Circulation

ORIGINAL RESEARCH ARTICLE

rECHOmmend: An ECG-Based Machine Learning Approach for Identifying Patients at Increased Risk of Undiagnosed Structural Heart Disease Detectable by Echocardiography

Alvaro E. Ulloa-Cerna, PhD; Linyuan Jing, PhD; John M. Pfeifer, MD, MPH; Sushravya Raghunath, PhD; Jeffrey A. Ruhl, MS; Daniel B. Rocha, MM; Joseph B. Leader, BA; Noah Zimmerman, PhD; Greg Lee, BS; Steven R. Steinhubl, MD; Christopher W. Good, DO; Christopher M. Haggerty, PhD; Brandon K. Fornwalt, MD; Ruijun Chen, MD

BACKGROUND: Timely diagnosis of structural heart disease improves patient outcomes, yet many remain underdiagnosed. While population screening with echocardiography is impractical, ECG-based prediction models can help target high-risk patients. We developed a novel ECG-based machine learning approach to predict multiple structural heart conditions, hypothesizing that a composite model would yield higher prevalence and positive predictive values to facilitate meaningful recommendations for echocardiography.

METHODS: Using 2 232 130 ECGs linked to electronic health records and echocardiography reports from 484 765 adults between 1984 to 2021, we trained machine learning models to predict the presence or absence of any of 7 echocardiography-confirmed diseases within 1 year. This composite label included the following: moderate or severe valvular disease (aortic/mitral stenosis or regurgitation, tricuspid regurgitation), reduced ejection fraction <50%, or interventricular septal thickness >15 mm. We tested various combinations of input features (demographics, laboratory values, structured ECG data, ECG traces) and evaluated model performance using 5-fold cross-validation, multisite validation trained on 1 site and tested on 10 independent sites, and simulated retrospective deployment trained on pre-2010 data and deployed in 2010.

RESULTS: Our composite rECHOmmend model used age, sex, and ECG traces and had a 0.91 area under the receiver operating characteristic curve and a 42% positive predictive value at 90% sensitivity, with a composite label prevalence of 17.9%. Individual disease models had area under the receiver operating characteristic curves from 0.86 to 0.93 and lower positive predictive values from 1% to 31%. Area under the receiver operating characteristic curves for models using different input features ranged from 0.80 to 0.93, increasing with additional features. Multisite validation showed similar results to cross-validation, with an aggregate area under the receiver operating characteristic curve of 0.91 across our independent test set of 10 clinical sites after training on a separate site. Our simulated retrospective deployment showed that for ECGs acquired in patients without preexisting structural heart disease in the year 2010, 11% were classified as high risk and 41% (4.5% of total patients) developed true echocardiography-confirmed disease within 1 year.

CONCLUSIONS: An ECG-based machine learning model using a composite end point can identify a high-risk population for having undiagnosed, clinically significant structural heart disease while outperforming single-disease models and improving practical utility with higher positive predictive values. This approach can facilitate targeted screening with echocardiography to improve underdiagnosis of structural heart disease.

Key Words: cardiomyopathies ■ echocardiography ■ electrocardiography ■ heart valve diseases ■ machine learning ■ ventricular dysfunction

Correspondence to: Ruijun Chen, MD, 100 North Academy Avenue, Danville, PA 17822. Email ruijun.chen@gmail.com

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCULATIONAHA.122.057869.

Continuing medical education (CME) credit is available for this article. Go to http://cme.ahajournals.org to take the quiz.

For Sources of Funding and Disclosures, see page 46.

Circulation is available at www.ahajournals.org/journal/circ

Clinical Perspective

What Is New?

- An ECG-based machine learning model can identify those at higher risk for previously undiagnosed structural heart disease with excellent performance.
- Using a composite end point of multiple structural heart disease end points aligned in the clinical response of diagnostic echocardiography markedly increases positive predictive value of detecting structural abnormality and improves clinical actionability.

What Are the Clinical Implications?

- This composite end point model overcomes the limitations of low prevalence and positive predictive value in previous single-disease machine learning models to improve performance and generate meaningful recommendations for echocardiography.
- Such predictions can help close the large diagnostic gap of structural heart disease for millions of patients with undetected yet clinically significant disease.
- Reducing the number of undiagnosed patients may allow more patients to appropriately receive evidence-based care to improve patient outcomes and prevent adverse events.

Nonstandard Abbreviations and Acronyms

AR	aortic regurgitation
AS	aortic stenosis
AUPRC	area under the precision-recall curve
AUROC	area under the receiver operating characteristic curve
CNN	convolutional neural network
EF	ejection fraction
IVS	interventricular septum
MS	mitral stenosis
PPV	positive predictive value
TR	tricuspid regurgitation

In patients with structural heart disease carrying a high burden of morbidity and mortality, echocardiography provides important evidence-based implications for diagnosis, prognosis, and management.[1–5] Echocardiography is the primary diagnostic test for many structural conditions, including valvular disease, left ventricular dysfunction, and various cardiomyopathies[6–8]; however, despite the growth of evidence-based therapies for patients with structural heart disease and the increasing availability of echocardiography, these conditions continue to be underdiagnosed.[6,9–12] Studies have shown that millions of patients have unrecognized disease, while many others are diagnosed only after the occurrence of adverse events or irreversible sequelae of undiagnosed disease.[11–15]

ECG-based machine learning models can help identify undiagnosed patients for targeted screening, yet limitations to their practical adoption remain. ECGs are more common, inexpensive, and broadly indicated than echocardiograms, and machine learning approaches using ECGs have been shown to identify patients at increased risk of individual diseases.[16–18] However, despite otherwise good performance, these models often suffer from low positive predictive values (PPVs) because of the low prevalence of individual target diseases.[18,19] This limits the practical utility of real-world implementations, given that many patients identified as high risk would need to undergo screening to diagnose one true case.

We therefore aimed to combine multiple disease outcomes into a single composite prediction to increase diagnostic yield. We developed a novel machine learning approach to identify patients at high risk for any of 7 structural heart disease end points within a single ECG platform, including moderate or severe valvular disease (aortic stenosis [AS]; aortic regurgitation; mitral stenosis [MS]; mitral regurgitation; and tricuspid regurgitation [TR]), reduced left ventricular ejection fraction (EF), and increased interventricular septal (IVS) thickness. We hypothesized that our model would generate a composite prediction with higher yield/PPV to facilitate a practical clinical recommendation for diagnostic echocardiography. Moreover, we simulated the utility of this model on a large retrospective dataset to assess expected real-world performance if implemented into clinical care.

METHODS

The data that support the findings of this study are available from the corresponding author on reasonable request.

Data

The institutional review board approved this study with a waiver of consent. We retrieved and processed data from 3 clinical sources at Geisinger, a large regional US health system providing both inpatient and outpatient care, including 2 110 332 patients from the Epic (Epic Systems, Madison, Wisconsin) electronic health record, 758 269 echocardiograms from Xcelera (Philips, Cambridge, Massachusetts), and 3 548 543 ECGs from MUSE (GE Healthcare). We included all 12-lead ECGs after 1984 from patients ≥18 years old, sampled at either 250 Hz or 500 Hz, and a corresponding Epic medical record, resulting in 2 925 925 ECGs from 631 710 patients. All data were collected through July 2021.

We obtained vitals, laboratory results, and patient demographics as of the index ECG acquisition date and time (Table S1). We used the closest past measurement unless the measurement was >1 year old, in which case we assigned a missing value. We extracted echocardiographic measurements and

diagnoses from Xcelera reports and ECG structured findings, measurements, and 12-lead traces from MUSE.[16,20] Structured ECG findings were directly obtained from the final, official interpretation by an attending cardiologist. We then labeled ECGs as detailed below. Overall, we included 2 232 130 ECGs with at least 1 label from 484 765 patients (Figure 1).

Echocardiography-Confirmed Disease Outcome Definitions

We defined 7 outcome labels using echocardiography reports: one for each disease outcome (AS, aortic regurgitation, mitral regurgitation, MS, TR, reduced EF, increased IVS thickness). We used regular expressions to extract key words and phrases identifying the diagnosis of valvular stenosis or regurgitation and its associated severity level, based on the final interpretation by an attending cardiologist (Table S2). We labeled each of the valvular conditions of interest as positive if moderate or severe and negative if normal or mild in severity. We assigned a missing label otherwise. Mild-to-moderate cases were labeled as moderate.

We defined positive labels for reduced EF as a reported EF of <50% on echocardiography. We defined increased IVS thickness as >15 mm. These criteria were chosen based on cardiologist and clinician consensus and in concordance with existing guidelines for potential diseases of interest, such as hypertrophic cardiomyopathy.[21] Echocardiograms not meeting those criteria were labeled as negative. We assigned a missing label if the measurement was missing.

Outcome labels extracted from echocardiography reports for AS, aortic regurgitation, mitral regurgitation, MS, and TR were randomly sampled in sets of 100 to 200 and validated by manual chart review.

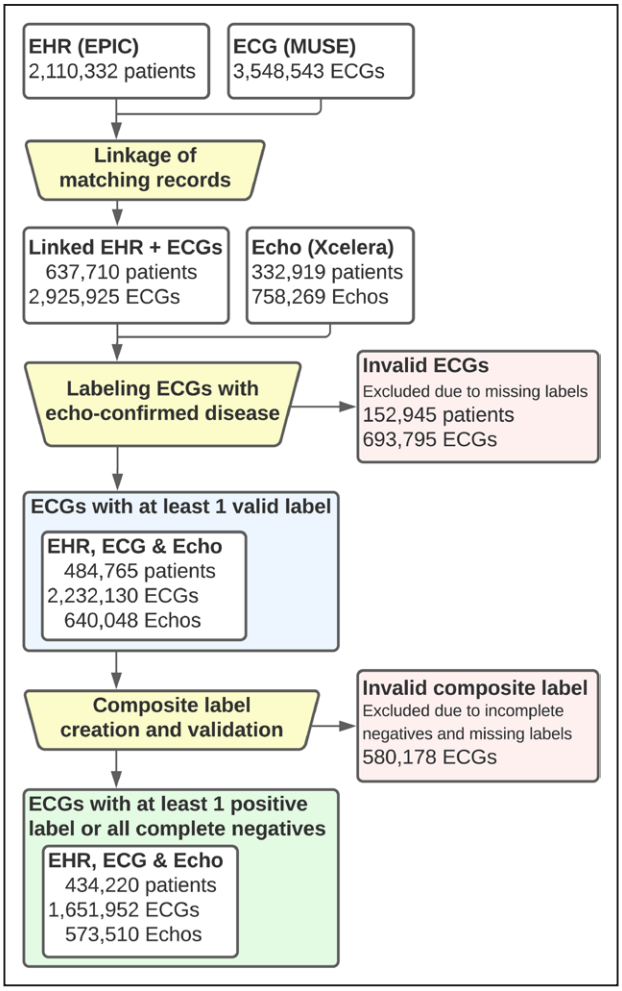

Figure 1. Flow diagram from source data to the study datasets.

We processed data from research repositories created using EHR data from Epic, ECG data from MUSE, and echocardiography data from Xcelera. The clinical MUSE database was processed to include 12-lead ECGs sampled at either 250 Hz or 500 Hz, acquired after 1984 from patients >18 years. ECG indicates electrocardiogram; EHR, electronic health record; and Echo, echocardiography.

ECG Labeling

For each given disease outcome, an ECG was labeled as positive if it was acquired within one year before the patient's first positive echocardiography report for that disease, or any time after the echocardiogram until a censoring event (Figure S1). Censoring events were defined as death, end of observation in the electronic health record, or any intervention that directly treated the disease and could modify the underlying physiology, such as valve replacement or repair. We also used a negative echocardiography report after a positive echocardiography report as a censoring event to conservatively eliminate the possibility that such interventions may have been performed at outside institutions and therefore not represented in our data.

For each given disease outcome, an ECG could be labeled as negative using 2 sets of criteria, depending on whether the patient did or did not have a history of previous echocardiography. For patients with history of echocardiography, ECGs acquired more than 1 year before the last negative echocardiogram with confirmed absence of that given disease were labeled as negative (Figure S1). In the absence of any patient history of echocardiography, an ECG was also labeled as negative if there was at least 1 year of subsequent follow-up without a censoring event and without any coded diagnoses for the relevant disease (Table S3).

For the composite end point, we labeled an ECG as positive if any of the 7 individual outcomes were positive and as negative if all 7 outcomes were negative.

Model Development

We developed 9 models using different combinations of input feature sets from structured data (demographics, vitals, labs, structured ECG findings and measurements) and ECG voltage traces. For ECG trace models, we developed a deep convolutional neural network (CNN) consisting of 6 1-dimensional CNN-Batch Normalization-ReLU layer blocks, followed by a multilayer perceptron and a final logistic output layer (Table S4).[22] The CNN used raw ECG trace data sampled at 500 Hz as input. ECGs sampled at 250 Hz were resampled to 500 Hz using linear interpolation. We used the same configuration to train 1 model per clinical outcome, resulting in 7 independently trained CNN models (Figure 2). We chose a minimalistic CNN architecture design as we sought to focus on the novel effect of the composite end point rather than the exploration or development of novel machine learning architectures.

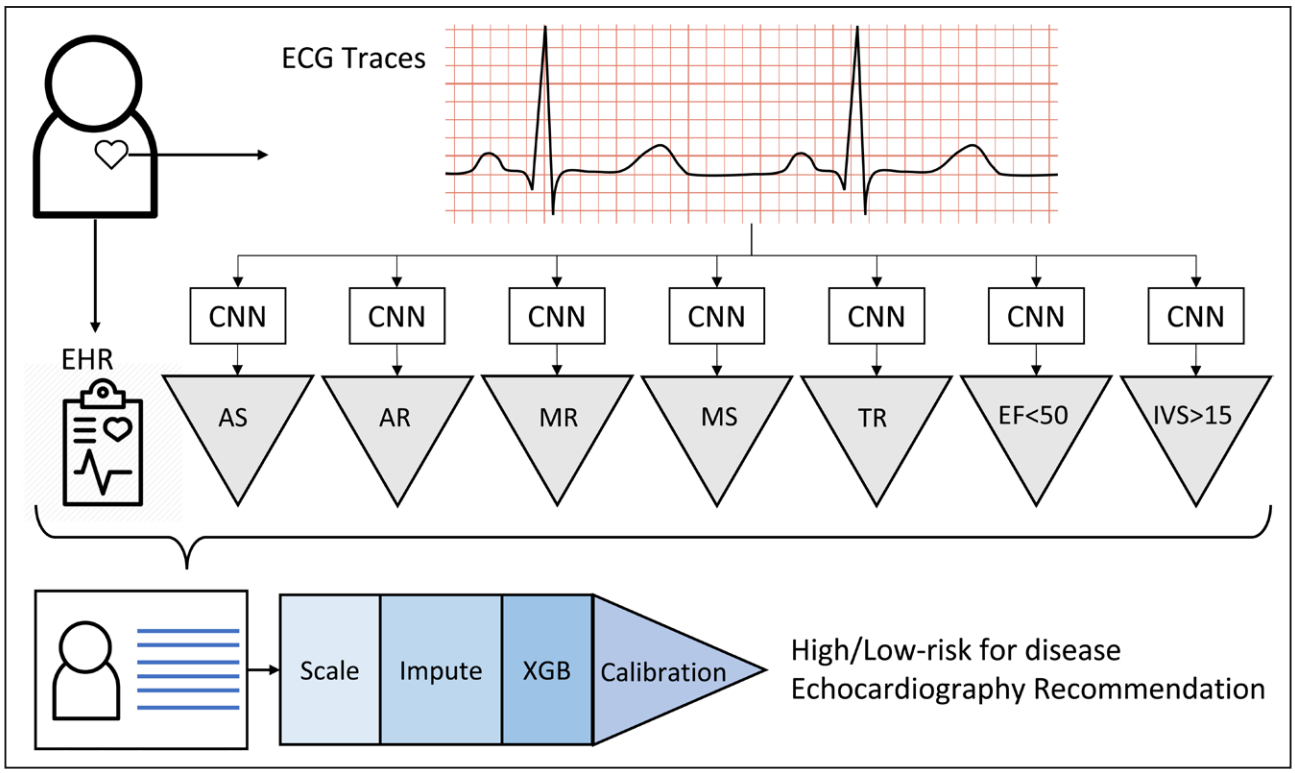

Figure 2. rECHOmmend model diagram showing the classification pipeline for ECG traces and other electronic health record data. The output (*gray triangle*) of each CNN applied to ECG trace data are concatenated with labs, vitals, and demographics to form a feature vector. This vector is the input to the classification pipeline (min–max scaling, mean imputation, XGBoost classifier, and calibration), which outputs a composite prediction for the patient. AR indicates aortic regurgitation; AS, aortic stenosis; CNN, convolutional neural network; ECG, electrocardiogram; EF, ejection fraction; EHR, electronic health record; IVS, interventricular septum; MR, mitral regurgitation; MS, mitral stenosis; and TR, tricuspid regurgitation.

To form the final model and combine ECG trace–based models with structured data, we concatenated the risk scores from the individual CNNs with the structured data. We used the concatenated feature vector to train a classification pipeline consisting of a minutes-max scaler (min 0, max 1), mean imputation, XGBoost classifier, and calibration (Figure 2).[23,24]

Model Evaluation

We evaluated the models using 3 approaches: (1) a traditional random cross-validation partition; (2) a multisite validation where the model was trained on data from Geisinger Medical Center and tested on 10 other independent clinical sites; and (3) a retrospective deployment scenario where, using 2010 as the simulated deployment year, we used past data to train and future data to test. We measured area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and other performance metrics (sensitivity, specificity, PPV and negative predictive value) at multiple operating points. For all experiments, data were split into training, internal validation, and test sets with no overlap of patients across these sets.

Cross Validation

We conducted a 5-fold cross validation by randomly sampling 5 mutually exclusive sets of patients. We expanded each set to all ECGs from each patient to form the training and test sets. When training the CNN models for each individual end point, we discarded samples with missing labels. We applied the model to all test samples and evaluated performance only on samples with complete labels that also satisfied the rECHOmmend labeling criteria, described above. Performance statistics were reported as means and 95% CIs across 5 folds for a random ECG per patient.

Multisite Validation

To perform multisite validation, we created mutually exclusive sets of patients from 11 clinical sites in the Geisinger Health System. We assigned each patient to a particular site by selecting the most common ECG site of origin for that given patient. We removed any ECGs taken outside of the assigned site for each patient.

We trained our model on data from patients at a single site—Geisinger Medical Center, a large quaternary teaching hospital in Danville, Pennsylvania. We then tested this model on 10 other independent clinical sites, ranging from outpatient centers to small community hospitals to large teaching hospitals, at various locations across Pennsylvania.

Retrospective Deployment

We retrospectively simulated a deployment of our model using a cutoff date of January 1, 2010, relabeling all ECGs with information available as of that date. We used this artificially constrained dataset to replicate the cross-validation experiments and train a deployment model using data prior to 2010. We then

applied the deployment model to the first ECG per patient for all patients seen from January 1, 2010 through December 31, 2010. We calibrated the XGBoost model using earliest ECGs from the at-risk population in 2005 and measured performance statistics on all patients at risk in 2010. We determined the true outcomes of the at-risk population using information up to July 23, 2021, adhering to the definitions for positive and negative outcome labels outlined previously in this article. We report the results for 2 thresholds for comparison: (1) a high sensitivity threshold (90%), which may be more ideal for a screening use case; and (2) a 50% sensitivity threshold, identifying half of the diseased population with higher PPVs, which may be more practical and actionable.

Sensitivity Analyses

To account for potential variation in what providers and patients may find to be clinically significant disease, we repeated the cross-validation experiment on a different set of labels representing severe disease only. These label definitions include severe valvular disease only (moderate valvular disease now considered a negative label) and changed the definition for reduced EF to be <35%.

To account for the possibility of patients with persistently undiagnosed disease in our definition of negative ECGs, we also repeated our cross-validation experiment using only echocardiography-confirmed negatives. Patients who never received an echocardiogram were excluded. All ECGs labeled negative were followed by a negative echocardiogram confirming the absence of that given disease outcome.

Both analyses were performed using models trained on new label definitions and an independent set of cross-validation folds.

RESULTS

We identified 758 269 echocardiography reports from 332 919 patients, of which 191 652 echocardiograms from 88 093 patients were positive for at least 1 disease outcome label. Disease prevalence ranged from 0.6% for MS to 17.2% for reduced EF (Table S5). We identified 2 232 130 ECGs from 484 765 patients who met criteria for ≥1 positive or negative individual disease label; of those, 1 651 952 ECGs from 434 220 patients qualified for the composite label (Table S6). Baseline across 2.23 million ECGs was as follows: median patient age, 64 years old; 50.1% male; and 97.1% White (Table 1). Patients with positive labels—compared with negative labels—were generally older and comprised a greater proportion of males and smokers. Baseline characteristics among patients with missing or undefined labels as compared to patients with ≥1 defined label were largely similar (Table S7).

Model Input Feature Evaluation

Table 2 shows the results of 5-fold cross-validation comparing model performance as a function of different input features. AUROCs ranged from 0.80 for the model using only age and sex to 0.93 for the model with all available inputs, including structured ECG findings and measurements, demographics, labs, vitals, and ECG traces (Figure 3). While the model with all available inputs provided the best performance, we focus the remainder of our results in this article on the models including only age, sex, and ECG traces since this input set best balances portability, objectivity, and performance with a 0.91 AUROC. These inputs are all directly available from MUSE or other ECG systems, without additional integration with other data sources, and do not require waiting for the official cardiologist interpretation, which may be subject to interrater variability. Complete, detailed results including all other input sets for every disease label across all folds and various subgroups are available at: http://www.rechommend.com.

Cross-Validation Performance of rECHOmmend Model

The rECHOmmend model using age, sex, and ECG traces for prediction of the composite disease label yielded a 0.91 AUROC (95% CI, 0.90–0.91) and 42% PPV at 90% sensitivity with 18% disease prevalence (Table 3). As hypothesized, the composite model yielded a higher PPV than any of the 7 models trained for an individual component end point, which ranged from 1% for MS to 31% for reduced EF (Table 3). We found the same trend for the AUPRC (0.71 [95% CI, 0.71–0.72]) for the rECHOmmend model, as compared with individual model AUPRCs, which ranged from 0.04 to 0.65 (Figure S2). Performance metrics for a wide range of alternate model operating points, including those with higher sensitivity or higher PPV, are presented in Table S8. Performance of the rECHOmmend model across various subgroups of age, sex, race, comorbidities, and ECG findings were largely comparable to the overall results in terms of AUROC except for slightly decreased AUROCs in patients with heart failure and patients with pacemakers (Table S9).

Multisite Validation Performance

The rECHOmmend model trained on Geisinger Medical Center and validated across 10 other clinical sites performed similarly well to our cross-validation experiment, yielding an AUROC of 0.91 in aggregate across all other sites (Table S10). Individual site AUROCs ranged from 0.79 at the Viewmont Imaging Center to 0.93 at the Scranton Community Medical Center, with 8 out of 10 sites having AUROCs >0.85 and 7 sites having AUROCs >0.90. The prevalence of the composite label for disease among sites varied from 1% at Viewmont to 39% at the Geisinger Commonwealth School of Medicine. Correspondingly, PPV varied from 15% at Viewmont to 54% at the Geisinger Commonwealth School of Medicine.

Table 1. Baseline Characteristics and Features at Time of Index ECG

Demographics and vitals	Mean±SD	Median [interquartile range]
Age, y	63±17	64 [52–76]
Body mass index, kg/m^2	31±9	30 [25–35]
Systolic blood pressure, mm Hg	129±20	128 [116–140]
Diastolic blood pressure, mm Hg	73±12	72 [64–80]
Heart rate, beats/min	76±15	74 [66–84]
Height, cm	168±11	168 [160–178]
Weight, kg	88±24	85 [70–101]
White*	97.1	
Male sex*	50.1	
Ever smoked*	59.7	
Laboratory values		
HbA1C, %	6.9±3	6.5 [5.8–7.5]
Bilirubin, mg/dL	0.57±0.60	0.5 [0.3–0.7]
BUN, mg/dL	20.5±12.8	17 [13–23]
Cholesterol, mg/dL	172±47	168 [140–200]
Creatine kinase-MB, ng/mL	8.9±32.2	2.9 [1.9–5]
Creatinine, mg/dL	1.2±1.4	0.9 [0.8–1.2]
C-reactive protein, mg/L	36.2±63.9	9 [2.6–38]
D-dimer, µg/mL	1.5±2.6	0.6 [0.3–1.5]
Glucose, mg/dL	119±48	104 [93–125]
High-density lipoprotein, mg/dL	48±16	45 [37–56]
Hemoglobin, g/dL	14±34	13 [11.7–14.3]
Lactate dehydrogenase, U/L	249±237	207 [171–264]
Low-density lipoprotein, mg/dL	95±38	91 [68,117]
Lymphocytes, %	23±11	22 [15–29]
Potassium, mmol/L	4.2±0.7	4.2 [3.9–4.5]
Pro-BNP, pg/mL	5002±10668	1369 [341–4377]
Sodium, mmol/L	139±3	140 [137–141]
Troponin I, ng/mL	1±13	0.03 [0.01–0.06]
Troponin T, ng/mL	0.16±0.84	0.01 [0.01–0.04]
Triglyceride, mg/dL	154±122	127 [90–183]
Uric acid, mg/dL	6.6±2.4	6.3 [4.9–7.9]
Very-low-density lipoprotein, mg/dL	29±16	25 [18–36]
eGFR, mL/(min·1.73 m^2)	54±12	60 [55–60]
Other comorbidities		
Heart failure*	17.2	
Previous myocardial infarction*	18.8	
Diabetes *	23.1	
Chronic obstructive pulmonary disease*	14.0	
Renal failure*	8.3	
Previous echocardiogram*	28.4	
Coronary artery disease*	23.1	
Hypertension*	46.4	
ECG findings and measurements		
R axis	22±50	21 [−10 to 54]
PR interval†	164±40	160 [144–182]
P axis	48±30	50 [33–64]
QRS duration	98±25	90 [82–104]
QT	400±51	398 [368–430]
QTC	445±4	440 [418–464]
T axis	52±53	46 [23–71]
Ventricular rate	77±20	74 [63–87]
Average RR interval	821±194	814 [688–946]
Normal*	43.8	
Previous infarct*	18.7	
Nonspecific T-wave changes*	16.0	
Sinus bradycardia*	14.1	
Nonspecific ST changes*	10.3	
Ischemia*	10.0	
Left axis deviation*	9.3	
Atrial fibrillation*	8.5	
Left ventricular hypertrophy*	8.0	
Tachycardia*	7.5	
Previous anterior myocardial infarction*	7.3	
Premature ventricular contractions*	6.8	
First-degree block*	6.3	
Right bundle-branch block*	6.0	
Prolonged QT*	5.0	
Poor tracing*	4.9	
Premature atrial contractions*	4.8	
Pacemaker*	4.6	
T-wave inversion*	4.6	
Low QRS voltage*	4.4	
Fascicular block*	3.2	
Incomplete right bundle-branch block*	3.1	
Left bundle-branch block*	2.8	
Intraventricular block*	2.3	
Right axis deviation*	2.2	
Atrial flutter*	1.3	
Acute myocardial infarction*	1.0	
Incomplete left bundle-branch block*	0.4	
Supraventricular tachycardia*	0.4	
Early repolarization*	0.3	
Complete heart block*	0.1	
Other bradycardia*	0.1	
Second-degree atrioventricular block*	0.1	
Ventricular tachycardia*	0.1	

Data reported as mean±SD and median [interquartile range] for continuous values, unless otherwise indicated. eGFR indicates estimated glomerular filtration rate; and proBNP, pro B-type natriuretic peptide.

*Data reported as percentages (%) for categorical values.

†Thirty-one extreme outlier values were removed for PR interval >2000 ms.

Table 2. Performance Comparison of Cross-Validated Models Across Various Input Features for Composite End Point

Input features	Area under receiver operating curve	Area under precision-recall curve	Positive predictive value*	Negative predictive value*	Specificity*
Age + sex	0.799 (0.795–0.802)	0.468 (0.462–0.473)	27.5 (27.0–28.0)	95.7 (95.6–95.7)	48.2 (47.5–49.0)
Demographics, labs, and vitals (structured EHR)	0.862 (0.860–0.865)	0.651 (0.644–0.657)	32.3 (31.8–32.8)	96.4 (96.4–96.5	58.9 (58.3–59.5)
ECG structured findings and measurements (structured ECG)	0.879 (0.877–0.881)	0.677 (0.672–0.683)	34.0 (33.4–34.5)	96.6 (96.6–96.6)	61.8 (61.0–62.6)
ECG traces	0.904 (0.902–0.906)	0.719 (0.714–0.724)	41.1 (40.4–41.9)	97.1 (97.1–97.1)	71.9 (71.3–72.6)
Available from ECG system					
Age + sex + ECG traces	0.907 (0.905–0.908)	0.714 (0.707–0.722)	42.0 (41.4–42.6)	97.1 (97.1–97.1)	72.9 (72.4–73.4)
Structured ECG + ECG traces	0.912 (0.910–0.913)	0.739 (0.733–0.744)	42.9 (42.0–43.8)	97.1 (97.1–97.1)	73.9 (73.2–74.6)
Available from ECG + EHR					
Age + sex + structured EHR + structured ECG	0.917 (0.915–0.919)	0.762 (0.757–0.767)	44.2 (43.5–44.9)	97.2 (97.2–97.2)	75.2 (74.6–75.8)
Age + sex + structured EHR + ECG traces	0.925 (0.923–0.926)	0.780 (0.775–0.784)	46.7 (46.0–47.4)	97.3 (97.2–97.3)	77.6 (77.0–78.2)
Age + sex + structured EHR + structured ECG + ECG traces (all inputs)	0.928 (0.927–0.930)	0.787 (0.783–0.792)	47.8 (47.2–48.4)	97.3 (97.3–97.3)	78.6 (78.2–79.0)

All values are shown as percentage (95% CI). Each model was tested based on a random ECG per patient. End points include valvular disease, reduced ejection fraction, and increased interventricular septal thickness. ECG indicates electrocardiogram; and EHR, electronic health record.

*All values are at 90% sensitivity.

Simulated Deployment Performance

We identified 692 273 ECGs with qualifying labels for any of the 7 clinical outcomes before 2010, of which 485 469 ECGs qualified for the composite label to train the deployment model. A cross-validation experiment for this pre-2010 subset showed similar, yet slightly reduced performance as compared with the full dataset (AUROC, 0.89; 31% PPV at 90% sensitivity; Table S11).

The 2010 deployment test set contained ECGs from 69 544 patients (Figure 4A). After excluding patients with a known history of disease, we identified 63 459 at-risk patients between January 1 and December 31, 2010. Of these patients, outcome labels for 20 395 were undefined because of inadequate follow-up or not meeting criteria for the composite label. As previously noted, the characteristics of patients with undefined labels were similar to those with defined labels. The AUROC among patients with defined labels was 0.86. Using a threshold estimated to yield 90% sensitivity based on the pre-2010 training data, the deployment model labeled 43.3% of patients as high risk and obtained a PPV of 15.1% and a negative predictive value of 98.5%.

For a more practical comparison, using a threshold estimated to yield 50% sensitivity, the deployment model labeled 10.7% of patients as high risk for any of the 7 disease outcomes. Among 2969 predicted high-risk patients with adequate follow-up who met our definition for the composite label, 1219 patients were diagnosed with ≥1 disease outcome within a year, with a PPV of 41.1%. Of these 1219 patients, 137 (11%) received a diagnosis of AS, 86 (7%) were diagnosed with aortic regurgitation, 387 (32%) with mitral regurgitation, 17 (1%) with MS, 375 (31%) with TR, 785 (64%) with reduced EF, and 280 (23%) with IVS thickening. Among 40 095 predicted low-risk patients with adequate follow-up and defined labels, 38 552 patients did not develop any of the outcomes within a year, with a negative predictive value of 96.2%.

Overall, at this model threshold, for every 100 at-risk patients who obtained an ECG, our model would identify 11 as high risk, of which 5 would truly have echocardiography-confirmed disease, and 89 as low risk, of which 86 would truly not have disease within 1 year (Figure 4B).

Sensitivity Analyses

When using severe-only disease labels, AUROCs across input feature combinations for the composite end point were similar to the primary results (Table 2), ranging from 0.79 for age and sex only to 0.94 for all inputs (Table S12). AUPRC and PPV at 90% sensitivity were lower given the lower prevalence of severe-only disease. Across the individual diseases, AUROCs of the age, sex, and ECG traces model were again similar, ranging from 0.84 to 0.96, and again with lower AUPRC and PPV attributable to the lower prevalence (Table S13). The overall rECHOmmend model using severe-only disease labels attained an AUROC of 0.92 with a PPV of 31.2% at 90% sensitivity with 10.6% disease prevalence.

When using echocardiography-confirmed labels only, AUROC was slightly lower than our primary results, while AUPRC and PPV at 90% sensitivity was higher (Tables S14–15). This was likely because of the artificially higher prevalence, as the number of negative patients decreased with this requirement for echocardiography-confirmed absence of disease. The overall rECHOmmend model

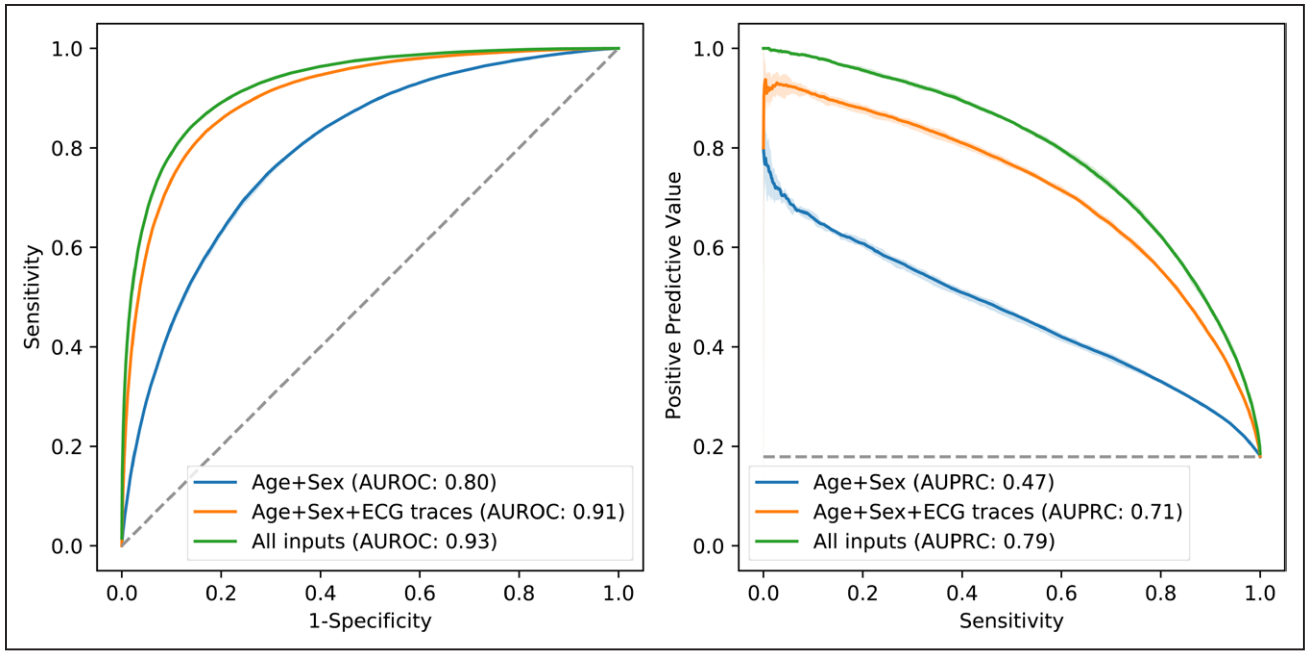

Figure 3. Performance of the rECHOmmend model in cross-validation experiments across various inputs.
The plot on the left shows the AUROC while the plot on the right shows the AUPRC. AUPRC indicates area under the precision-recall curve; and AUROC, area under the receiver operating curve.

obtained an AUROC of 0.88 with a 74% PPV at 90% sensitivity with 53% disease prevalence.

Because of the potential for lead-induced TR in patients with pacemakers, we excluded TR from the composite end point and showed similar performance for our composite model (Table S19).

DISCUSSION

We developed a machine learning platform called "rECHOmmend" that can predict clinically significant valvular disease, reduced left ventricular EF, or pathologically increased septal thickness with excellent performance (AUROC, 0.91) by using only ECG traces, age, and sex. Furthermore, we demonstrated that the combination of these distinct end points into a single platform tied to a recommendation for a singular, practical clinical response—follow-up echocardiography—resulted in an overall PPV of 42% for clinically meaningful disease while maintaining high sensitivity (90%) and specificity (73%). This suggests that for the millions of patients who receive an ECG each year without preexisting structural heart disease, nearly half of those deemed high risk by this model would be found to have true disease within a year. We confirmed the validity of this approach through multisite validation on nonoverlapping data sets from multiple clinical sites across the Geisinger system. Moreover, we confirmed the clinical utility of this approach in our retrospective deployment, as our model trained on pre-2010 data and deployed on all patients without preexisting disease who obtained an ECG in 2010 maintained similarly high performance as compared to the main cross-validation results based only on passive observation and standard clinical care. With an active deployment of the rECHOmmend platform, even higher yields/PPV are anticipated once clinicians can pursue active intervention in the form of follow-up echocardiogram or more detailed history-taking and physical examination.

Clinically, this model enables targeted echocardiographic screening to help detect unrecognized and underdiagnosed diseases. Currently echocardiography is not used for population screening because of the low prevalence of disease in the general population—previous attempts were shown to be largely ineffective.[25,26] Therefore, indicated use of echocardiography is typically triggered by a symptom, adverse event, physical examination, or incidental finding leading to suspicion of heart disease, raising the pretest probability and likelihood of a clinically impactful or actionable finding.[6,7,27] However, a significant gap remains: in meeting the triggered indication for suspected disease, a large number of patients will have already suffered an adverse event, a symptom affecting their quality of life, or an irreversible pathophysiologic change from their undiagnosed disease. For example, in severe AS, the initial presenting symptom is reduced EF for 8% of patients, angina for 35 to 41%, and syncope for 10 to 11% of patients, which may lead to falls, hip fractures, or reduced functional status.[13–15] Previous studies have also shown that up to one-half of elderly patients have undiagnosed valvular disease, including 11.3% with moderate or severe disease, while the majority of patients with hypertrophic cardiomyopathy may be undiagnosed and nearly 50% of patients with EF <40% are asymptomatic.[11,12,28] The rECHOmmend

Table 3. Age + Sex + ECG Traces Model Results for Cross-Validation Experiments for Each Individual Disease Outcome and Composite rECHOmmend Model

Disease	Prevalence	Area under receiver operating curve	Area under precision-recall curve	Positive predictive value*	Negative predictive value*	Specificity*
Aortic stenosis	2.4 (2.3–2.5)	0.908 (0.900–0.915)	0.221 (0.204– 0.239)	8.4 (7.7–9.1)	99.7 (99.7–99.7)	75.7 (73.6–77.7)
Aortic regurgitation	1.8 (1.8–1.9)	0.849 (0.844–0.855)	0.120 (0.114– 0.127)	3.9 (3.6–4.2)	99.7 (99.7–99.7)	58.9 (57.2–60.7)
Mitral regurgitation	4.5 (4.4–4.6)	0.911 (0.908–0.914)	0.367 (0.347– 0.388)	15.2 (14.7–15.7)	99.4 (99.4–99.4)	76.4 (75.8–77.0)
Mitral stenosis	0.3 (0.2–0.3)	0.918 (0.905–0.930)	0.039 (0.036– 0.044)	1.1 (1.0–1.3)	100 (100–100)	79.4 (75.3–82.9)
Tricuspid regurgitation	4.7 (4.6–4.9)	0.915 (0.909–0.920)	0.415 (0.393– 0.438)	16.1 (14.7–17.7)	99.4 (99.3–99.4)	76.9 (74.7–78.9)
Ejection fraction <50%	9.2 (9.1–9.2)	0.929 (0.926–0.931)	0.647 (0.633– 0.662)	31.4 (30.2–32.7)	98.8 (98.7–98.8)	80.2 (79.1–81.2)
Interventricular septum >15 mm	4.0 (3.9–4.1)	0.862 (0.856–0.868)	0.223 (0.213– 0.234)	9.4 (8.8–10.1)	99.4 (99.3–99.4)	64.2 (61.7–66.6)
rECHOmmend (composite)	17.9 (17.8–18.0)	0.907 (0.905–0.908)	0.714 (0.707–0.722)	42.0 (41.4–42.6)	97.1 (97.1–97.1)	72.9 (72.4–73.4)

Results are shown based on a random ECG per patient and averaged across 5 folds. All values are shown as percentages (95% CI). ECG indicates electrocardiogram.

*All values are at 90% sensitivity.

model, with both high sensitivity and precision, can enable targeted screening and guide the decision to obtain an echocardiogram even for asymptomatic patients. This may be particularly relevant given the increasing availability and growing indications for interventions showing benefit in targeting earlier disease, such as recent evidence for earlier intervention in asymptomatic severe AS or novel therapeutics for hypertrophic cardiomyopathy and amyloid.[29] Depending on the use case for a practical, real-world deployment of this model, different thresholds may be used to optimize certain metrics, such as higher sensitivity with the tradeoff of lower specificity and PPV if prioritizing a screening scenario to capture as much of the diseased population as possible, or a threshold prioritizing higher PPV if seeking to optimize a limited amount of resources such as the practical limitations of the total volume of echocardiograms which can be obtained. Together, this machine learning approach can help shift the balance to a scenario where echocardiography can be an effective, targeted screening tool to help clinicians diagnose patients at the right time to prevent downstream adverse events, optimize the timing of interventions, and better implement evidence-based monitoring or management.

Our findings also suggest a path toward overcoming some of the existing challenges with clinical implementation of ECG prediction models. This novel approach of combining multiple end points that align under the same recommended clinical action enables the model to leverage the increased prevalence and probability of any one disease state occurring to improve predictive performance for potential clinical implementation. Previous studies have shown that CNN-based ECG prediction models can predict a variety of cardiovascular outcomes including atrial fibrillation, AS, and left ventricular dysfunction with good performance, with AUROCs from 0.80 to 0.93.[16–19,30] However, concerns often exist around real-world implementation of such models because of limitations in precision and recall, concerns regarding the negative effect of false positives, and limited actionability or portability.[31] Our model compares favorably to those in the literature, with a similar or higher AUROC, comparable performance at similar thresholds, and consistently higher precision or PPV (Tables S16–S17), but also results in a clearly actionable recommendation while remaining highly portable. Our featured model results of 0.91 AUROC, 42% PPV, and 90% sensitivity on cross-validation is based on age, sex, and ECG traces alone as inputs, which we believe represents the optimal balance between performance and portability. While the addition of electronic health record data did slightly improve performance, there would be a major tradeoff in decreased portability with the need for electronic health record or clinical data warehouse integration. This model uses data readily available from any ECG system, such as MUSE, and could be easily deployed across most health care systems.

We also found that simulated deployment on large retrospective datasets can shed light on important questions and estimate true clinical impact before the costly implementation of prediction models in practice or clinical trials, where performance may differ from strictly cross-validation performance of the same models.[16,32] In our simulated deployment on ECGs from 2010, 11% of at-risk patients without history of disease were predicted to be high risk; 41% of patients with adequate follow-up were truly diagnosed with disease within the next year, through only standard clinical care and without any clinician-directed behavior change or active intervention that true deployment may elicit. This suggests that this 41% PPV is likely a lower bound for the expected real-world performance of the rECHOmmend model. Our simulated real-world deployment scenario compares favorably with a recent pragmatic trial for predicting reduced EF which identified a real-world PPV of 39% using an EF cutoff of ≤50%, of which

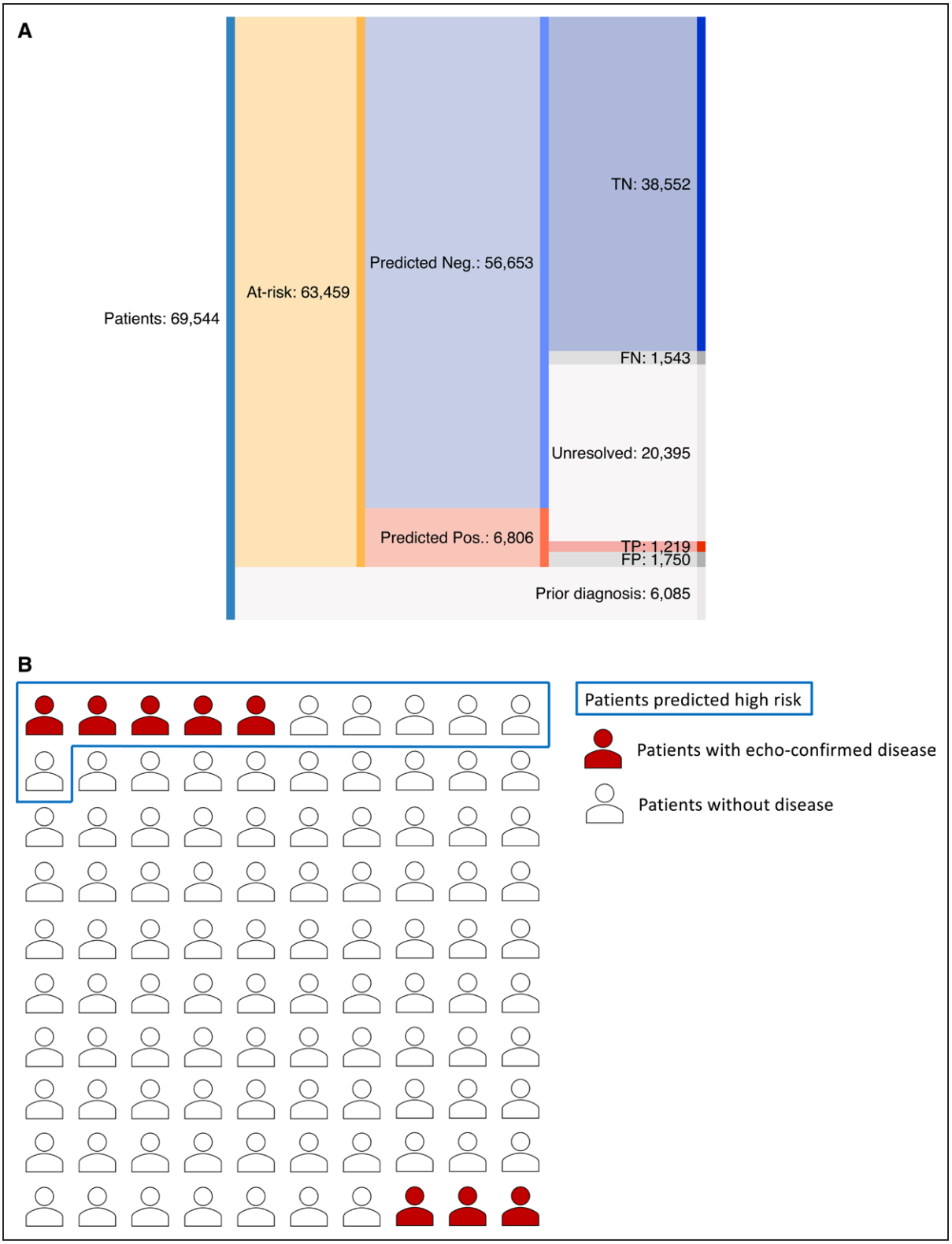

Figure 4. Results of retrospective deployment scenario from 2010.
A, Results for all patients. **B**, Relative results per 100 at-risk patients. These results are based on a threshold yielding 50% sensitivity from the pre-2010 cross-validation experiment, resulting in 41.1% positive predictive value, 96.2% negative predictive value, 95.7% specificity, 44.1% sensitivity, and 6.4% prevalence in 2010. For 100 patients without known history of disease obtaining an ECG, the rECHOmmend model will identify 11 patients at high risk of disease, of which 5 are expected to have true disease within 1 year. The model will identify 89 patients not at high risk of disease, of which 86 are not expected to have true disease within 1 year. ECG indicates electrocardiogram; FN, false negative; FP, false positive; TN, true negative; and TP, true positive.

24% of patients meeting this definition qualified with an EF of exactly 50%.[32] Deployment scenarios also demonstrate that cross-validation metrics that depend on prevalence likely overestimate real-world performance as seen in recent studies, including for the aforementioned reduced EF trial which lagged behind the original cross-validation results (reported PPV of 63%).[18,32] Simulated deployment showed lower prevalence and PPV as compared with cross-validation experiments because of the inclusion of only at-risk patients, those with no previous history of any of the 7 disease labels, as well as the more limited 1-year time window, resulting in essentially a 1-year incidence rate versus a decades-long incidence rate. Cross-validation experiments that restrict the analysis to at-risk patients (ie, with no previous positive diagnosis) may resemble a closer estimate to a real-world deployment. We propose that simulated retrospective deployment be carried out for future prediction models to better gauge feasibility and real-world performance before clinical implementation.

Limitations

Our study has several limitations. Training and evaluation were limited to a regional health system where most patients are White, so results may not be generalizable to hospitals or regions with more diversity. We are not aware of any physiologic differences across race/ethnicity that would lead these ECG-based models to perform differently across groups, as corroborated by previous studies,[33] and subgroup analyses across racial groups in our data showed similar performance (Table S9). In addition, we used echocardiography-confirmed diagnoses to generate our positive labels, which were confirmed on chart review to have a high PPV; however, there may be additional patients with disease (ie, false negatives) who were not captured using this method. Given the low prevalence of each disease in the general population and that echocardiography is the diagnostic standard, the negatives are likely true negatives, as seen in the retrospective deployment where we leveraged up to a decade of follow-up to determine negative outcomes. There may also be potential for misclassification because we are unable to ascertain the exact time at which a patient develops a disease or no longer qualifies for a disease label attributable to physiologic changes such as spontaneous recovery or newly reduced EF. We sought to mitigate this through our methods for labeling positive ECGs, as well as no longer including ECGs after a censoring event such as normal repeat echocardiography or treatment which may alter the physiology of the heart or improve function. Fortunately, despite any potential misclassification, our performance on echocardiography-confirmed disease was excellent. In addition, this machine learning approach has limited interpretability in identifying feature importance. Our model may not perform as well in patients with pacemakers or preexisting diagnoses of heart failure (Table S9); however, these patients are likely to undergo echocardiography as part of standard clinical care and may therefore represent a population for whom this predictive model is neither beneficial nor actionable in a prospective deployment. Last, increased IVS thickness may represent infiltrative diseases or hypertrophic cardiomyopathy, or may largely represent concentric remodeling related to longstanding, poorly controlled hypertension; however, these conditions are all clinically actionable.

This study demonstrates that a machine learning model using only ECG-based inputs can predict multiple important cardiac end points within a single platform with both good performance and high PPV, thereby representing a practical tool with which to better target echocardiography for the detection of undiagnosed disease. We confirmed these results through retrospective real-world deployment scenarios to show the large effect that such a model can have on patients when deployed across a health system. In future work, we plan to execute a prospective study deploying this model across multiple health care systems to better understand the performance, feasibility, and optimal implementation of this approach in the real world. These approaches to both clinical predictions and simulated deployment represent practical solutions for existing limitations in the implementation of machine learning in health care, hopefully bringing this technology one step closer to standard clinical practice.

ARTICLE INFORMATION

Received October 7, 2021; accepted April 5, 2022.

Affiliations

Department of Translational Data Science and Informatics (A.E.U-C., L.J., J.M.P., S.R., J.A.R., C.W.G., C.M.H., B.K.F., R.C.); Phenomic Analytics and Clinical Data Core (D.B.R., J.B.L.); Department of Radiology (B.K.F.); Department of Medicine (R.C.); and Heart Institute (C.M.H., B.K.F.); Geisinger, Danville, PA. Heart and Vascular Center, Evangelical Hospital, Lewisburg, PA (J.M.P.). Tempus Labs Inc, Chicago, IL (J.M.P., S.R., J.B.L., N.Z., G.L., S.R.S., B.K.F., R.C.). Scripps Research Translational Institute, La Jolla, CA (S.R.S.). Heart and Vascular Institute, University of Pennsylvania Medical Center, Hamot, PA (C.W.G.).

Sources of Funding

This work is supported by a grant from Tempus.

Disclosures

Drs Ulloa-Cerna, Jing, Raghunath, and Haggerty, J.A. Ruhl, D.B. Rocha, and J.B. Leader are Geisinger investigators and receive funding from Tempus for ongoing development of predictive modeling technology. Tempus and Geisinger have jointly applied for predictive modeling patents. None of the Geisinger investigators have ownership interest in any of the intellectual property resulting from the partnership. Tempus did not have any input in the design, execution, interpretation of results or decision to publish. Drs Pfeifer, Zimmerman, Raghunath, Chen, and Fornwalt, and G. Lee are Tempus employees. Dr Steinhubl is a consultant for Tempus and an employee of physIQ, and reports personal fees from Otsuka and Janssen outside the submitted work. Dr Fornwalt reports personal fees from Novartis outside the submitted work.

Supplemental Material

Figure S1 and S2
Table S1–S20

REFERENCES

1. Ross J Jr, Braunwald E. Aortic stenosis. *Circulation.* 1968;38(1 Suppl):61–67. doi: 10.1161/01.cir.38.1s5.v-61
2. Cheitlin MD, Gertz EW, Brundage BH, Carlson CJ, Quash JA, Bode RS Jr. Rate of progression of severity of valvular aortic stenosis in the adult. *Am Heart J.* 1979;98:689–700. doi: 10.1016/0002-8703(79)90465-4
3. Davies SW, Gershlick AH, Balcon R. Progression of valvar aortic stenosis: a long-term retrospective study. *Eur Heart J.* 1991;12:10–14. doi: 10.1093/oxfordjournals.eurheartj.a059815
4. Curtis JP, Sokol SI, Wang Y, Rathore SS, Ko DT, Jadbabaie F, Portnay EL, Marshalko SJ, Radford MJ, Krumholz HM. The association of left ventricular ejection fraction, mortality, and cause of death in stable outpatients with heart failure. *J Am Coll Cardiol.* 2003;42:736–742. doi: 10.1016/s0735-1097(03)00789-7
5. Martinez-Naharro A, Baksi AJ, Hawkins PN, Fontana M. Diagnostic imaging of cardiac amyloidosis. *Nat Rev Cardiol.* 2020;17:413–426. doi: 10.1038/s41569-020-0334-7
6. Otto CM, Nishimura RA, Bonow RO, Carabello BA, Erwin JP 3rd, Gentile F, Jneid H, Krieger EV, Mack M, McLeod C, et al. 2020 ACC/AHA guideline for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation.* 2021;143:e72–e227. doi: 10.1161/CIR.0000000000000923
7. Cheitlin MD, Alpert JS, Armstrong WF, Aurigemma GP, Beller GA, Bierman FZ, Davidson TW, Davis JL, Douglas PS, Gillam LD. ACC/AHA guidelines for the clinical application of echocardiography. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Clinical Application of Echocardiography). Developed in collaboration with the American Society of Echocardiography. *Circulation.* 1997;95:1686–1744. doi: 10.1161/01.cir.95.6.1686
8. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, Flachskampf FA, Foster E, Goldstein SA, Kuznetsova T, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur Heart J Cardiovasc Imaging.* 2015;16:233–270. doi: 10.1093/ehjci/jev014
9. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. *Lancet.* 2006;368:1005–1011. doi: 10.1016/S0140-6736(06)69208-8
10. Alexander KM, Orav J, Singh A, Jacob SA, Menon A, Padera RF, Kijewski MF, Liao R, Di Carli MF, Laubach JP, et al. Geographic disparities in reported US amyloidosis mortality from 1979 to 2015: potential underdetection of cardiac amyloidosis. *JAMA Cardiol.* 2018;3:865–870. doi: 10.1001/jamacardio.2018.2093
11. d'Arcy JL, Coffey S, Loudon MA, Kennedy A, Pearson-Stuttard J, Birks J, Frangou E, Farmer AJ, Mant D, Wilson J, et al. Large-scale community echocardiographic screening reveals a major burden of undiagnosed valvular heart disease in older people: the OxVALVE Population Cohort Study. *Eur Heart J.* 2016;37:3515–3522. doi: 10.1093/eurheartj/ehw229
12. Maron MS, Hellawell JL, Lucove JC, Farzaneh-Far R, Olivotto I. Occurrence of clinically diagnosed hypertrophic cardiomyopathy in the United States. *Am J Cardiol.* 2016;117:1651–1654. doi: 10.1016/j.amjcard.2016.02.044
13. Park SJ, Enriquez-Sarano M, Chang SA, Choi JO, Lee SC, Park SW, Kim DK, Jeon ES, Oh JK. Hemodynamic patterns for symptomatic presentations of severe aortic stenosis. *JACC Cardiovasc Imaging.* 2013;6:137–146. doi: 10.1016/j.jcmg.2012.10.013
14. Goliasch G, Kammerlander AA, Nitsche C, Dona C, Schachner L, Öztürk B, Binder C, Duca F, Aschauer S, Laufer G, et al. Syncope: the underestimated threat in severe aortic stenosis. *JACC Cardiovasc Imaging.* 2019;12:225–232. doi: 10.1016/j.jcmg.2018.09.020
15. Selzer A. Changing aspects of the natural history of valvular aortic stenosis. *N Engl J Med.* 1987;317:91–98. doi: 10.1056/NEJM198707093170206
16. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, vanMaanen DP, Hartzel DN, Ruhl JA, Lagerman BF, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke. *Circulation.* 2021;143:1287–1298. doi: 10.1161/CIRCULATIONAHA.120.047829
17. Kwon JM, Lee SY, Jeon KH, Lee Y, Kim KH, Park J, Oh BH, Lee MM. Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. *J Am Heart Assoc.* 2020;9:e014717. doi: 10.1161/JAHA.119.014717
18. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med.* 2019;25:70–74. doi: 10.1038/s41591-018-0240-2
19. Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko WY, Murphree DH, Michelena HI, Enriquez-Sarano M, Carter RE, et al. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J.* 2021;42:2885–2896. doi: 10.1093/eurheartj/ehab153
20. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, Williams BA, Haggerty CM, Fornwalt BK. Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. *JACC Cardiovas Imaging.* 2019;12:681–689.
21. 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: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation.* 2020;142:e558–e631. doi: 10.1161/CIR.0000000000000937
22. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. *PMLR.* 2015;37:448–456.
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: Association for Computing Machinery; 2016:785–794. doi: 10.1145/2939672.2939785
24. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers.* 1999;10:61–74.
25. Bodison SA, Wesley YE, Tucker E, Green KJ. Results of screening a large group of intercollegiate competitive athletes for cardiovascular disease. *J Am Coll Cardiol.* 1987;10:1214–1221.
26. Hada Y, Sakamoto T, Amano K, Yamaguchi T, Takenaka K, Takahashi H, Takikawa R, Hasegawa I, Takahashi T, Suzuki J. Prevalence of hypertrophic cardiomyopathy in a population of adult Japanese workers as detected by echocardiographic screening. *Am J Cardiol.* 1987;59:183–184. doi: 10.1016/s0002-9149(87)80107-8
27. Baumgartner H, Hung J, Bermejo J, Chambers JB, Edvardsen T, Goldstein S, Lancellotti P, LeFevre M, Miller F Jr, Otto CM. Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *J Am Soc Echocardiogr.* 2017;30:372–392. doi: 10.1016/j.echo.2017.02.009
28. Davies M, Hobbs F, Davis R, Kenkre J, Roalfe AK, Hare R, Wosornu D, Lancashire RJ. Prevalence of left-ventricular systolic dysfunction and heart failure in the Echocardiographic Heart of England Screening study: a population based study. *Lancet.* 2001;358:439–444. doi: 10.1016/s0140-6736(01)05620-3
29. Banovic M, Putnik S, Penicka M, Doros G, Deja MA, Kockova R, Kotrc M, Glaveckaite S, Gasparovic H, Pavlovic N, et al; AVATAR Trial Investigators*. Aortic valve replacement versus conservative treatment in asymptomatic severe aortic stenosis: the AVATAR Trial. *Circulation.* 2022;145:648–658. doi: 10.1161/CIRCULATIONAHA.121.057639
30. Kagiyama N, Piccirilli M, Yanamala N, Shrestha S, Farjo PD, Casaclang-Verzosa G, Tarhuni WM, Nezarat N, Budoff MJ, Narula J, et al. Machine learning assessment of left ventricular diastolic function based on electrocardiographic features. *J Am Coll Cardiol.* 2020;76:930–941. doi: 10.1016/j.jacc.2020.06.061
31. Wolters FJ. An AI-ECG algorithm for atrial fibrillation risk: steps towards clinical implementation. *Lancet.* 2020;396:235–236. doi: 10.1016/S0140-6736(20)31062-X
32. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, Bernard ME, Rosas SL, Akfaly A, Misra A, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat Med.* 2021;27:815–819. doi: 10.1038/s41591-021-01335-4
33. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, Friedman PA, Lopez-Jimenez F. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circ Arrhythm Electrophysiol.* 2020;13:e007988. doi: 10.1161/CIRCEP.119.007988