: 278–84.

6. Maron BJ, Friedman RA, Kligfield P, et al. Assessment of the 12-lead electrocardiogram as a screening test for detection of cardiovascular disease in healthy general populations of young people (12-25 years of age): a scientific statement from the American Heart Association and the American College of Cardiology. J Am Coll Cardiol 2014;64:1479–514.

7. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25:70–4.

8. Gersh BJ, Maron BJ, Bonow RO, et al. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2011;58:2703–38.

9. Siontis KC, Geske JB, Ong K, Nishimura RA, Ommen SR, Gersh BJ. Atrial fibrillation in hypertrophic cardiomyopathy: prevalence, clinical correlations, and mortality in a large high-risk population. J Am Heart Assoc 2014;3:e001002.

10. Geske JB, Ong KC, Siontis KC, et al. Women with hypertrophic cardiomyopathy have worse survival. Eur Heart J 2017;38:3434–40.

11. Campbell MJ, Zhou X, Han C, et al. Pilot study analyzing automated ECG screening of hypertrophic cardiomyopathy. Heart Rhythm 2017;14: 848–52.

12. Rahman QA, Tereshchenko LG, Kongkatong M, Abraham T, Abraham MR, Shatkay H. Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification. IEEE Trans Nanobioscience 2015;14:505–12.

13. Ouyang N, Yamauchi K. Using a neural network to diagnose the hypertrophic portions of hypertrophic cardiomyopathy. MD Comput 1998;15: 106–9.

14. Lyon A, Ariga R, Minchole A, et al. Distinct ECG phenotypes identified in hypertrophic cardiomyopathy using machine learning associate with arrhythmic risk markers. Front Physiol 2018;9:213.

15. Rowin EJ, Maron BJ, Appelbaum E, et al. Significance of false negative electrocardiograms in preparticipation screening of athletes for hypertrophic cardiomyopathy. Am J Cardiol 2012;110: 1027–32.

16. Corrado D, Pelliccia A, Heidbuchel H, et al. Recommendations for interpretation of 12-lead electrocardiogram in the athlete. Eur Heart J 2010;31:243–59.

17. Uberoi A, Stein R, Perez MV, et al. Interpretation of the electrocardiogram of young athletes. Circulation 2011;124:746–57.

18. Drezner JA, Ackerman MJ, Anderson J, et al. Electrocardiographic interpretation in athletes: the "Seattle criteria." Br J Sports Med 2013;47:122–4.

19. Sheikh N, Papadakis M, Ghani S, et al. Comparison of electrocardiographic criteria for the detection of cardiac abnormalities in elite black and white athletes. Circulation 2014;129:1637–49.

20. Pickham D, Zarafshar S, Sani D, Kumar N, Froelicher V. Comparison of three ECG criteria for athlete pre-participation screening. J Electrocardiol 2014;47:769–74.

21. Brosnan M, La Gerche A, Kumar S, Lo W, Kalman J, Prior D. Modest agreement in ECG interpretation limits the application of ECG screening in young athletes. Heart Rhythm 2015;12:130–6.

22. Dhutia H, Malhotra A, Gabus V, et al. Cost implications of using different ECG criteria for screening young athletes in the United Kingdom. J Am Coll Cardiol 2016;68:702–11.

23. Corrado D, Pelliccia A, Bjornstad HH, et al. Cardiovascular pre-participation screening of young competitive athletes for prevention of sudden death: proposal for a common European protocol: consensus statement of the Study Group of Sport Cardiology of the Working Group of Cardiac Rehabilitation and Exercise Physiology and the Working Group of Myocardial and Pericardial Diseases of the European Society of Cardiology. Eur Heart J 2005;26:516–24.

24. Wheeler MT, Heidenreich PA, Froelicher VF, Hlatky MA, Ashley EA. Cost-effectiveness of preparticipation screening for prevention of sudden cardiac death in young athletes. Ann Intern Med 2010;152:276–86.

25. Riding NR, Sharma S, McClean G, Adamuz C, Watt V, Wilson MG. Impact of geographical origin upon the electrical and structural manifestations of the black athlete's heart. Eur Heart J 2019;40: 50–8.

26. Olivotto I, Maron MS, Adabag AS, et al. Gender-related differences in the clinical presentation and outcome of hypertrophic cardiomyopathy. J Am Coll Cardiol 2005;46:480–7.

KEY WORDS artificial intelligence, diagnostic performance, electrocardiogram, hypertrophic cardiomyopathy