ORIGINAL RESEARCH ARTICLE

Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection

Shinichi Goto, MD, PhD; Divyarajsinhji Solanki, MS; Jenine E. John, MD; Ryuichiro Yagi, MD, MPH; Max Homilius, PhD; Genki Ichihara, MD; Yoshinori Katsumata, MD, PhD; Hanna K. Gaggin, MD, MPH; Yuji Itabashi, MD, PhD; Calum A. MacRae, MD, PhD; Rahul C. Deo, MD, PhD

BACKGROUND: Novel targeted treatments increase the need for prompt hypertrophic cardiomyopathy (HCM) detection. However, its low prevalence (0.5%) and resemblance to common diseases present challenges that may benefit from automated machine learning–based approaches. We aimed to develop machine learning models to detect HCM and to differentiate it from other cardiac conditions using ECGs and echocardiograms, with robust generalizability across multiple cohorts.

METHODS: Single-institution HCM ECG models were trained and validated on external data. Multi-institution models for ECG and echocardiogram were trained on data from 3 academic medical centers in the United States and Japan using a federated learning approach, which enables training on distributed data without data sharing. Models were validated on held-out test sets for each institution and from a fourth academic medical center and were further evaluated for discrimination of HCM from aortic stenosis, hypertension, and cardiac amyloidosis. Last, automated detection was compared with manual interpretation by 3 cardiologists on a data set with a realistic HCM prevalence.

RESULTS: We identified 74 376 ECGs for 56 129 patients and 8392 echocardiograms for 6825 patients at the 4 academic medical centers. Although ECG models trained on data from each institution displayed excellent discrimination of HCM on internal test data (C statistics, 0.88–0.93), the generalizability was limited, most notably for a model trained in Japan and tested in the United States (C statistic, 0.79–0.82). When trained in a federated manner, discrimination of HCM was excellent across all institutions (C statistics, 0.90–0.96 and 0.90–0.96 for ECG and echocardiogram model, respectively), including for phenotypic subgroups. The models further discriminated HCM from hypertension, aortic stenosis, and cardiac amyloidosis (C statistics, 0.84, 0.83, and 0.88, respectively, for ECG and 0.93, 0.94, 0.85, respectively, for echocardiogram). Analysis of electrocardiography-echocardiography paired data from 11 823 patients from an external institution indicated a higher sensitivity of automated HCM detection at a given positive predictive value compared with cardiologists (0.98 versus 0.81 at a positive predictive value of 0.01 for ECG and 0.78 versus 0.59 at a positive predictive value of 0.24 for echocardiogram).

CONCLUSIONS: Federated learning improved the generalizability of models that use ECGs and echocardiograms to detect and differentiate HCM from other causes of hypertrophy compared with training within a single institution.

Key Words: cardiomyopathy, hypertrophic ■ echocardiography ■ electrocardiography ■ machine learning

---

ORIGINAL RESEARCH ARTICLE

<div style="border">

Clinical Perspective

What Is New?

- Although the ECG is thought to be a well-standardized modality, machine learning models to discriminate hypertrophic cardiomyopathy (HCM) using ECG did not generalize well to data from external sources.
- Federated learning across multiple institutions improved the generalizability of the model to discriminate HCM using ECG without the need to transfer the raw data.
- Compared with detection by cardiologists, a machine learning pipeline combining electrocardiographic and echocardiographic data was able to detect HCM with higher sensitivity at a given specificity.

What Are the Clinical Implications?

- External validation is crucial to evaluate the generalizability of machine learning models, and federated learning can be considered to improve generalizability by training models across multiple institutions without data sharing.
- An ECG and echocardiogram HCM machine learning model may improve high-throughput detection of HCM by automatically analyzing data and indicating the need for further clinical review.

</div>

Nonstandard Abbreviations and Acronyms

AS	aortic valve stenosis
AUROC	area under the receiver-operating characteristics curve
BWH	Brigham and Women's Hospital
CA	cardiac amyloidosis
GRAD-CAM	gradient-weighted class activation mapping
HCM	hypertrophic cardiomyopathy
LVH	left ventricular hypertrophy
MGH	Massachusetts General Hospital
ML	machine learning
MRI	magnetic resonance imaging
PPV	positive predictive value
2D	2-dimensional
UCSF	University of California San Francisco
UMAP	Uniform Manifold Approximation and Projection

Hypertrophic cardiomyopathy (HCM) is a genetic disease of the myocardium arising from mutations in genes encoding proteins that constitute the sarcomere apparatus.[1,2] The heart in HCM is typically hypercontractile and hypertrophied and has reduced compliance. Although initially believed to be a rare dis-ease, recent reports have suggested an HCM prevalence close to 0.5% in the general population.[3]

HCM manifests primarily as shortness of breath with reduced activity and carries a risk of adverse events such as atrial fibrillation, stroke, and sudden cardiac death. Once diagnosed, patients with HCM should be treated with comprehensive strategies, including screening of at-risk family members, pharmacological management of symptoms, and assessment and mitigation of risk of sudden death, including consideration of implantable cardioverter defibrillators.[4] Although there is no approved medical therapy specific to HCM, a newly developed myosin inhibitor, mavacamten, appears to ameliorate symptoms and biomarkers in defined subsets of patients with HCM.[5,6] There is thus an increasing need for systematic detection of this disease. However, because the myocardium also hypertrophies in response to more prevalent stresses such as hypertension and aortic valve stenosis (AS) or infiltrative diseases such as cardiac amyloidosis (CA) and Fabry disease, HCM remains underdiagnosed.[7]

Both echocardiogram and cardiac magnetic resonance imaging (MRI) are used to make a definitive diagnosis of HCM. In addition to these imaging modalities, interrogation of family history, physical examination, and genetic testing play important roles in identifying patients with HCM. Approximately 60% of patients have a clearly recognizable familial disease, and various causal mutations in genes such as *MYH7* and *MYBPC3* have been identified.[1] HCM is also characterized by a distinct histopathological pattern with disarray in the overall architecture of the hypertrophied myocytes.[8] Given the underdetection of HCM, typical workflows involving these diagnostic modalities are clearly not sufficient because of cost, subtlety of findings in early disease, invasiveness of the approach, or availability of the modality, as well as the need for nonspecialist providers either to recall the diagnostic steps needed or to make a timely referral.

HCM can also be detected from characteristic changes in the ECG,[9-13] a more widely available modality. We and others have developed machine learning (ML) strategies to detect HCM from both electrocardiography and echocardiography,[12-17] but to date, these approaches have involved only a single modality and have been evaluated at single centers, which can limit the ability to generalize owing to biases in the distribution and ascertainment of patients. Improving generalizability requires large quantities of data from various institutions. However, in the medical field, training data may include identifiable information, leading to a reluctance to share. Federated learning is an ML technique that allows training of an ML model using data sets from multiple institutions without sharing raw data.[18] The approach was initially adopted to train models on small devices such as smartphones[18] but has begun to be used in medical applications, including training

models for brain tumor segmentation[19] and prediction of COVID-19 outcomes.[20]

In the present study, we aimed to develop an automated workflow to detect HCM from ECGs and echocardiograms using ML models that generalize across multiple clinical settings. To this end, we tested the hypothesis that federated learning would improve generalizability compared with training within a single institution and developed a stepwise approach integrating ECGs and echocardiograms (Figure 1).

METHODS

Data Availability

The data supporting the findings are available on approval of data-sharing committees at the respective institutions. The availability of code is detailed in the Supplemental Methods section.

Definition of HCM Cases and Controls

For all institutions, patients with HCM were first identified from diagnostic codes (*International Classification of Diseases, 10th Revision* codes I42.1 or I42.2) or echocardiography reports and then manually confirmed by chart review. Cases of HCM were defined as patients with maximum left ventricular wall thickness exceeding 15 mm (except for Keio University Hospital, where 13 mm was used as a cutoff because the interventricular septum thickness was measured excluding the right ventricular wall thickness) without any other explanation for ventricular hypertrophy. All patients were required to have been diagnosed with HCM at a specialized clinic within that institution. Controls were selected from those without diagnostic codes I42.1 or I42.2. No other exclusion criteria were applied; thus, no chart review was undertaken for the controls.

Data Collection for ECG Training and Testing Cohorts

To build ECG models for HCM detection, we collected electrocardiographic data from 4 institutions (3 institutions from the United States and 1 from Japan). Although the diagnosis of HCM is made when cardiac hypertrophy is present, because HCM may have subtle manifestations before overt hypertrophy,[21] we included all historic 12-lead electrocardiographic studies for identified cases. After exclusion of ECGs after myectomy or ablation of myocardium, with electronic pacing (both atrial and ventricular), with left bundle-branch block patterns, or with recording <10 seconds, the ECGs from cases were matched on age and sex in a 1:5 ratio (reason detailed in the Supplemental Methods) to ECGs from controls. All ECGs were preprocessed to match 250-Hz sampling (described in Supplemental Methods). After the process, all electrocardiographic data were represented as time series of voltages recorded for 10 seconds at 250 Hz for each lead, which results in 12 sets of voltage vectors of length $250 \times 10 = 2500$ (Figure S1). We converted these data into a 2-dimensional (2D) matrix of shape 12×2500 saved in a binary format designed to hold multidimensional matrix (ie, NumPy array format).

Evaluation of Electrocardiographic Data Heterogeneity Across Institutions

Uniform Manifold Approximation and Projection (UMAP) is an unsupervised ML technique for dimension reduction.[22] This approach can be used to project high-dimensional vector onto a 2D space to allow visual inspection. In UMAP, similar data are placed closer, and data with different features are placed farther apart, allowing identification of data clusters.

To visualize the heterogeneity of the electrocardiographic data across institutions, raw electrocardiographic voltage recordings were projected on a 2D map using UMAP. Because UMAP takes vectors rather than matrices as input, the 2D matrices of ECGs with a shape of 12×2500 were flattened into a vector of length 30000 by first separating the 12-lead data into 12 sets of vectors with a length of 2500 and then concatenating it (Figure S2). ECGs from all 4 institutions were projected onto a single map. The UMAP projection was colored according to institution, HCM status, sampling rate, age, sex, race, and heart rate.

Training and Evaluation of ECG Model Trained With Single-Institution Data

To evaluate the generalizability of the ECG models trained at individual institutions, we trained a convolutional neural network model (detailed in the Supplemental Methods) with 5-fold cross-validation for each institution resulting in 20 models (5 models each for 4 institutions). To this end, data sets from each institution were randomly separated into 5 equal chunks (folds).

Data leakage is a common error in ML. It happens when cases (or controls) used in the testing phase are more similar to cases (or controls) in the training phase than expected by chance. This causes exaggerated performance. In our case, data leakage could happen if ECGs from a single patient were distributed among both training (derivation and validation data set) and test (including >1-fold in cross-validation) data. The model then learns and uses patient-specific rather than disease-specific features. To avoid this, splitting of training data was performed on the patient level. Models were trained for 150 epochs (an epoch indicates 1 pass over the entire training data set completed by the ML algorithm). To avoid overfitting, the model with the highest area under the receiver-operating characteristics curve (AUROC) on the validation fold within the 150 epochs was chosen for evaluation (150 was chosen as a long enough number to avoid underfitting). All 20 models were tested on held-out data for each institution, along with the complete data set from each external institution, which resulted in 5 predictions per institution for each model. The mean and 95% CI of the AUROC were calculated for each model on each institution from these predictions. Subgroup analyses on age, sex, heart rate, and race were performed using the external data set to understand the heterogeneity of the models.

Training and Evaluation of ECG Model With Federated Learning

To train a model generalizable across multiple institutions without explicit comingling of data, we used a federated learning approach.[18] The same architecture and data sets as the individual institution models were used to train the federated learning model. Data from each of the 3 institutions (Massachusetts General
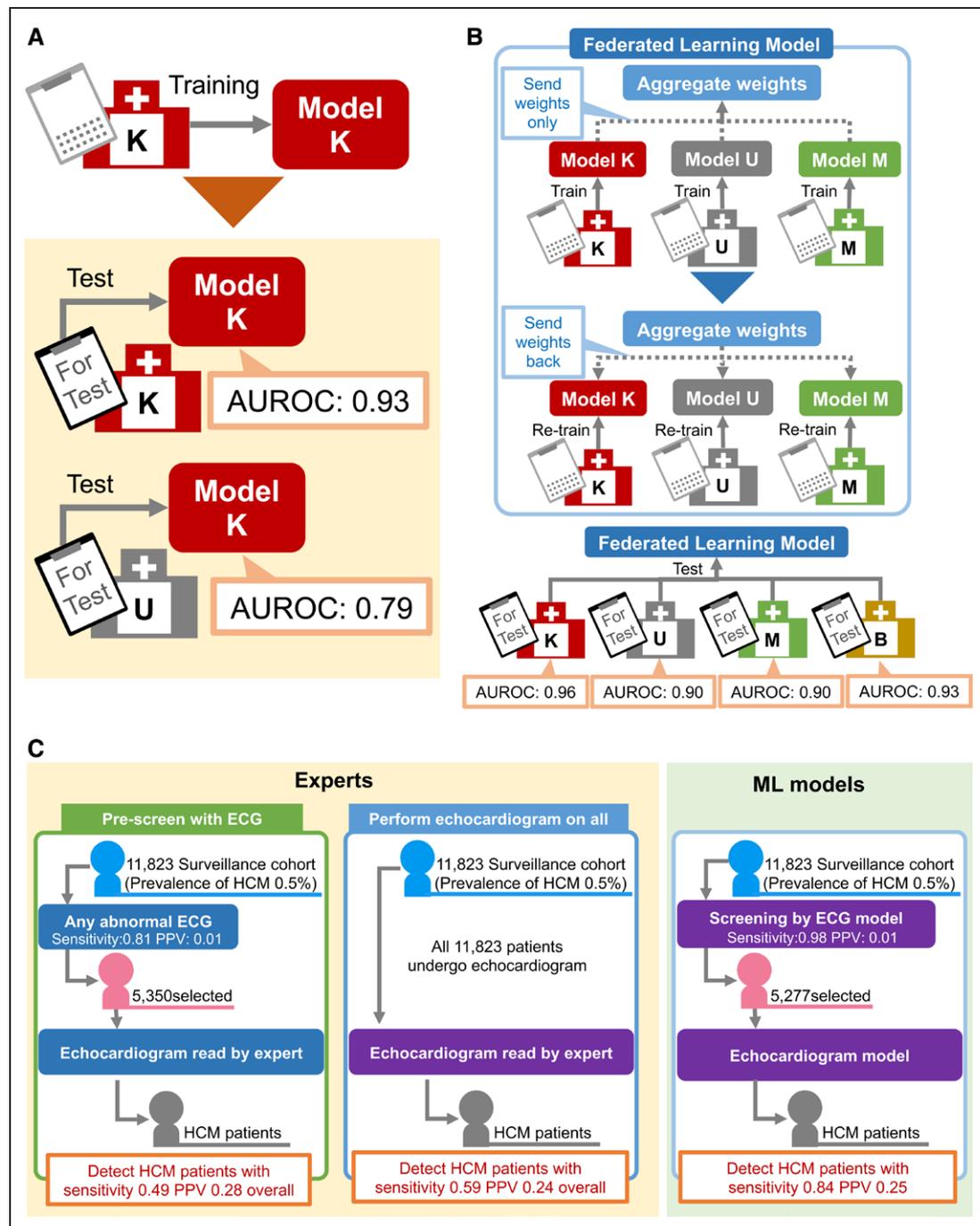
Figure 1. Study overview.

A, The ECG–hypertrophic cardiomyopathy (HCM) models trained on data from a single institution discriminated HCM excellently on a held-out data set from that same institution, but some models generalized poorly on an external data set. **B**, Schematic of the process of federated learning. Multi-institutional models can be trained without data leaving any institution. In our case, ECG and echocardiogram models trained using federated learning not only discriminated HCM well on held-out test sets but also had excellent discrimination on external validation data sets from an independent institution. Schematic of deployment simulation. **C**, A cohort with an HCM prevalence of 0.5% was constructed to reflect prevalence in the general population. The stepwise approach with ECG followed by echocardiogram model achieved a sensitivity of 0.84 at a positive predictive value (PPV) of 0.25, whereas expert cardiologists could achieve a sensitivity of only 0.59 at a PPV of 0.24 even by performing echocardiograms on all patients in this cohort. AUROC indicates area under the receiver-operating characteristics curve; B, Brigham and Women's Hospital; K, Keio University Hospital; M, Massachusetts General Hospital; and U, University of California San Francisco.

Hospital [MGH], University of California San Francisco [UCSF], and Keio University Hospital) were separated into derivation, validation, and test data sets with a 3:1:1 ratio. The split was done at a patient level to ensure that no data from a single patient were allocated to >1 data set. The model was trained by performing multiple steps of individual model training and central aggregation of the models across institutions. A single step consisted of the following sequence. (1) A separate model is trained on data from

each institution for 1 epoch. (2) The models from all institutions are sent to the central server and are aggregated (Supplemental Methods). (3) The central server sends back the aggregated model to each institution. (4) The institution updates the model by training another epoch on their data. The final model (chosen as described previously) was evaluated on the test data set from each of the 3 institutions, along with an external validation data set from Brigham and Women's Hospital (BWH). Analyses were performed comparing various subgroups: HCM with and without outflow tract obstruction (defined as a pressure gradient > 50 mm Hg), apical and nonapical HCM, HCM with and without pathological genetic mutations, and HCM with and without confirmation by MRI. Furthermore, we performed analyses limiting the data to outpatient ECGs and restricting controls to patients with non–left ventricular hypertrophy (LVH) confirmed by echocardiogram. In addition, the discrimination for cases of nonapical HCM before developing LVH was tested. To help understand the features used by the model, activation patterns weighted by the gradient of classification (GRAD-CAM[23]) were visualized.

Hypertension and AS are 2 common reasons for LVH. Because these diseases are much more prevalent compared with HCM, they are a source of misclassification in real-world practice. Although rare, CA resembles HCM morphologically yet requires different treatment. The following comparisons were thus evaluated on BWH data:

1. HCM versus AS with LVH (aortic valve area <1 cm² and left ventricular mass index to body surface area ≥95 for women and ≥115 for men on an echocardiogram within 90 days of the ECG);
2. HCM versus hypertension (median systolic blood pressure >140 mm Hg) with LVH; and
3. HCM versus CA (as described previously[24]).

Training and Evaluation of the Echocardiogram Model With Federated Learning

To train the echocardiogram model to detect HCM, we collected echocardiogram data from the same 4 institutions as the ECGs and defined HCM case status as above. All echocardiogram studies for identified cases were extracted. After exclusion of studies after myectomy/ablation of the myocardium and those with pacemaker or implantable cardioverter defibrillator leads, cases were matched with a 1:3 ratio (reason detailed in Supplemental Methods) on the basis of age and sex to echocardiograms from control patients. Before training, the videos were standardized to 30 frames with 30 frames per second and a squared size of 299×299 (Supplemental Methods). A 3-dimensional convolutional neural network–based model (Supplemental Methods) was trained using federated learning. Echocardiogram models were trained and tested using the same multi-institutional strategy as with the ECG model. A GRAD-CAM image was also used to explore model interpretability. The same subgroups and discrimination between other causes of LVH were tested as the ECG model.

Comparing Sensitivity and Positive Predictive Value of ECG-Echocardiography Models With Cardiologist Interpretation

To reflect real-world prevalence, we assembled data to construct a surveillance cohort with an HCM prevalence of 0.5%.[1]

Patients with an ECG and echocardiogram taken within 30 days were identified within our BWH cohort. A single ECG-echocardiography study pair having the shortest time between electrocardiography and echocardiography studies was selected for each patient, and confirmed cases of HCM and controls were randomly extracted at a 1:200 ratio until the controls were exhausted. We deployed the ECG and echocardiogram HCM models on each study and assessed the discrimination of the ECG and echocardiogram model for HCM using precision recall curve plots. The same data set (ECG and echocardiogram) was labeled by 3 cardiologists for comparison with the models (Supplemental Methods). Because diagnoses by physician are binary rather than continuous values, a specific cutoff (a value at which all observations higher are defined positive and all lower are negative) for the model was required to enable comparison. Cutoffs were selected to match the positive predictive value (PPV) by the physicians' reading, and sensitivity at that cutoff was compared. We additionally evaluated a stepwise approach of an ECG model followed by echocardiogram model.

Statistical Analysis

Continuous values are presented as mean±SD, and categorical values are presented as numbers and percentages if not otherwise specified. The 95% CIs were calculated by the bootstrap method with 2000 bootstrap samples, except for the analyses of models trained at individual institutions, for which the 95% CIs were calculated from the SE of the 5 models generated with 5-fold cross-validation. A 2-tailed value of $P<0.05$ was considered significant.

Ethics Statement

This study complies with all ethics regulations and guidelines. The study protocol was approved by local institutional review boards of Massachusetts General Brigham (2019P002651), UCSF (10-03386), and Keio University Hospital (20200030). Because the study collected data retrospectively, a waiver of informed consent was approved by the institutional review board.

RESULTS

Limitation in Generalizability of HCM-ECG Models

From the 4 participating institutions, 3932, 3802, 1461, and 3201 eligible ECGs from 447, 324, 196, and 141 patients with HCM were identified at BWH, MGH, UCSF, and Keio University Hospital, respectively (Figures S3–S6). The ECGs from patients with HCM showed lower heart rates, longer PR intervals, longer QRS durations, longer corrected QT intervals, higher amplitudes of R wave in V_5/V_6 leads, and deeper S waves in V_1/V_2 leads compared with controls (Tables S1 and S2). In BWH, 96 (21.5%) had apical HCM, 159 (61.9%) had left ventricular outflow tract obstruction, 92 (48.2%) had a pathological genetic mutation, and 275 (61.5%) were confirmed for HCM by MRI. The UMAP projection of

the ECGs revealed heterogeneity of the data, which was most pronounced in ECGs from Keio University Hospital, for which almost no data were present in the middle-left region (Figure 2A). This region could not be mapped unambiguously to patient factors or electrocardiographic characteristics (Figure S7). Although the models trained at individual institutions discriminated HCM excellently with an AUROC of 0.88 to 0.93 on the internal held-out data set (ie, from the same institution), some models did not generalize well to data from other institutions (Figure 2B). Specifically, the model from Keio University Hospital, despite having larger sample size than UCSF and an internal test set AUROC of 0.93, had an AUROC of 0.79, 0.79, and 0.82 on MGH, UCSF, and BWH, respectively. The difference visible in the UMAP partially explained this result: The model trained on data from Keio University Hospital showed significantly degraded performance on external ECG samples that were projected into the cluster, where internal samples were lacking (cluster with versus without internal samples: AUROC, 0.86 versus 0.81, 0.82 versus 0.77, and 0.81 versus 0.79 for BWH, MGH, and UCSF, respectively; Figure S8). No obvious patient subgroup was responsible for the disparate performance (detailed in Supplemental Results and Figure S9).

Federated Learning Improves the Generalizability of the HCM-ECG Model

Federated learning on data from 3 institutions enhanced the overall discrimination of the ECG model and greatly improved generalizability to external cohorts (AUROC, 0.90, 0.90, and 0.96 for MGH, UCSF, and Keio University Hospital, respectively, for the internal test; and AUROC, 0.93 for BWH external validation; Figure 3A and 3B). HCM discrimination varied across phenotypic subgroups, with AUROC values of 0.94, 0.97, 0.92, and 0.92 for HCM with outflow tract obstruction, apical HCM, no outflow tract obstruction, and nonapical HCM, respectively (Figure 3C). The discrimination was not affected by the limitation of cases to those with known pathological genetic mutations (AUROC, 0.94 and 0.93 for with and without mutation, respectively) or with MRI confirmation (AUROC, 0.92 and 0.93 for with and without confirmation, respectively), nor was it affected by limiting the entire cohort to ECGs obtained in outpatient settings (AUROC, 0.93; 16 040 patients) or limiting the controls to those without LVH confirmed by echocardiography (AUROC, 0.93; Figure S10; Figure S11 shows selection of no-LVH controls). Furthermore, the model discriminated HCM from hypertension, severe AS, and CA, although with a slight performance drop (AUROC, 0.84, 0.83, and 0.88, respectively; Figure 3D; Figures S12–S14 and Tables S3–S5 show patient selection). The model was also able to discriminate cases with non-

apical HCM before developing LVH with a slight drop in discrimination (AUROC, 0.88; Figure 3E; Figure S15 shows patient selection). GRAD-CAM revealed that the model was focusing primarily on the QRS complex for those ECGs fulfilling voltage criteria for LVH. For those without high voltage, the model appeared to focus mainly on the QT interval of the ECG (Figure 3F).

Combining data from 3 institutions increases overall sample size and may explain the benefit of a federated learning approach. We thus performed a sensitivity analysis by training a federated learning model after subsampling the training data set to match the sample size from Keio University Hospital. The results revealed much a higher AUROC (0.87, 0.88, 0.96, and 0.91 for MGH, UCSF, Keio University Hospital, and BWH, respectively) compared with the model trained at Keio University Hospital alone (Figure S16).

To understand the additional value of raw electrocardiographic data over traditional electrocardiographic measurements and patient characteristics, 4 additional models were trained (Supplemental Methods). All showed only moderate discrimination for HCM (AUROC, 0.81–0.82 on ECGs from BWH; Table S6).

Excellent Generalizability of an HCM-Echocardiogram Model Trained by Federated Learning

From the 4 participating institutions, 760, 514, 296, and 528 eligible echocardiogram studies from 327, 242, 167, and 172 patients with HCM were identified at BWH, MGH, UCSF, and Keio University Hospital, respectively (Figures S17–S20). Patients with HCM (at the time of the echocardiogram) had lower heart rates, higher ejection fraction, greater interventricular septum thickness, and greater posterior wall thickness compared with controls (Tables S7 and S8). Of the cases with HCM in the BWH cohort, 64 (19.6%) had apical HCM, 125 (61.8%) had left ventricular outflow tract obstruction, 74 (50.0) had pathological genetic mutation, 218 (66.7%) were confirmed by MRI, and 137 (63.4%) had LVH. The echocardiogram model trained with data from 3 institutions using a federated learning approach discriminated HCM excellently across the internal test data sets and external validation data set (AUROC, 0.91, 0.92, and 0.90 for MGH, UCSF, and Keio University Hospital, respectively, for the internal test; and AUROC, 0.96 on BWH external validation; Figure 4A and 4B). There was modest variation across phenotypic subgroups: AUROC of 0.98 and 0.96 for detecting HCM with and without outflow tract obstruction, respectively (Figure 4C); AUROC of 0.94 versus 0.96 for apical versus nonapical HCM; AUROC of 0.96 for HCM with and without genetic mutations; and AUROC of 0.96 versus 0.97 for patients with and without MRI confirmation, respectively. HCM
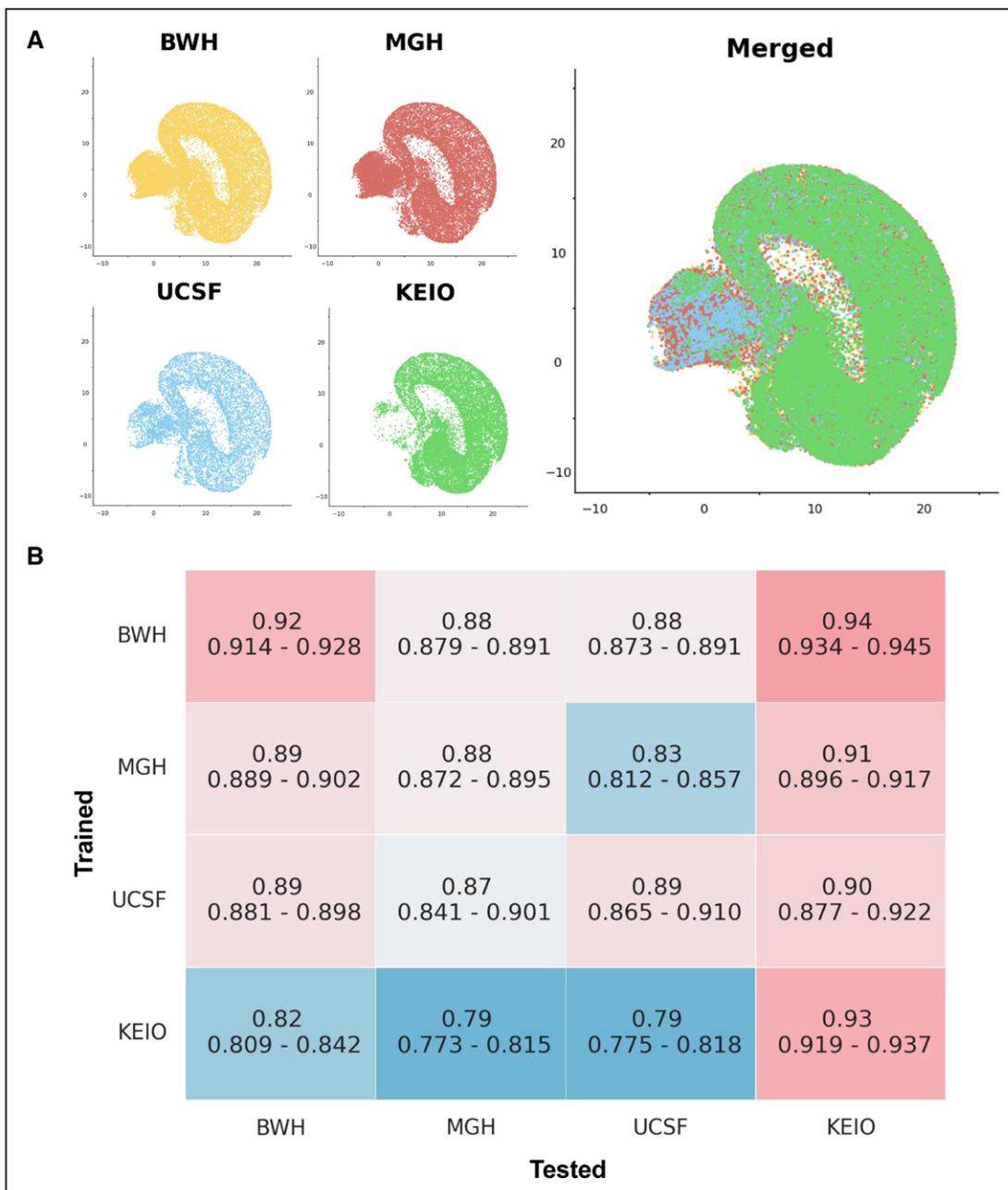
Figure 2. Heterogeneity of ECG and AUROCs of models trained at individual institutions.
A, Uniform Manifold Approximation and Projection (UMAP) projection of raw electrocardiographic recording stratified by institution. **B**, Heat map showing the performance of models trained at individual institutions. Held-out test data sets were used to evaluate model performance. The models were trained using 5-fold cross-validation for each institution, and all models were tested on their own institution test set along with 3 external data sets. The area under the receiver-operating characteristics curve (AUCROC) and 95% CI based on the SE for the 5 models are shown. BWH indicates Brigham and Women's Hospital; Keio, Keio University Hospital; MGH, Massachusetts General Hospital; and UCSF, University of California San Francisco.

discrimination was comparable using only outpatient echocardiograms (AUROC, 0.96) or limiting the controls to cases without LVH (AUROC, 0.97; Figure S21). Unlike the ECG model, the echocardiogram model could discriminate HCM from hypertension or AS without a significant drop in AUROC (0.93 and 0.94 for hypertension and AS, respectively; Figure 4D; Figures S22–S24 and Tables S9–S11 show patient selection). For cases with CA, the discrimination between HCM was slightly lower (AUROC, 0.85). As with ECG, the model
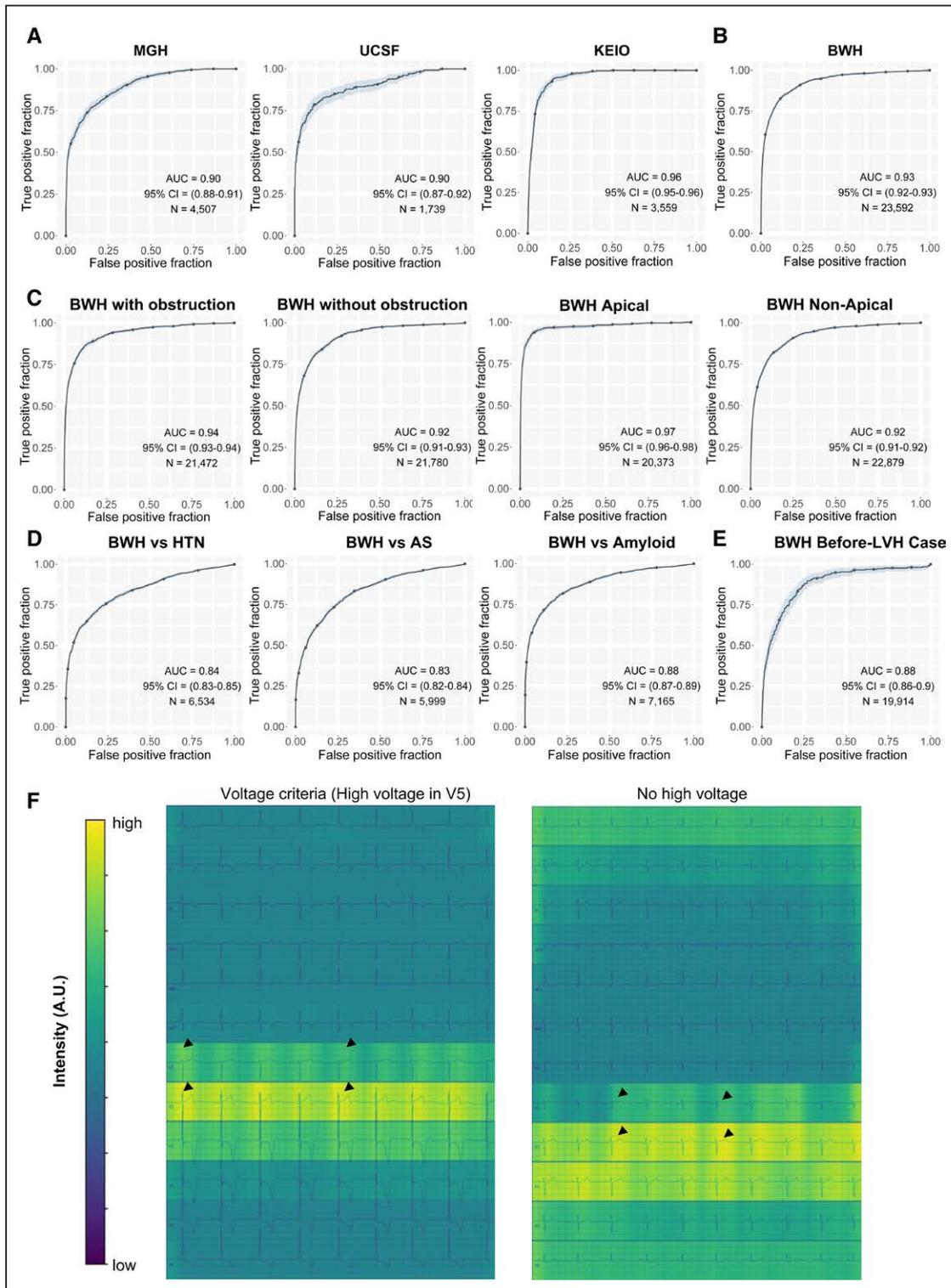
Figure 3. Discrimination of HCM by the ECG model trained with federated learning.
A, Receiver-operating characteristics (ROC) plots for hypertrophic cardiomyopathy (HCM) discrimination of the ECG model trained in a federated manner on a held-out internal test data set for each institution (3569, 1375, and 2656 patients in test data set from Massachusetts General Hospital [MGH], University of California San Francisco [UCSF], and Keio University Hospital [Keio], respectively) and on an (**B**) external data set (18 118 patients from Brigham and Women's Hospital [BWH]). **C**, ROC curves for discriminating HCM with and without outflow tract obstruction (17 830 and 17 769 patients, respectively) and apical and nonapical HCM (17 767 and 18 022 patients, respectively). **D**, ROC curves for discriminating HCM with hypertension, aortic valve stenosis (AS), or cardiac amyloidosis (1020, 746, and 811 patients, respectively). **E**, An ROC curve for discrimination of HCM before developing HCM (17 760 patients). The 95% CI of the true-positive fraction for a given false-positive fraction is shown as a blue ribbon (N is the number of studies). **F**, Gradient-weighted class activation mapping images for HCM samples with and without high voltage. Areas of primary focus of the model are indicated by black arrowheads. AUC indicates area under the curve; HTN, hypertension; and LVH, left ventricular hypertrophy.
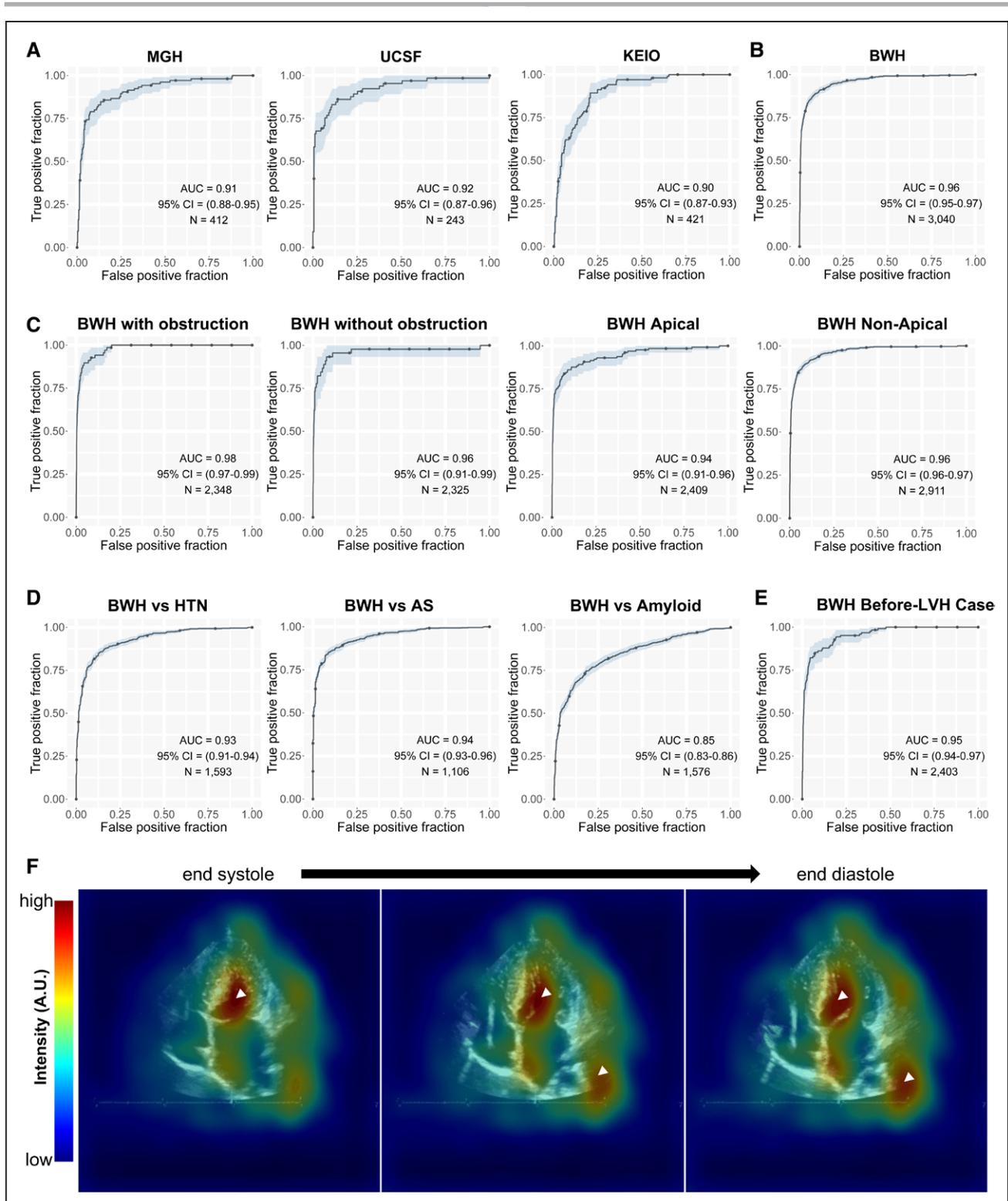
Figure 4. Discrimination of HCM by the echocardiogram model trained with federated learning.
A, Receiver-operating characteristics (ROC) plots for detecting hypertrophic cardiomyopathy (HCM) using echocardiogram model trained in a federated manner tested on held-out internal data for each institution (1700, 1031, and 1639 patients in test data set from Massachusetts General Hospital [MGH], University of California San Francisco [UCSF], and Keio University Hospital [Keio], respectively) and on an (**B**) external data set (2455 patients from Brigham and Women's Hospital [BWH]). **C**, ROC plots for detecting HCM with and without outflow tract obstruction (2253 and 2205 patients, respectively) and apical and nonapical HCM (2192 and 2391 patients, respectively). **D**, ROC curves for discriminating HCM with hypertension (HTN), aortic valve stenosis (AS), cardiac amyloidosis (1491, 611, and 640 patients, respectively). **E**, ROC curve for discrimination of HCM before developing HCM (2403 patients). The 95% CI of the true-positive fraction for a given false-positive fraction is shown as a blue ribbon (N is the number of studies). **F**, Gradient-weighted class activation mapping images for HCM sample. Areas of primary focus of the model are indicated by white arrowheads. AUC indicates area under the curve; and LVH, left ventricular hypertrophy.

Table 1. Sensitivities and Specificities for HCM Detection at Various Cutoffs

Model	Target sensitivity	Cutoff	Combined internal test cohort			
			Sensitivity	Specificity	PPV	NPV
ECG	0.90	5.0	0.92 (0.91–0.94)	0.67 (0.66–0.68)	0.30 (0.30–0.31)	0.98 (0.98–0.99)
	0.85	10.0	0.86 (0.84–0.87)	0.80 (0.79–0.81)	0.41 (0.39–0.42)	0.97 (0.97–0.98)
	0.80	17.5	0.81 (0.79–0.73)	0.87 (0.86–0.88)	0.50 (0.48–0.51)	0.97 (0.97–0.86)
	0.75	25.0	0.76 (0.74–0.79)	0.90 (0.90–0.91)	0.55 (0.53–0.57)	0.96 (0.96–0.96)
	0.70	40.0	0.70 (0.67–0.72)	0.94 (0.93–0.94)	0.63 (0.61–0.65)	0.95 (0.95–0.96)
Echocardiogram	0.90	2.0	0.91 (0.87–0.94)	0.72 (0.69–0.75)	0.53 (0.50–0.56)	0.96 (0.94–0.97)
	0.85	6.0	0.86 (0.82–0.90)	0.83 (0.81–0.86)	0.64 (0.60–0.68)	0.95 (0.93–0.96)
	0.80	9.0	0.81 (0.77–0.86)	0.86 (0.83–0.88)	0.66 (0.62–0.70)	0.93 (0.91–0.95)
	0.75	18.0	0.75 (0.7–0.81)	0.90 (0.88–0.92)	0.73 (0.68–0.77)	0.92 (0.90–0.93)
	0.70	32.0	0.70 (0.65–0.75)	0.93 (0.91–0.95)	0.77 (0.73–0.82)	0.90 (0.89–0.92)

(*Continued*)

discriminated HCM cases before overt LVH (AUROC, 0.95; Figure 4E; Figure S25 shows patient selection). GRAD-CAM analysis revealed a focus on the left ventricular septum and a region located caudal and posterior to the left atrium and prioritization of end diastole (Figure 4F).

As with electrocardiographic data, the 4 baseline models using readily available echocardiogram measurements and baseline patient characteristics showed only moderate discrimination of HCM (AUROC, 0.81–0.82 on echocardiograms from BWH; Supplemental Methods and Table S12).

Screening Approaches Using the ECG and Echocardiogram Models Detect HCM at Higher Sensitivity Than Cardiologists

After exclusion of patients with invalid ECGs or echocardiograms, 11 823 patients (59 cases with HCM) with ECG-echocardiogram pairs were included in the constructed surveillance cohort (Figure S26 and Table S13). Assuming that the models would be used in screening settings, cutoffs at high sensitivity ranges were analyzed, showing comparable sensitivities/specificities for the internal test sets, external validation set, and the surveillance cohort at various cutoff points (Table 1).

Using ECGs alone, the model achieved much higher sensitivities (98%) compared with all 3 cardiologists (73%–81%) at the same PPV of 1% (Supplemental Results provides details; see also Table 2 and Figure 5A). A decision curve analysis revealed that the net benefit of the ECG ML model surpassed other heuristic strategies (Figure S27).

Although there was considerable variability in the ability of experts to detect HCM from echocardiogram, the sensitivity of the cardiologists for detecting patients with HCM from echocardiogram was low in this cohort (37%–59%). The PPV of the experts was 19% to 24%, with lower sensitivity on MRI-confirmed cases and cases without LVH (sensitivities, 31%–50% and 25%–42% respectively; Table S14). When cutoffs were adjusted to achieve the same or higher PPV as the experts, the echocardiogram model consistently showed higher sensitivities (78% sensitivity at 24% PPV; Table 2). As with the ECG, the observation was evident across the entire PPV range (Figure 5A) and surpassed other approaches (Figure S27).

We further compared the use of the models in a setting where ECGs were used as a screening tool to select patients for echocardiogram. First, we compared the stepwise approach using the ECG model to prescreen patients before the echocardiogram model (ECG-ECHO) versus using an echocardiogram model for all patients (ECHO). The results showed that by preselecting with ECG, a higher sensitivity could be achieved (sensitivity, 75% and 66% for ECG-ECHO and ECHO, respectively, at a PPV of 30%) despite performing smaller numbers of echocardiograms (5277 of 11 823 patients undergoing echocardiogram evaluation; Table 3). In comparison, the best expert would select 5350 patients to perform an echocardiogram based on ECG and would detect HCM with 49% sensitivity at a 28% PPV (Table 3 and Figure 5B). This resulted in improved negative likelihood (negative likelihood ratios, 0.26, 0.34, and 0.51 for ECG-ECHO, ECHO, and stepwise approach by the best expert; for all comparisons with stepwise approach, $P<0.01$).

DISCUSSION

We describe here a multimodality approach to automate detection of HCM enabled by training on multi-institutional data using federated learning with low-cost inputs that can be gathered in a primary care setting. The 2-stage screening strategy using the models showed a much higher sensitivity compared with cardiologists.

Table 1. Continued

BWH matched cohort				Surveillance cohort			
Sensitivity, %	Specificity, %	PPV	NPV	Sensitivity	Specificity	PPV	NPV
0.95 (0.94–0.96)	0.63 (0.63–0.74)	0.34 (0.34–0.35)	0.99 (0.98–0.99)	0.98 (0.95–1.00)	0.56 (0.55–0.57)	0.01 (0.01–0.01)	1.00 (1.00–1.00)
0.90 (0.89–0.91)	0.78 (0.78–0.79)	0.45 (0.45–0.46)	0.98 (0.97–0.98)	0.97 (0.92–1.00)	0.72 (0.71–0.72)	0.02 (0.02–0.02)	1.00 (1.00–1.00)
0.85 (0.84–0.86)	0.86 (0.86–0.87)	0.55 (0.54–0.56)	0.97 (0.96–0.97)	0.97 (0.92–1.00)	0.82 (0.81–0.82)	0.03 (0.02–0.03)	1.00 (1.00–1.00)
0.81 (0.79–0.82)	0.90 (0.90–0.90)	0.62 (0.61–0.63)	0.96 (0.96–0.96)	0.97 (0.92–1.00)	0.87 (0.86–0.87)	0.03 (0.03–0.04)	1.00 (1.00–1.00)
0.74 (0.72–0.75)	0.94 (0.93–0.94)	0.69 (0.68–0.71)	0.95 (0.94–0.95)	0.90 (0.81–0.97)	0.91 (0.91–0.92)	0.05 (0.04–0.05)	1.00 (1.00–1.00)
0.93 (0.92–0.95)	0.83 (0.81–0.85)	0.93 (0.92–0.95)	0.98 (0.97–0.98)	0.98 (0.95–1.00)	0.80 (0.79–0.80)	0.02 (0.02–0.02)	1.00 (1.00–1.00)
0.89 (0.87–0.91)	0.90 (0.89–0.91)	0.75 (0.72–0.77)	0.96 (0.95–0.97)	0.97 (0.92–1.00)	0.88 (0.88–0.89)	0.04 (0.04–0.04)	1.00 (1.00–1.00)
0.88 (0.86–0.90)	0.92 (0.91–0.93)	0.79 (0.76–0.81)	0.96 (0.95–0.97)	0.97 (0.92–1.00)	0.90 (0.90–0.91)	0.05 (0.04–0.05)	1.00 (1.00–1.00)
0.84 (0.82–0.87)	0.95 (0.94–0.96)	0.85 (0.82–0.87)	0.95 (0.94–0.96)	0.97 (0.92–1.00)	0.94 (0.93–0.94)	0.07 (0.07–0.08)	1.00 (1.00–1.00)
0.80 (0.77–0.83)	0.96 (0.95–0.97)	0.88 (0.85–0.90)	0.94 (0.93–0.94)	0.93 (0.86–0.98)	0.96 (0.95–0.96)	0.10 (0.09–0.11)	1.00 (1.00–1.00)

Results are shown as value (95% CI). Cutoffs are selected to achieve the highest specificity for a sensitivity at or higher than the target sensitivity. Cutoff scores are presented as 10 for the electrocardiogram model and 10 for the echocardiogram model. Combined internal test cohort: combined test set of Massachusetts General Hospital, University of California San Francisco, and Keio University Hospital. Brigham and Women's Hospital: matched cohort: external validation cohort constructed by age and sex matching for Brigham and Women's Hospital. HCM indicates hypertrophic cardiomyopathy; NPV, negative predictive value; and PPV, positive predictive value.

We previously developed a convolutional neural network–guided approach (combined with gradient boosting) to detect HCM by ECG,[14] and others have subsequently published an end-to-end solution toward the same end.[16] We and others have also published a 2D convolutional neural network model using a frame-by-frame approach for echocardiographic data.[17,25] These studies were all based on data from a single center and used a single modality. We show here that models trained on data from a single institution may not perform as expected on external data even when the performance on the held-out test data set is excellent. We have also demonstrated that the discrimination of HCM and generalizability of the model can be substantively improved by using a federated learning approach across multiple institutions without the need for data sharing. Although ECG is generally considered a well-standardized modality, our results suggest that there is still heterogeneity across institutions. The heterogeneity of the ECG could be attributable to (1) a difference in population factors or (2) a difference in technical factors such as vendor signal processing approaches. Our analysis visualizing patient characteristics on UMAP projections or subgroup analysis of models from individual institutions did not find any patient factors that explain the heterogeneity.

Although federated learning eliminates the need for centralizing the data, the model still needs to be transferred during training, and aspects of the original data can be partially extracted from neural network models by reverse engineering.[26,27] However, the federated learning approach still has significant merits. First, as described, reverse engineering is required to extract data from the model. In our scenario, the models are shared between trusted entities. Thus, if an agreement is made not to reverse engineer the model, the data are not visible to the recipient. Second, the amount of information transferred and stored in the central server is greatly different. For example, the current analysis required 10 to 100 terabytes of data from each institution, whereas the models are gigabytes in size. Last, encryption techniques can further protect the original data, albeit at the expense of model performance.[28]

Downstream workflows will inform how to implement the models in clinical practice. In the case of HCM, this would include gathering of family information and, if required, confirmatory cardiac MRI, genetic testing, or biopsy if the echocardiogram is not clearly diagnostic. A PPV for screening can be considered a pretest probability of the next test or action; a more costly or invasive next action often requires a higher pretest probability. The strength of using ML model is that the choice of cutoff point can be tuned to adapt to situation. Our models displayed consistently higher sensitivity compared with cardiologists across a wide range of PPVs, suggesting that a substantial number of HCM cases that would otherwise be missed by cardiologists could be detected by the models at a given pretest probability requirement.

We have considered how to incorporate data from both ECG and echocardiogram to perform screening for HCM effectively. We believe the best way to apply such disease detection models in clinical settings is through a systematic screening with ECG followed by a more informative evaluation by echocardiography. Our data suggest that performing electrocardiographic screening with the ML model improves sensitivity at the same PPV compared with a strategy of performing an echocardiogram on all patients and reduces the number of echocardiograms performed. Because echocardiogram provides information beyond an HCM diagnosis, the modality should certainly still be used if otherwise indicated. However, in a

Table 2. HCM Discrimination by the Models Compared With Expert Cardiologists

| Modality | Experts | | | | | | Model | | | |
	By	Criteria	Sensitivity	Specificity	PPV		Cutoff	Sensitivity	Specificity	PPV
ECG	R.Y.	Any abnormality	0.73	0.55	0.01		5.0	0.98 (0.95–1.00)	0.56 (0.55–0.57)	0.01 (0.01–0.01)
		LVH (voltage criteria)	0.17	0.99	0.06		52.5	0.86 (0.78–0.95)	0.93 (0.93–0.94)	0.06 (0.05–0.07)
		Suspected HCM	0.08	1.00	0.13		155.0	0.63 (0.49–0.75)	0.98 (0.98–0.98)	0.13 (0.10–0.16)
	S.G.	Any abnormality	0.73	0.55	0.01		5.0	0.98 (0.95–1.00)	0.56 (0.55–0.57)	0.01 (0.01–0.01)
		LVH (voltage criteria)	0.31	0.99	0.10		100.0	0.75 (0.63–0.96)	0.97 (0.96–0.97)	0.10 (0.09–0.12)
		Suspected HCM	0.15	1.00	0.20		242.5	0.54 (0.41–0.68)	0.99 (0.99–0.99)	0.20 (0.15–0.24)
	J.E.J.	Any abnormality	0.81	0.55	0.01		5.0	0.98 (0.95–1.00)	0.98 (0.98–0.98)	0.01 (0.01–0.01)
		LVH (voltage criteria)	0.47	0.99	0.14		167.5	0.63 (0.49–0.75)	0.98 (0.98–0.98)	0.14 (0.11–0.17)
		Suspected HCM	0.17	1.00	0.22		297.5	0.51 (0.37–0.64)	0.99 (0.99–0.99)	0.22 (0.17–0.28)
Echocardiogram	R.Y.	Suspected HCM	0.37	0.99	0.19		63.0	0.85 (0.75–0.93)	0.98 (0.98–0.98)	0.19 (0.16–0.22)
	S.G.	Suspected HCM	0.56	0.99	0.22		71.0	0.82 (0.71–0.90)	0.99 (0.98–0.99)	0.22 (0.19–0.26)
	J.E.J.	Suspected HCM	0.59	0.99	0.24		75.0	0.78 (0.68–0.88)	0.99 (0.99–0.99)	0.25 (0.21–0.29)

Results are shown as value (95% CI). Cutoffs scores are presented as 10 for the electrocardiogram model and 10 for the echocardiogram model. Combined internal test cohort: combined test set of Massachusetts General Hospital, University of California San Francisco, and Keio University Hospital. HCM indicates hypertrophic cardiomyopathy; and PPV, positive predictive value.

resource-limited setting, an ECG-based screening strategy may be of value. Furthermore, although the criteria for HCM diagnosis are based on imaging, there may also be electrocardiography features that help distinguish HCM from other forms of hypertrophy. For example, it appears that apical HCM is detected better by the ECG than by the echocardiogram model. In this way, one may consider ECG and echocardiography as complementary rather than elevating one modality over the other.

One interesting finding in our analysis on expert reading of echocardiogram is the relatively low sensitivity for detecting HCM, which, unsurprisingly, was more apparent in patients with HCM without overt LVH or those confirmed by MRI, presumably because the latter included more challenging cases. In contrast, the ECG and echocardiogram model trained with the federated learning approach robustly discriminated HCM in these cases. The results, along with the good discrimination between other causes of LVH (hypertension, AS, and CA) and before overt LVH, suggest that the model could aid detection of cases that are hard to detect without cardiac MRI.

From a pragmatic standpoint, ML solutions are most helpful if they increase the number of patients likely to benefit from detection, a function of available clinical workflows for prevention and treatment. Models that are trained on highly exaggerated patient phenotypes such as those that may be found in national or international referral centers are likely to demonstrate the best statistical model performance (especially on internal data sets) but are unlikely to be of greatest utility for most cases found in the community because they may detect disease in patients at late stages such as those requiring complex procedures like myectomy. In the case of inherited disorders such as HCM, including studies from more minimally affected relatives may help, although again

there may be bias toward more penetrant mutations and the resulting phenotypic patterns. The fact that our models are effective at multiple institutions in different countries is encouraging, although all are tertiary academic centers, so they may be enriched in some more extreme phenotypic manifestations.

There are some limitations to the study. First, because our disease detection approach involved convolutional neural networks, the features used by the model remain obscure. We attempted to partly address this by using GRAD-CAM. However, GRAD-CAM provides only "where" the model is weighting and does not provide information on "what" the feature is. Furthermore, the GRAD-CAM outputs differ from sample to sample, resulting in some ambiguity. Second, although the AUROCs of the models trained at individual institutions were calculated with 5-fold cross-validation, it was calculated on the test data set in a federated learning model. Thus, the comparison is not direct. However, because the test data set was constructed by random split, it is not unreasonable to assume that the AUROC on the test data set is representative. Third, because HCM is an underdiagnosed disease, there was a possibility that undiagnosed cases of HCM have been falsely included in the controls. However, given that HCM is relatively rare, we believe that the impact of this misclassification was small. Fourth, Keio University Hospital measured the interventricular septum thickness excluding the right ventricular wall; thus, the measurement could not be compared directly with other institutions. Fifth, although the surveillance cohort was constructed to mimic the prevalence of HCM in the general population, it was randomly constructed from a tertiary care center–based cohort and may differ from a truly unselected population. Sixth, because patients after myectomies or myocardial ablation or those after
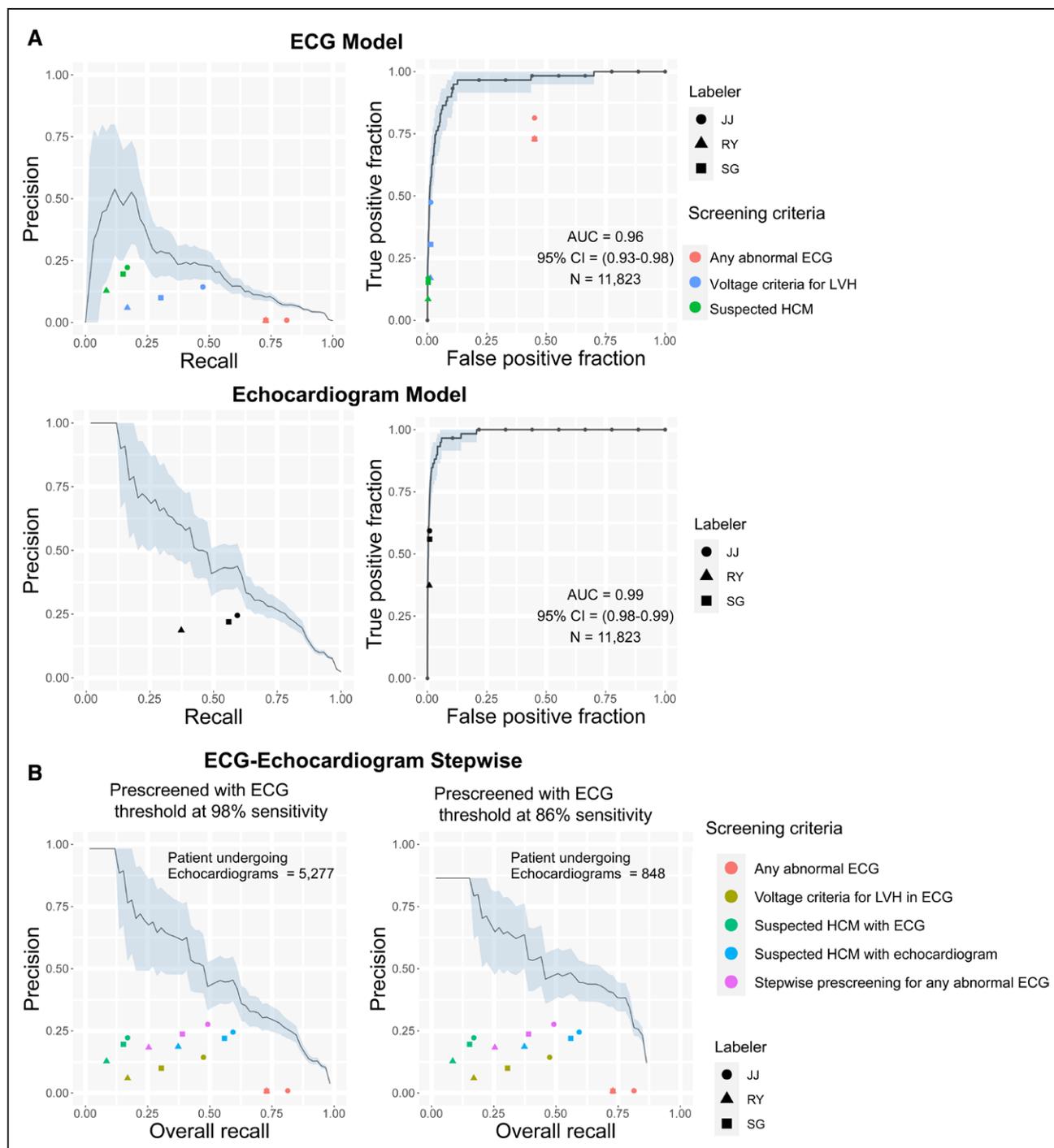
Figure 5. Deployment simulation of the models on surveillance populations for detecting HCM.
A, Precision recall curve (PRC) and receiver-operating characteristics (ROC) plots for the ECG model and echocardiography model for discrimination of patients with hypertrophic cardiomyopathy (HCM) in the surveillance populations. The 95% CIs of precision and true-positive fraction are shown as blue ribbons in the PRC and ROC curves, respectively (N is the number of patients). The sensitivity, specificity, and positive predictive value (PPV) for detecting patients with HCM by human experts are plotted with the curves. **B**, PRC curves for the stepwise approach applying echocardiogram model after prescreening with ECG model using 2 cutoffs corresponding to the PPV of the any abnormal ECG findings by human experts. The overall sensitivity, specificity, and PPV for detecting patients with HCM by human experts are plotted. Overall recall is the number of HCMs detected after all the processes divided by the total number of HCM cases in the original cohort. AUC indicates area under the curve; and LVH, left ventricular hypertrophy.

implantation of an implantable cardioverter defibrillator or a pacemaker were excluded from the analysis, the model may not discriminate HCM cases with these conditions.

However, because these cases are usually already diagnosed and evaluated, the influence on the model utility is minimal.

Table 3. Comparison of Discrimination Performance for HCM in Deployment of the Echocardiogram Model Alone or With Stepwise Approach Against Expert Readings

PPV requirement	Model	No. undergoing echocardiogram	PPV	Overall recall
0.10	Echocardiogram alone	11 823	0.10 (0.09–0.11)	0.90 (0.81–0.97)
	Stepwise (electrocardiogram model cutoff=5.0)	5277	0.10 (0.09–0.11)	0.97 (0.93–1.00)
	Stepwise (electrocardiogram model cutoff=52.5)	848	0.10(0.09–0.10)	0.86(0.86–0.86)
0.15	Echocardiogram alone	11 823	0.15 (0.13–0.17)	0.86 (0.76–0.95)
	Stepwise (electrocardiogram model cutoff=5.0)	5277	0.15 (0.13–0.17)	0.88 (0.80–0.95)
	Stepwise (electrocardiogram model cutoff=52.5)	848	0.15 (0.13–0.16)	0.85 (0.81–0.86)
0.20	Echocardiogram alone	11 823	0.20 (0.18–0.23)	0.83 (0.73–0.92)
	Stepwise (electrocardiogram model cutoff=5.0)	5277	0.20 (0.17–0.23)	0.85 (0.76–0.93)
	Stepwise (electrocardiogram model cutoff=52.5)	848	0.20 (0.18–0.22)	0.85 (0.81–0.86)
0.25	Echocardiogram alone	11 823	0.25 (0.21–0.30)	0.76 (0.64–0.86)
	Stepwise (electrocardiogram model cutoff=5.0)	5277	0.25 (0.22–0.29)	0.84 (0.73–0.92)
	Stepwise (electrocardiogram model cutoff=52.5)	848	0.25 (0.22–0.27)	0.85 (0.81–0.86)
0.30	Echocardiogram alone	11 823	0.30 (0.24–0.36)	0.66 (0.54–0.78)
	Stepwise (electrocardiogram model cutoff=5.0)	5277	0.30 (0.25–0.35)	0.75 (0.64–0.85)
	Stepwise (electrocardiogram model cutoff=52.5)	848	0.30 (0.26–0.34)	0.81 (0.76–0.86)
Experts	R.Y. (echocardiogram alone)	11 823	0.19	0.37
	R.Y. (preselected with any abnormal ECG)	5345	0.18	0.25
	S.G. (echocardiogram alone)	11 823	0.22	0.56
	S.G. (preselected with any abnormal ECG)	5344	0.24	0.39
	J.E.J. (echocardiogram alone)	11 823	0.24	0.59
	J.E.J. (preselected with any abnormal ECG)	5350	0.28	0.49

Results are shown as value (95% CI). Overall recall is the number of HCMs detected after all of the processes divided by the total number of HCM cases in the original cohort. Cutoffs scores are presented as 10 for the electrocardiogram model. HCM indicates hypertrophic cardiomyopathy; and PPV, positive predictive value.

Conclusions

We have developed models, using federated learning strategies, that detect HCM from ECGs and echo-cardiograms across multiple institutions. In addition, we have shown that a screening strategy using both models could potentially improve screening of patients with HCM.

REFERENCES

1. Marian AJ, Braunwald E. Hypertrophic cardiomyopathy. *Circ Res*. 2017;121:749–770. doi: 10.1161/CIRCRESAHA.117.311059
2. Seidman JG, Seidman C. The genetic basis for cardiomyopathy from mutation identification to mechanistic paradigms. *Cell*. 2001;104:557–567. doi: 10.1016/s0092-8674(01)00242-2
3. Maron BJ. Clinical course and management of hypertrophic cardiomyopathy. *N Engl J Med*. 2018;379:655–668. doi: 10.1056/NEJMra1710575
4. Ommen SR, Mital S, Burke MA, Day SM, Deswal A, Elliott P, Evanovich LL, Hung J, Joglar JA, Kantor P, et al. 2020 AHA/ACC guideline for the diagnosis and treatment of patients with hypertrophic cardiomyopathy. *Circulation*. 2020;142:e558–e631. doi:10.1161/CIR.0000000000000937
5. Iacopo O, Artur O, Roberto B-V, Theodore PA, Ahmad M, Pablo G-P, Sara S, Neal KL, Matthew TW, Anjali O, et al. Mavacamten for treatment of symptomatic obstructive hypertrophic cardiomyopathy (EXPLORER-HCM): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet*. 2020;396:759–769. doi: 10.1016/S0140-6736(20)31792-X

6. Ho CY, Mealiffe ME, Bach RG, Bhattacharya M, Choudhury L, Edelberg JM, Hegde SM, Jacoby D, Lakdawala NK, Lester SJ, et al. Evaluation of mavacamten in symptomatic patients with nonobstructive hypertrophic cardiomyopathy. *J Am Coll Cardiol*. 2020;75:2649–2660. doi: 10.1016/j.jacc.2020.03.064

7. Maron BJ, Desai MY, Nishimura RA, Spirito P, Rakowski H, Towbin JA, Rowin EJ, Maron MS, Sherrid MV. Diagnosis and evaluation of hypertrophic cardiomyopathy: JACC state-of-the-art review. *J Am Coll Cardiol*. 2022;79:372–389. doi: 10.1016/j.jacc.2021.12.002

8. Hughes SE. The pathology of hypertrophic cardiomyopathy. *Histopathology*. 2004;44:412–427. doi: 10.1111/j.1365-2559.2004.01835.x

9. López-Cuenca D, Muñoz-Esparza C, Peñalver MN, Alberola AG, Blanes JRG. Hypertrophic or hypertensive cardiomyopathy? *Int J Cardiol*. 2016;203:891–892.

10. Maron MS, Rowin EJ, Maron BJ. How to image hypertrophic cardiomyopathy. *Circ Cardiovasc Imaging*. 2017;10:e005372. doi: 10.1161/CIRCIMAGING.116.005372

11. Lemery R, Kleinebenne A, Nihoyannopoulos P, Aber V, Alfonso F, McKenna WJ. Q waves in hypertrophic cardiomyopathy in relation to the distribution and severity of right and left ventricular hypertrophy. *J Am Coll Cardiol*. 1990;16:368–374. doi: 10.1016/0735-1097(90)90587-f

12. Maron BJ, Wolfson JK, Ciró E, Spirito P. Relation of electrocardiographic abnormalities and patterns of left ventricular hypertrophy identified by 2-dimensional echocardiography in patients with hypertrophic cardiomyopathy. *Am J Cardiol*. 1983;51:189–194. doi: 10.1016/s0002-9149(83)80034-4

13. Usui M, Inoue H, Suzuki J, Watanabe F, Sugimoto T, Nishikawa J. Relationship between distribution of hypertrophy and electrocardiographic changes in hypertrophic cardiomyopathy. *Am Heart J*. 1993;126:177–183. doi: 10.1016/s0002-8703(07)80026-3

14. Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005289. doi: 10.1161/CIRCOUTCOMES.118.005289

15. Morita SX, Kusunose K, Haga A, Sata M, Hasegawa K, Raita Y, Reilly MP, Fifer MA, Maurer MS, Shimada YJ. Deep learning analysis of echocardiographic images to predict positive genotype in patients with hypertrophic cardiomyopathy. *Front Cardiovasc Med*. 2021;8:669860. doi: 10.3389/fcvm.2021.669860

16. Ko W-Y, Siontis KC, Attia ZI, Carter RE, Kapa S, Ommen SR, Demuth SJ, Ackerman MJ, Gersh BJ, Arruda-Olson AM, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol*. 2020;75:722–733. doi: 10.1016/j.jacc.2019.12.030

17. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation*. 2018;138:1623–1635. doi: 10.1161/CIRCULATIONAHA.118.034338

18. McMahan HB, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. *arXiv*. Preprint posted online February 17, 2016. doi: 10.48550/arXiv.1602.05629

19. Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, Baust M, Cheng Y, Ourselin S, Cardoso MJ, et al. Machine Learning in Medical Imaging, 10th international workshop, MLMI 2019, held in conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings. *Lect Notes Comput Sci*. 2019:133–141.

20. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, Liu A, Costa AB, Wood BJ, Tsai C-S, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27:1735–1743. doi: 10.1038/s41591-021-01506-3

21. Ho CY, Day SM, Colan SD, Russell MW, Towbin JA, Sherrid MV, Canter CE, Jefferies JL, Murphy AM, Cirino AL, et al. The burden of early phenotypes and the influence of wall thickness in hypertrophic cardiomyopathy mutation carriers: findings from the HCMNet study. *JAMA Cardiol*. 2017;2:419. doi: 10.1001/jamacardio.2016.5670

22. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv*. Preprint posted online December 6, 2018. doi: 10.48550/arXiv.1802.03426

23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *arXiv*. Preprint posted online December 3, 2019. doi: 10.48550/arXiv.1610.02391

24. Goto S, Mahara K, Beussink-Nelson L, Ikura H, Katsumata Y, Endo J, Gaggin HK, Shah SJ, Itabashi Y, MacRae CA, et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nat Commun*. 2021;12:2726. doi: 10.1038/s41467-021-22877-8

25. Morita SX, Kusunose K, Haga A, Sata M, Hasegawa K, Raita Y, Reilly MP, Fifer MA, Maurer MS, Shimada YJ. Deep learning analysis of echocardiographic images to predict positive genotype in patients with hypertrophic cardiomyopathy. *Front Cardiovasc Med*. 2021;8:669860. doi: 10.3389/fcvm.2021.669860

26. Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown TB, Song DX, Erlingsson Ú, Oprea A, Raffel C. Extracting Training Data from Large Language Models. *arXiv*. Preprint posted online December 14, 2020. doi: 10.48550/arXiv.2012.07805

27. Ray I, Li N, Kruegel C, Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. *Proc 22nd Acm Sigsac Conf Comput Commun Secur*. 2015:1322–1333.

28. Ma C, Poor HV. On safeguarding privacy and security in the framework of federated learning. *IEEE Network*. 2020;34:242–248. doi: 10.1109/MNET.001.1900506

ORIGINAL RESEARCH ARTICLE