Electrocardiogram screening for aortic valve stenosis using artificial intelligence

Michal Cohen-Shelly [1], **Zachi I. Attia** [1], **Paul A. Friedman**[1], **Saki Ito**[1], **Benjamin A. Essayagh** [1], **Wei-Yin Ko**[1], **Dennis H. Murphree** [1], **Hector I. Michelena** [1], **Maurice Enriquez-Sarano**[1], **Rickey E. Carter** [2], **Patrick W. Johnson** [2], **Peter A. Noseworthy**[1], **Francisco Lopez-Jimenez** [1], and **Jae K. Oh**[1]*

[1]Department of Cardiovascular Medicine, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA; and [2]Health Sciences Research, Mayo Clinic, 4500 San Pablo Rd S, Jacksonville, FL 32224, USA

Received 17 October 2020; revised 22 December 2020; editorial decision 23 February 2021; accepted 4 March 2021; online publish-ahead-of-print 22 March 2021

Listen to the audio abstract of this contribution.

|	
Aims	Early detection of aortic stenosis (AS) is becoming increasingly important with a better outcome after aortic valve replacement in asymptomatic severe AS patients and a poor outcome in moderate AS. We aimed to develop artificial intelligence-enabled electrocardiogram (AI-ECG) using a convolutional neural network to identify patients with moderate to severe AS.
Methods and results	Between 1989 and 2019, 258 607 adults [mean age 63 ± 16.3 years; women 122 790 (48%)] with an echocardiography and an ECG performed within 180 days were identified from the Mayo Clinic database. Moderate to severe AS by echocardiography was present in 9723 (3.7%) patients. Artificial intelligence training was performed in 129 788 (50%), validation in 25 893 (10%), and testing in 102 926 (40%) randomly selected subjects. In the test group, the AI-ECG labelled 3833 (3.7%) patients as positive with the area under the curve (AUC) of 0.85. The sensitivity, specificity, and accuracy were 78%, 74%, and 74%, respectively. The sensitivity increased and the specificity decreased as age increased. Women had lower sensitivity but higher specificity compared with men at any age groups. The model performance increased when age and sex were added to the model (AUC 0.87), which further increased to 0.90 in patients without hypertension. Patients with false-positive AI-ECGs had twice the risk for developing moderate or severe AS in 15 years compared with true negative AI-ECGs (hazard ratio 2.18, 95% confidence interval 1.90–2.50).
Conclusion	An AI-ECG can identify patients with moderate or severe AS and may serve as a powerful screening tool for AS in the community.

* Corresponding author. Tel: 507-266-1376, Fax: 507-266-9142, Email: oh.jae@mayo.edu

Graphical Abstract
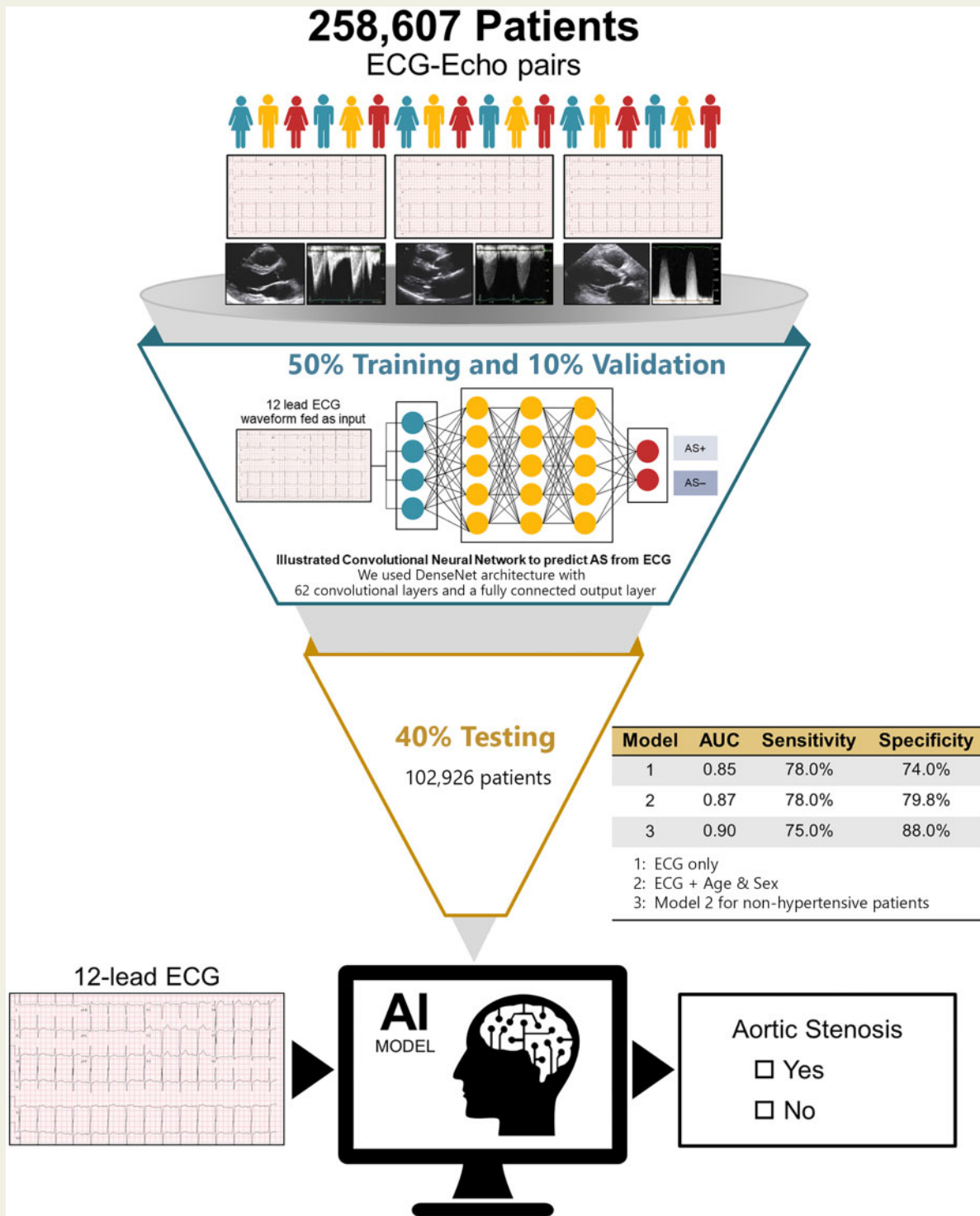
AI-ECG for Aortic Stenosis screening using convolutional neural network (CNN).

Introduction

Aortic stenosis (AS) has been managed by aortic valve replacement (AVR) usually when a patient develops symptoms.[1] If AVR is not performed at an appropriate time, severe AS can lead to heart failure or death.[2] Recently, benefit of an early AVR was demonstrated even in asymptomatic patients with severe AS[3] and long-term outcome was found to be poor in patients with moderate AS.[4,5] Together with an increasing number of patients being treated by transcatheter AVR, the current management paradigm for AS will continue to evolve for those patients with asymptomatic severe AS or less than severe AS.

During its early or asymptomatic period, AS is usually suspected by characteristic systolic murmur and confirmed by echocardiography study. However, careful auscultation may not be performed in asymptomatic patients and auscultation skill has been declining with advances in cardiac imaging. Actually, systolic murmur is documented in <50% of patients with moderate or severe AS.[6] Hence, there is an important clinical need for developing a novel and simple tool for identifying these patients with AS.

In this study, we sought to develop an artificial intelligence-enabled electrocardiogram (AI-ECG) as a screening tool for moderate to severe AS. The ECG is inexpensive and universally available, making it an excellent screening tool, but it does not detect AS in its standard form. However, AS increases left ventricular (LV) afterload, impairs systolic and diastolic function, and leads to cardiac remodelling with concomitant electrocardiographic changes.[7,8] We hypothesized that application of artificial intelligence (AI) in the form of a convolutional neural network (CNN) to the ECG would allow detection of individuals with AS of at least moderate severity.

This study has two objectives: (i) to develop and test the ability of the AI-ECG to identify patients with moderate to severe AS and (ii)

to assess the prognostic performance of the AI-ECG to identify the risk of future moderate or severe AS in individuals without significant AS at the time of initial screening.

Methods

Cohort identification

We identified patients aged ≥18 years who had at least one transthoracic echocardiogram (TTE) and ECG performed at our institution between January 1989 and September 2019 using the Mayo Clinic Unified Data Platform (UDP) that includes tests from the Minnesota, Arizona, and Florida locations. The details are described in *Figure 1*. Of the patients with TTE, included were only those with at least one of the following AS measurements: aortic valve area (AVA), mean transaortic pressure gradient, peak transaortic velocity, or dimensionless velocity index (DVI).[9,10] Of those, patients who had at least one digital, standard 12-lead ECG acquired within 180 days prior to their TTE exams were identified; all patients thus had ECG exams before TTE. When multiple TTEs and ECGs were available, we selected the earliest pair while minimizing the time interval. Patients missing measurements in the TTE report or with incomplete or corrupted ECG waveforms were excluded. Patients with previous cardiac surgery, a prosthetic valve or pacemaker were also excluded. Patients in whom TTE measurements (greater than or equal to moderate AS) and physician echocardiographers' final impression (less than moderate AS) were discrepant were excluded. The final cohort was assigned via outcome-stratified random sampling to training, validation, and testing subsets of 50%, 10%, and 40%, respectively. None of the patients were assigned to more than one group, thus no patients in the test set are seen by the model during training. Random selection across all centres was performed.

The Mayo Clinic Institutional Review Board approved this study and all patients had authorized research participation.
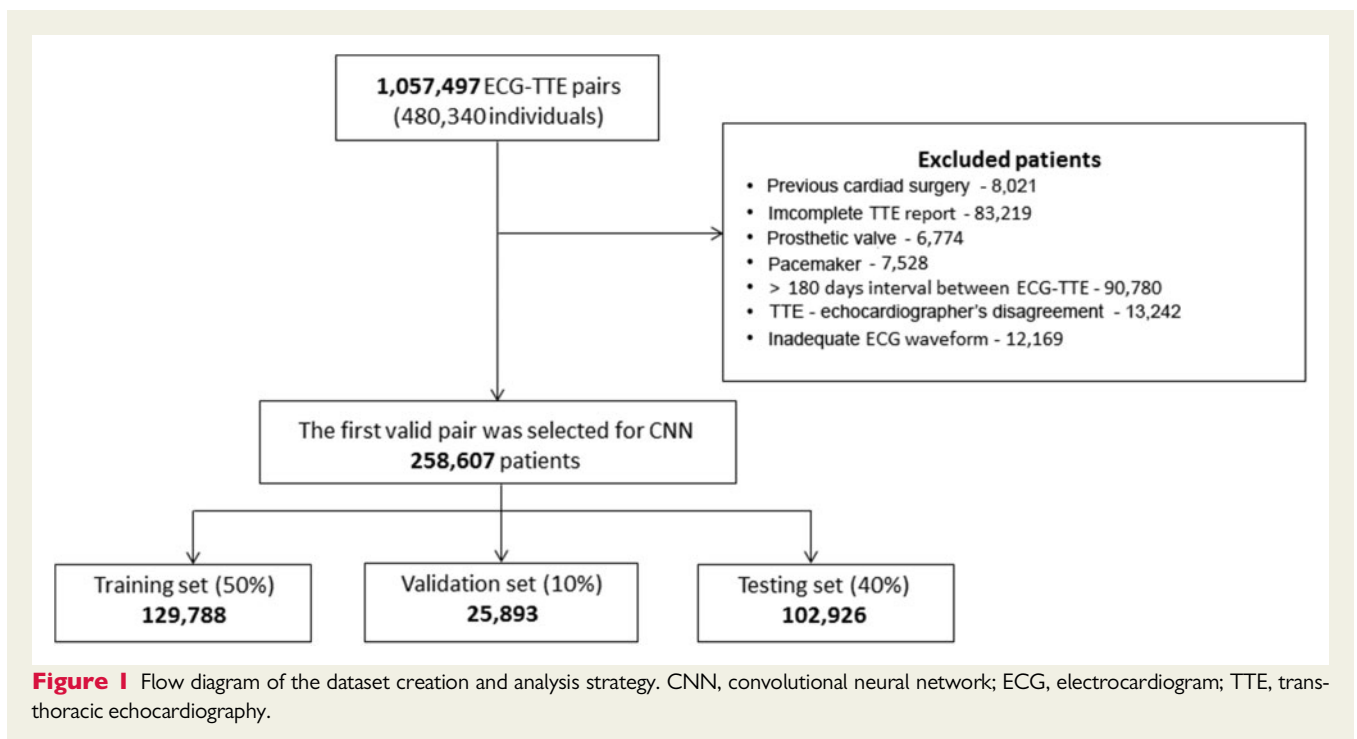

Figure 1 Flow diagram of the dataset creation and analysis strategy. CNN, convolutional neural network; ECG, electrocardiogram; TTE, transthoracic echocardiography.

Data sources and labelling

TTE data were used to classify patients into two groups: echo-positive AS (+) were those with moderate to severe AS and echo-negative AS (-) were those with mild or no AS by using published guidelines (Supplementary material online, Table S1).[1,10] If a patient satisfied one of the following echocardiography parameters, AS severity was classified as moderate to severe: peak velocity $\geq$3.0 m/s, mean gradient $\geq$20 mmHg, DVI $\leq$0.35, or AVA $\leq$1.5 cm^2.

Transthoracic echocardiogram

Mean pressure gradient and peak velocity were acquired by continuous-wave Doppler from all available transducer positions to obtain the highest values.[10,11] Left ventricular outflow tract (LVOT) velocity was measured by pulsed wave Doppler. DVI was calculated as the ratio between LVOT and aortic valve velocity time integral.[9] AVA was calculated using the continuity equation.[9,10]

Electrocardiogram

All ECGs were acquired as digital standard 12-lead ECGs using a Marquette ECG machine (GE Healthcare, WI, USA). Their raw data were stored using the MUSE data management system for later retrieval.

Outcomes

The primary outcome was the ability of the AI-ECG to identify moderate to severe AS (echo-positive AS). The second outcome was the ability of the AI network to determine if subjects that are deemed positive by the AI model (AI-ECG-positive AS) but actually had echo-negative AS at the time of screening (false positives) had higher risk of developing echo-positive AS in the future compared with those who were truly negative and were deemed negative by the AI model.

Demographic data tabulation

In order to characterize the study population and identify comorbidities at the time of ECG, the UDP was queried using standardized ICD-9 and ICD-10 billing codes for each diagnosis within 30 days post-ECG. The UDP was also queried to assess the referral reasons for TTE.

Overview of artificial intelligence model development

We developed a CNN model using Keras framework with Tensorflow (Google; Mountain View, CA, USA) backend implemented in Python.[12] This framework was used successfully for creating models to screen LV contractile dysfunction and to estimate age as well as sex from standard 12-lead ECG.[13,14] Each ECG was considered a matrix of the following dimensions: 12 × 5000 (representing 12 leads for 10 s duration sampled at 500 Hz), ECGs that were originally sampled in 250 Hz were up-sampled to 500 Hz using the 'Resample' function of the SCIPY python package[15]; the 1st dimension is spatial dimension and represents the different ECG leads and the 2nd dimension is temporal. The CNN model is based on a smaller version of DenseNet with 62 convolutional layers and 1 classification layer (Figure 2).[16] DenseNet uses densely connected convolutional blocks to concatenate the result of each convolutional output within the block in order to extract detailed features. The features extracted by the dense blocks were fully connected to the final layer that had two neurons activated using a softmax function, and later were represented as the probabilities of the ECG being from an AS vs. non-AS patient.[13,14] Minor modifications regarding zero-padding were made to the original network to account for the difference in image and ECG matrix inputs.
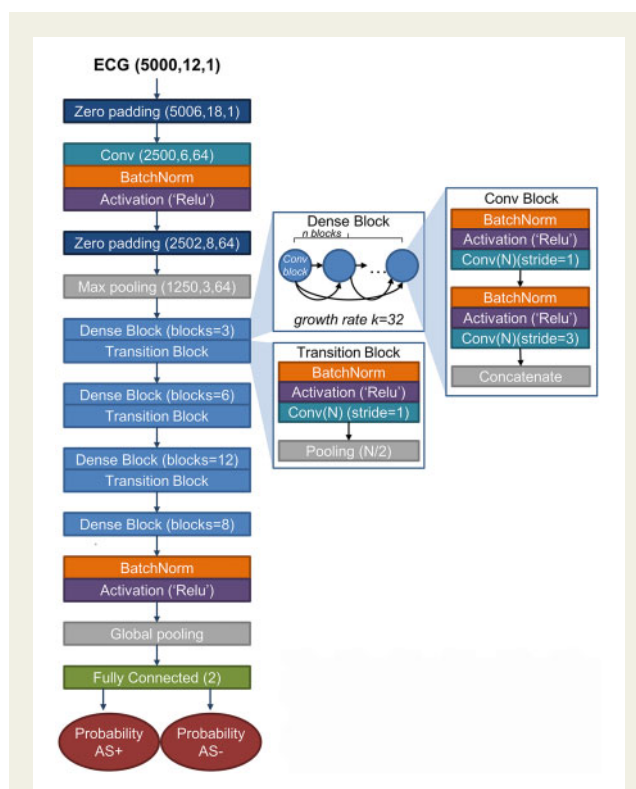

Figure 2 Selected model architecture. Selected model architecture is shown, four layers dense block with a growth rate of $k = 32$. Each layer takes all preceding feature-maps as input which connects each layer to every other layer in a feed-forward fashion. The layers between two dense blocks are referred to as transition layers and change feature map sizes via convolution and pooling. AS, aortic stenosis; BatchNorm, batch normalization, Conv, convolution; ECG electrocardiogram.

To create a model that included tabular variables such as age and sex, the same CNN architecture and hyper-parameters from the ECG-only model were used, with the last layer's extracted ECG features concatenated with age and sex prior to the final layer.

Adam optimizer was used for training, with categorical cross-entropy as the loss function, categorical cross-entropy used even though it is a binary classifier due to the use of one hot encoding and having one output neuron for AI-ECG-positive AS and one for AI-ECG-negative AS. Hyper-parameters such as learning rate (1e-3) and batch size (64) were tuned using a validation set. An area under the receiver operating characteristic curve (ROC AUC) was calculated for the internal validation set after each epoch. After training completion, the model with the highest validation set AUC was evaluated on the holdout 40% test set. The number of training epochs was governed by an early stopping strategy, where the network weights were updated as long as the loss was improving. After 10 epochs without improvement, training was discontinued.

Training of the algorithm was multifaceted. To protect against biasing our estimate of the model performance, the training data were used exclusively for developing the model architecture. The threshold for classifying an ECG as either a positive or negative screen was determined using Youden index in the validation data.[17] After training, the model was used to make predictions on the validation set, and an ROC was calculated. For each point in the curve, the Youden index was calculated, and the

Table I Patients characteristics and comorbidities

	Training set (n = 129 788)	Validation set (n = 25 893)	Testing set (n = 102 926)
Age, years	62.99 ± 16.3	63.09 ± 16.3	62.97 ± 16.3
Age groups			
<40	12 674 (9.8)	2508 (9.7)	10 094 (9.8)
40–49	12 978 (10.0)	2542 (9.8)	10 234 (9.9)
50–59	22 301 (17.2)	4466 (17.2)	17 909 (17.3)
60–69	31 231 (24.1)	6202 (24.0)	24 970 (24.2)
70–79	30.984 (23.9)	6242 (24.1)	24 077 (23.3)
≥80	19 620 (15.1)	3929 (15.2)	15 642 (15.2)
Female sex	61 514 (47.3)	12 288 (47.4)	48 988 (47.5)
AS measurement severity level			
No AS	114 646 (88.3)	22960 (88.7)	90 763 (88.1)
Mild AS	10 194 (7.9)	1991 (7.7)	8330 (8.1)
Moderate AS	1605 (1.2)	300 (1.5)	1225 (1.2)
Severe AS	3343 (2.6)	642 (2.5)	2608 (2.5)
Hypertension	63 244 (48.7)	12 621 (48.7)	50 486 (49.1)
Congestive heart failure	23 399 (18.0)	4733 (18.3)	18 531 (18.0)
Renal disease	15 641 (12.1)	3168 (12.2)	12 394 (12.0)
Chronic pulmonary disease	26 312 (20.3)	5210 (20.1)	20 932 (20.3)
Myocardial infarction	12 097 (9.3)	2446 (9.4)	9843 (9.6)
Diabetes mellitus	22 591 (17.4)	4563 (17.6)	18 186 (17.7)

Values are expressed as mean ± standard deviation, or n (%). Any observed difference in comorbidities is a result of random chance.
AS, aortic stenosis.

threshold corresponding to the point of highest Youden index was selected as the optimal threshold. Subsequently, the final model performance was assessed using the testing data. In the training process, we trained multiple CNN model architectures and selected the one with the highest AUC of the ROC, using the internal validation set. Precision-recall and calibration curves were created to assess model performance.

Saliency maps
In order to understand which portion of the ECG weighed in our model's prediction of AS, we created saliency maps using keras-vis python package.[18]

Statistical analysis
All measures of performance are based on the testing data. The ROC curve was constructed for assessing the model performance formed by modelling the CNN's prediction of the probability of AI-ECG-positive AS in relationship to the TTE-positive AS. Applying a threshold determined in the validation data indicating a positive screen, standard measures of diagnostic performance (AUC, sensitivity, specificity, and accuracy) were computed. Except for AUC, 95% exact confidence intervals (CI) were calculated for all measures of diagnostic performance using the large sample approximation of the DeLong method with optimization by the Sun and Xu method.[19] The diagnostic odds ratio, which is the ratio of positive likelihood ratio [sensitivity/(1 - specificity)] to the negative likelihood ratio [(1 - sensitivity)/specificity], and the associated 95% CI were also calculated.

To investigate the prognostic performance of the AI algorithm at detecting TTE-positive AS over time, all TTE-negative AS patients were classified into either true negative- or false-positive screens. Of those, individuals with ≥2 TTE–ECG pairs were modelled longitudinally until

the first date of reported as TTE-positive AS. The date of censoring was set at the last date of all available paired TTE–ECGs and the date of the first TTE-positive AS during follow-up, if applicable. The log-rank test was applied. The hazard ratio (HR) and 95% CI were calculated to quantify the relative differences in the hazard for the development of AS.

Continuous variables were summarized as mean and standard deviation or median [interquartile range (IQR)] when appropriate. Categorical variables were summarized using numbers and percentages. To compare subject characteristics across model prediction results (e.g. true positives, false negatives, etc.), one-way analysis of variance or the Wilcoxon rank-sum test (non-normal data) was employed. Binary data were compared with a χ^2 test. Statistical analyses were computed using Python 3.6, R version 3.6.2 (The R Foundation, Vienna, Austria) and JMP software, version 14.1 (SAS Institute, Cary, NC, USA).

Results

Of 480 340 patients who had both TTE and ECG, 258 607 patients (54%) had valid ECG–TTE pairs (*Figure 1*). Mean age was 62.9 ± 16.3 years with 122 790 (48%) women. Median time interval between ECG and TTE was 0 days (IQR 0, 4); 169 252 (65%) and 232 724 (90%) had them within 1 and 30 days of each other, respectively. The prevalence of moderate to severe AS (TTE-positive AS) was 3.7%. Of 258 607 patients, 129 788 (50%) were used for training, 25 893 (10%) for validation, and 102 926 (40%) for testing (*Graphical abstract*). The proportion of patients from each medical centre was similar between training, validation, and testing groups (e.g. testing group: Minnesota 70%, Florida 9%, and Arizona 21%). Patients'

characteristics and AS severity distribution were similar among the three groups (*Table 1*). In the testing group, most patients were Caucasian (88%, n = 90 938) followed by Black (3%, n = 2874), Hispanic (2%, n = 2129), Asian (1%, n = 1313), and others. The most common referral reason for TTE was dyspnea followed by hypertension, coronary artery disease, atrial fibrillation, and cardiac murmur (Supplementary material online, *Table S2*). Of 6351 patients with cardiac murmur as the reason for TTE, moderate or severe AS was diagnosed in 446 (7%). There was a significant difference in several

comorbidities including hypertension, congestive heart failure, and diabetes mellitus between AS (+) and AS (-) patients (*Table 2*).

The performance of the artificial intelligence-enabled electrocardiogram for detecting aortic stenosis

The threshold for the probability of classifying an ECG into TTE-positive AS and TTE-negative AS screen was established in the validation data as 0.0243. Applying the threshold of ≥0.0243 indicating a positive screen, the AUC for identifying echo-positive AS (+) and echo-negative AS (-) subjects was 0.85 in both validation and testing groups (*Figure 3A*). In the testing group, 3833 (3.7%) patients were labelled as AI-ECG-positive AS with sensitivity, specificity, and accuracy for predicting echo-positive AS of 78%, 74%, and 74%, respectively. Positive predictive value was low at 10.5%, but negative predictive value was 98.9%. Of a total of 102 926 patients in the testing group, true positive was present in 3% (n = 2995), true negative in 71% (n = 73 624), false positive in 25% (n = 25 469), and false negative in 1% (n = 838) (*Table 3*). Clinical characteristics, echocardiographic data, and ECG parameters are compared between the four groups (*Table 4*). Patients with false positive more frequently have hypertension and renal disease compared with other groups (*P* < 0.0001). Electrocardiogram features, such as QRS duration, QT interval, R-wave axis, and T-wave axis were significantly different between the four groups (*P* < 0.0001 for all).

Further analysis stratifying based on age and sex indicates (*Figure 3B*) that the sensitivity gradually increased and the specificity decreased as age increased. Women had lower sensitivity but higher

Table 2 Patient characteristics and comorbidities by aortic stenosis (entire cohort)

	AS (−) (n = 248 884)	AS (+) (n = 9723)	P
Hypertension (%)	120 797 (48.5)	5554 (57.1)	<0.0001
Congestive heart failure (%)	44 031 (17.7)	2632 (27.1)	<0.0001
Renal disease (%)	29 655 (11.9)	1548 (15.9)	<0.0001
Chronic pulmonary disease (%)	50 335 (20.2)	2119 (21.8)	0.00016
Myocardial infarction (%)	23 386 (9.4)	1000 (10.3)	0.0033
Diabetes mellitus (%)	43 097 (17.3)	5554 (57.1)	<0.0001

Values are expressed as *n* (%).
AS, aortic stenosis.
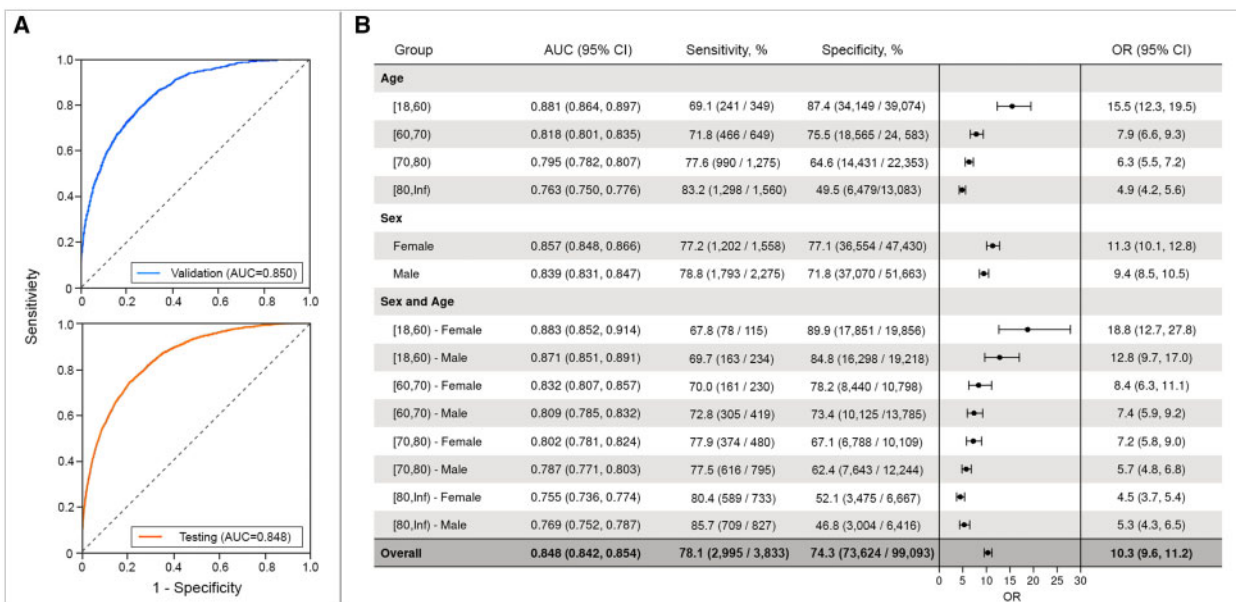
Figure 3 Receiver operating characteristic curves, sensitivity and specificity across age and sex subsets. (A) The receiver operating characteristic curve of the convolutional neural network for identifying patients with moderate to severe aortic stenosis is shown for the validation group (upper panel) and testing cohort (lower panel). The area under the curve (AUC) is calculated. (B) The sensitivity and specificity for the detection of moderate to severe aortic stenosis labelled by artificial intelligence electrocardiogram are tabulated across a range of sex and age combinations for testing dataset. The diagnostic odds ratio (OR) and 95% confidence interval (CI) are shown. (Parenthesis) excludes the numbers shown and [bracket] includes them.

Table 3 Artificial intelligence-enabled electrocardiogram model performance

Testing group (n = 102 926)	Aortic stenosis severity			
	Normal	**Mild**	**Moderate**	**Severe**
True positive (n = 2995, 3%)	0 (0)	0 (0)	830 (28)	2165 (72)
True negative (n = 73 624, 71%)	68 976 (94)	4648 (6)	0 (0)	0 (0)
False positive (n = 25 469, 25%)	21 787 (86)	3682 (15)	0 (0)	0 (0)
False negative (n = 838, 1%)	0 (0)	0 (0)	395 (47)	443 (53)

Values are expressed as n (%).

Table 4 Clinical characteristics, echocardiography, and electrocardiogram parameters for four artificial intelligence-enabled electrocardiogram groups

	True positive (n = 2995)	True negative (n = 73 624)	False positive (n = 25 469)	False negative (n = 838)	P
Age, years	76.9 ± 11.1	59.6 ± 16.2	70.7 ± 13.5	73.5 ± 11.8	<0.0001
Female sex	1202 (40)	36 554 (50)	10 876 (43)	356 (42)	<0.0001
Hypertension	1723 (58)	31 643 (43)	16 682 (66)	438 (52)	<0.0001
CHF	883 (28)	10 176 (14)	7347 (29)	175 (21)	<0.0001
Renal disease	493 (16)	7008 (10)	4767 (19)	126 (15)	<0.0001
COPD	665 (22)	13 921 (19)	6167 (24)	179 (21)	<0.0001
MI	326 (11)	6060 (8)	3374 (13)	83 (10)	<0.0001
Diabetes mellitus	707 (24)	11 071 (15)	6244 (25)	164 (20)	<0.0001
PVD	738 (25)	9489 (13)	5709 (22)	198 (24)	<0.0001
Echocardiography					
Aortic valve area, cm^2	0.96 ± 0.25	3.08 ± 0.90	2.87 ± 0.94	1.07 ± 0.25	<0.0001
Peak velocity, m/s	4.06 ± 0.78	1.41 ± 0.32	1.51 ± 0.40	3.67 ± 0.63	<0.0001
Mean pressure gradient, mmHg	41.0 ± 16.5	5.10 ± 2.93	6.46 ± 3.69	32.6 ± 11.5	<0.0001
DVI	0.25 ± 0.07	0.79 ± 0.18	0.74 ± 0.20	0.28 ± 0.06	<0.0001
ECG measurement (II lead)					
QRS duration, ms	105.1 ± 24.5	92.9 ± 16.7	104.9 ± 25.7	97.1 ± 19.7	<0.0001
QT interval, ms	407.6 ± 42.7	397.9 ± 45.8	404.4 ± 46.1	407.1 ± 50.4	<0.0001
QTc, ms	443.0 ± 33.6	434.1 ± 32.5	448.2 ± 36.7	440.9 ± 34.1	<0.0001
P axis	46.4 ± 29.5	47.9 ± 26.1	46.5 ± 30.6	47.9 ± 28.7	<0.0001
R axis	8.8 ± 46.3	25.6 ± 41.6	13.2 ± 53.1	17.5 ± 44.8	<0.0001
T axis	64.8 ± 64.1	42.9 ± 40.6	56.6 ± 60.1	50.6 ± 52.9	<0.0001
Ventricular rate, b.p.m.	73.4 ± 0.34	74.5 ± 0.07	76.6 ± 0.12	73.3 ± 0.65	<0.0001
Atrial fibrillation/flutter	484 (16)	5593 (8)	4584 (18)	99 (12)	<0.0001

Values are expressed as mean ± standard deviation, or n (%).
CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; DVI, dimensionless velocity index; ECG, electrocardiogram; MI, myocardial infarction; PG, pressure gradient; PVD, peripheral vascular disease.

specificity compared with men at any age groups. In all subgroups, the odds ratio remained >5. Since hypertension shares similar ECG changes with those from AS, the AUC was calculated separately in patients with and without hypertension. It was 0.81 and 0.88, respectively. The AUC was 0.89 for those without any comorbidities [n = 31 484 (31%)]. Of note, there was a positive correlation between AI-ECG and echocardiographic diagnosis of AS severity (*Figure 4*).

The precision-recall curve is in line with expectations for trying to model a clinically rare outcome; the calibration curve shows excellent linearity (Supplementary material online, *Figure S1*).

Model comparison

We tested if additional variables could improve our model performance. Compared with ECG morphology alone (AUC 0.85), AUC improved modestly when age and sex were added to the current model (AUC 0.87, sensitivity 78%, specificity 80%). Other variables, such as height, weight, body mass index as well as available ECG measurements (e.g. cardiac rhythm, heart rate, QT interval, QRS duration, QTc, R-wave axis, T-wave axis) did not improve the model performance. In the model of ECG with age and sex, the AUC improved further to 0.90 for non-hypertensive patients (sensitivity 75%, specificity 88%).
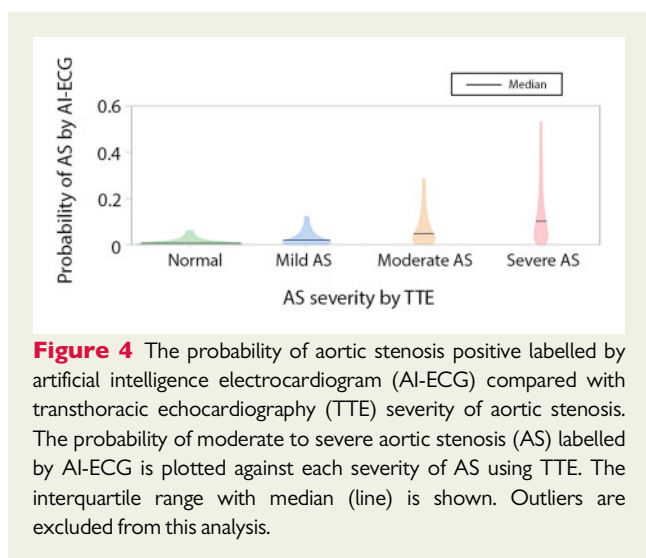
Figure 4 The probability of aortic stenosis positive labelled by artificial intelligence electrocardiogram (AI-ECG) compared with transthoracic echocardiography (TTE) severity of aortic stenosis. The probability of moderate to severe aortic stenosis (AS) labelled by AI-ECG is plotted against each severity of AS using TTE. The interquartile range with median (line) is shown. Outliers are excluded from this analysis.
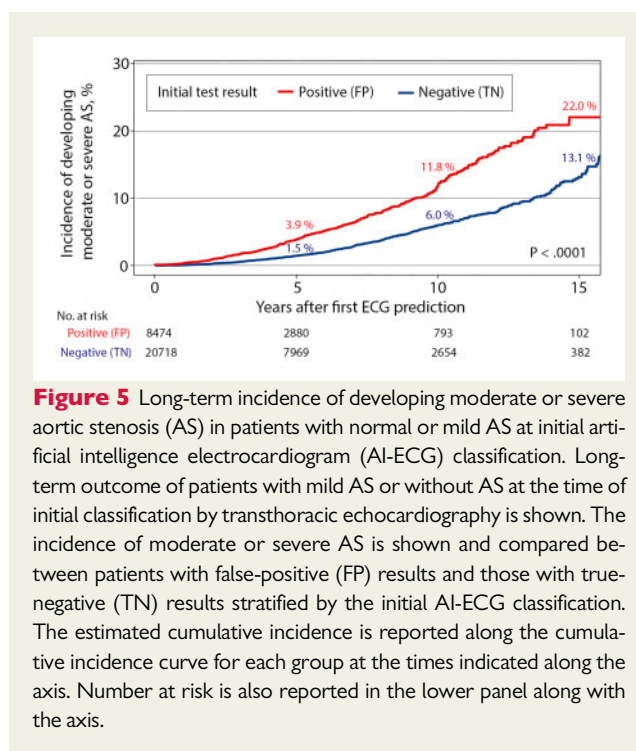

Figure 5 Long-term incidence of developing moderate or severe aortic stenosis (AS) in patients with normal or mild AS at initial artificial intelligence electrocardiogram (AI-ECG) classification. Long-term outcome of patients with mild AS or without AS at the time of initial classification by transthoracic echocardiography is shown. The incidence of moderate or severe AS is shown and compared between patients with false-positive (FP) results and those with true-negative (TN) results stratified by the initial AI-ECG classification. The estimated cumulative incidence is reported along the cumulative incidence curve for each group at the times indicated along the axis. Number at risk is also reported in the lower panel along with the axis.

Prediction of future aortic stenosis

Of 99 093 (96.3%) echo-negative AS patients in the testing cohort, 29 192 had ≥2 ECG–TTE pair with 8474 false-positive AI-ECG (AI-ECG-positive AS) and 20 718 true negative AI-ECG (AI-ECG-negative AS). With 15 years of follow-up, there were 1796 incident cases of echo-positive AS from the time of first TTE–ECG pair. The event rate in the false-positive group was 22.0% vs. 13.1% in the true-negative group (*Figure 5*). The false-positive group had almost twice the risk for development of moderate or severe AS compared with the true-negative group (HR 2.18, 95% CI 1.90–2.50; *P* < 0.0001). Of note, in the testing cohort, prevalence of mild AS was 15% (*n* = 3682) in the false-positive group and 6% (*n* = 4648) in the true-negative group (*P* < 0.001, *Table 3*).

Saliency map

Representative ECG example for a patient from the true-positive group is shown in *Figure 6*. The precordial leads (especially V1–V3) are more weighted compared with limb leads and the higher 'saliency' is frequently located from the end of T wave to the beginning of P wave (TP segment). Surprisingly, typical ECG findings for LV hypertrophy (e.g. increased R-wave amplitude, increased S-wave depth) that has been shown to be associated with high afterload in AS are not weighted.

Discussion

Our study demonstrated that the AI-ECG, a simple routine test, can successfully identify patients with moderate to severe AS with high performance (AUC 0.85), comparable or superior to currently used medical tests including pap smear detecting cervical uterine cancer (AUC 0.71)[20] and AI-interpret mammography detecting breast cancer (AUC 0.76–0.89).[21]

Aortic stenosis is usually suspected by a characteristic systolic murmur on auscultation and diagnosed by echocardiography, but may not be detected until symptoms develop[6,22] as demonstrated in the illustrative case (Supplementary material online, *Figure S2*). Auscultation of cardiac murmur is an important clinical finding in detecting valvular heart disease, but it was found that murmur was a key leading to the diagnosis of AS only in 62% of asymptomatic patients.[22] Kattoor *et al.*[6] reviewed 95 patients with moderate to severe AS diagnosed by echocardiography, and they found that murmur of AS was identified by only 39% of clinicians. They also reported that auscultation skill varies among physicians based on their specialty or experiences; AS murmur was detected in 87% of the patients seen by cardiology specialists, 50% by other medical specialists and <20% by non-medical clinicians. Thus, training in auscultation can potentially improve detection rate for AS; however, our physical examination skills have been eroding in the era of advanced imaging techniques. Therefore, our AI-ECG will be helpful and expected to increase the detection of AS even without symptoms or documented AS murmur.

Our AI-ECG model performed well (AUC 0.85) with high sensitivity and specificity in the identification of patients with moderate to severe AS. However, a careful interpretation is necessary especially when the model indicates presence of AS. With a prevalence of moderate or severe AS of about 4%, the positive predictive value was low and there may be a concern for performing unnecessary echocardiography examinations in the AI-ECG positive patients. The AI-ECG result thus needs to be integrated into clinical evaluation including comorbidities associated with AS, symptomatic status, and more careful auscultation before performing additional tests. Another important clinical value of AI-ECG is its high negative predictive value close to 99%. It will be thus helpful if AI-ECG can be used for excluding AS. Systolic murmur is non-specific and a minority of the patients
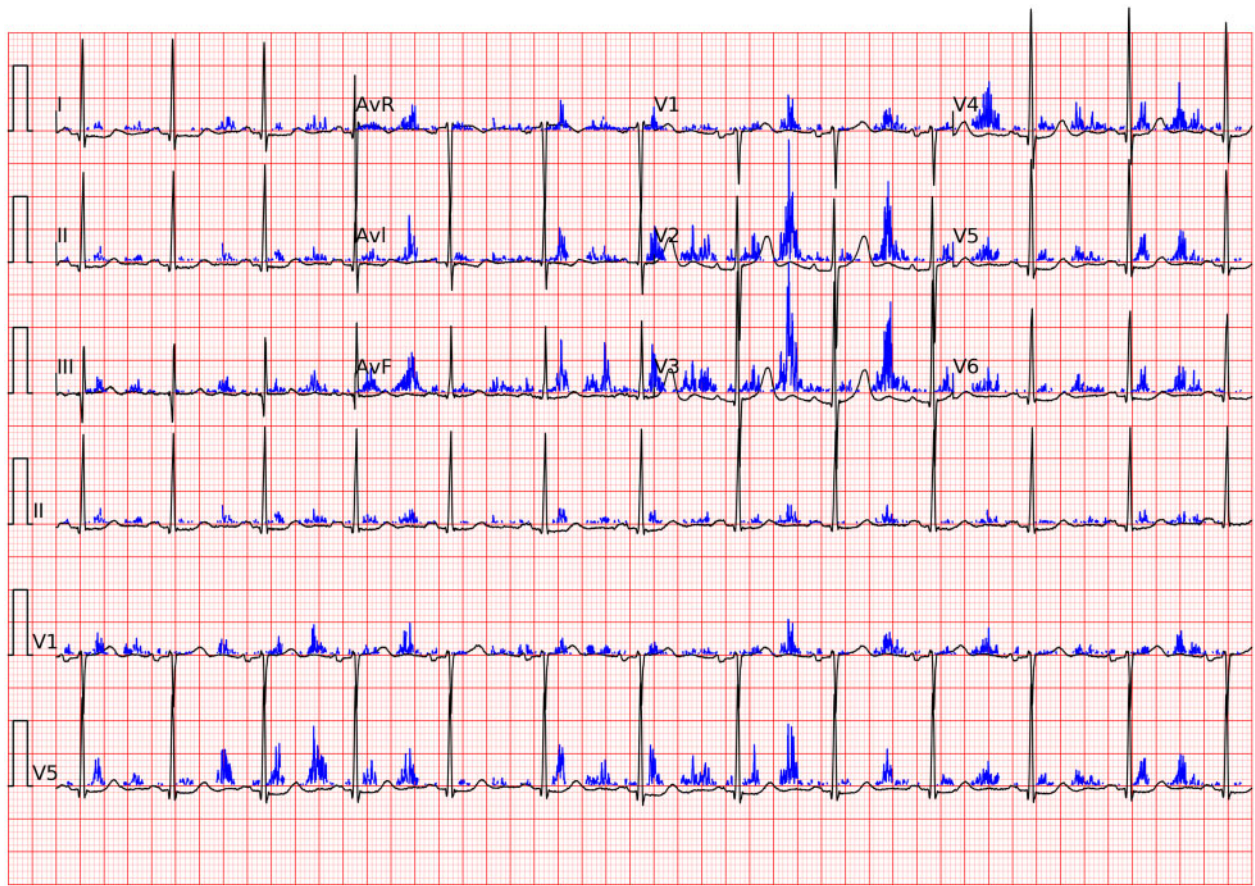
Figure 6 Saliency map. A representative electrocardiogram example for true positive is shown. Probability of moderate or severe aortic stenosis by artificial intelligence electrocardiogram is 0.92 in the presented case. The blue lines are the 'saliency' guiding the selection of attended locations.

referred to echocardiography laboratory for evaluation of cardiac murmur are diagnosed with AS. When we reviewed 6351 patients who were referred to echocardiography because of cardiac murmur among the testing group, only 446 patients (7%) were diagnosed to have moderate or severe AS. Our AI-ECG model with the excellent negative predictive value can reduce the number of unnecessary imaging in patients with non-significant murmur. There was a significant false-positive rate in our model, but a false-positive test indicates elevated future risk of developing significant AS (*Figure 5*). This is probably due to the fact that the prevalence of mild AS was more frequent in the false-positive group compared with the true-negative group (*Table 3*). Also, in the false-positive group, prevalence of hypertension and renal disease was higher compared with other groups (*Table 4*). These comorbidities are known to be associated with LV hypertrophy, and it may be thus possible that the AI-ECG identified them as AS positive. Patients with AI-ECG positive for AS warrant a close follow-up and surveillance even if AS murmur is absent.

Until recently, little effort has been made for detecting patients with AS in its earlier asymptomatic stage since patients are usually treated by AVR when they become symptomatic. However, symptoms are subjective and the lack of AS-related symptoms is not benign. Sudden death was reported in 4.1% in patients with asymptomatic severe AS,[23] and a recent clinical trial demonstrated a better clinical outcome in early AVR in this population.[3] Moreover, emerging risk factors such as brain natriuretic peptide and imaging evidence of myocardial fibrosis will likely expand the indication for AVR in asymptomatic severe AS patients.[24] Irreversible replacement by fibrosis is a major consequence of LV response to the pressure overload imposed by AS[7,25] which often does not normalize after AVR.[26] Reduced LV ejection fraction (LVEF) has been shown to be associated with worse survival outcomes despite AVR.[27] Patients with reduced LVEF and severe AS have evidence of abnormal systolic function as early as 10 years prior to developing severe AS with more precipitous decline at AVA of 1.2 cm^2.[27] As shown in the illustrative case (Supplementary material online, *Figure S2*), LVEF was reduced to 25% when the patient presented with pulmonary oedema. His LVEF did not improve after AVR and heart failure symptoms continued. In such patients, early detection of AS by an AI-ECG could potentially provide improved timing of AVR with better clinical outcome. Besides AVR, earlier detection of AS can potentially provide another treatment option in the coming future. An ideal management strategy of AS is a medical therapy to delay or prevent the progression of aortic valve calcification or haemodynamic severity. Although several trials using statin or lipid-lowering agents have failed to prove its

beneficial effect, dipeptidyl peptidase-4 inhibitor has recently been shown to delay the progression of aortic valve calcification in an animal model and to delay progression of AS in diabetic patients in a retrospective study.[28,29] A *post hoc* analysis of FOURIER exploratory data showed that proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitor, evolocumab, showed a significant 52% relative reduction in AS events compared with control patients.[30] If their beneficial effect can be confirmed in a prospective clinical trial, it will be even more critical to identify patients with earlier stage of AS.

It will be helpful to know which features of the ECG contribute to the AI-ECG's ability to detect AS. As demonstrated in *Table 4*, there are significant differences in ECG features between true-positive, true-negative, false-positive, and false-negative groups. However, these values fall within the normal range and poorly contributed to the model performance. It is thus unlikely that these would be useful features for identification of AS during manual ECG interpretation. On the other hand, a saliency map shows that TP segment or U wave in the right precordial leads is weighted most heavily for determining the presence of AS in our model. Surprisingly, the QRS complex (e.g. increased R-wave amplitude or increased S-wave depth that is typical for patients with LV hypertrophy in response to AS) is not weighted. Somewhat paradoxically, even though LV hypertrophy is a known key adaptation mechanism of the heart to AS, it does not appear to play a major role in this AI-ECG model. This finding has been observed in other similar studies. The AI-ECG model for AS from a Korean population using CNN showed that the initial area of T wave in V2–V5 was the most important region in their model for determining the presence of AS using a sensitivity map.[31] Although the number was small ($n = 700$), the AI-ECG model for AS from the Japanese population showed that ST-T segment is weighted using the gradient-weighted class activation mapping.[32] Neither of these studies demonstrated importance of the QRS complex or typical LV hypertrophy findings on ECG. Further studies are necessary to understand the physiologic underpinnings of these ECG findings.

Our model performance improved when age and sex were added to ECG morphology. Furthermore, our data showed that there was a sex difference in sensitivity and specificity of AI-ECG; higher sensitivity and lower specificity in men than in women across all age groups (*Figure 3*). Sex difference in LV remodelling in AS has been shown, which may explain the finding.[25] Furthermore, the AI-ECG had higher sensitivity and lower specificity in older population. With ageing, there is increased incidence of hypertension, diabetes, coronary artery disease, and diastolic dysfunction. Electrocardiogram changes perhaps related to these conditions may do share some features of AS-related changes. Comorbidities can produce similar ECG changes as AS. Indeed, there was a significant difference in the prevalence of comorbidities between AS (+) and AS (-) patients. It is therefore interesting to find that AI-ECG performed better in patients without any comorbidities including hypertension.

The network our model created successfully characterized the levels of AS severity (*Figure 4*). There was a positive correlation between the probabilities of AS labelled by the neural network and AS severity by TTE. It means more distinct ECG changes develop as the severity of AS progresses. This also explains a better performance of AI-ECG for detecting patients with severe AS. In addition, patients who had false-positive results were shown to have a significantly higher risk of developing moderate or severe AS in future compared with those with true-negative results (*Figure 5*). In certain clinical situations, reporting a probability may be more meaningful than reporting a dichotomous result (AS positive or negative) for longitudinal patient management.

Kwon *et al.*[31] described a similar concept as our study in a Korean population. Their model is well designed with high AUC of 0.861 on external test group and included age, sex, height, weight, body mass index, and ECG measurements in addition to ECG morphology. We thus tested several models; when we added clinical and ECG data to our model just based on ECG morphology, AI-ECG with age, and sex improved the AUC (from 0.85 to 0.87). Our model performance was comparable with the model from the Korean population (0.87 vs. 0.86); however, our main purpose is developing a model as a simple screening tool, thus, the number of variables in our model was minimum. Our study embodies several subtle but potentially important differences relative to the work of Kwon *et al.* While our network's training population is significantly larger, an aspect traditionally considered an advantage, there is an accompanying diversity in the patients included. Notably, the study by Kwon *et al.* focuses on a predominantly Asian population. Our cohort included three geographically different centres in the USA with the majority of patients being Caucasian (88%), followed by Black (3%), Hispanic (2%), and Asian (1%) in the testing group. Lastly, for the training of the AI model, Kwon group used several ECGs per patient.[31] On the other hand, our study used single ECG per patient in the training for preventing the biased testing and providing spurious high accuracy. Regardless of these differences, both studies successfully confirmed that AI-ECG is able to screen patients with AS with clinically useful power.

While in this study we focus on the AI-ECG's ability to screen for AS, using the AI-ECG to simultaneously screen for other cardiac disease such as LV dysfunction, hypertrophic cardiomyopathy, and silent atrial fibrillation,[13,14,33] may allow for a comprehensive identification of the individuals at high risk for having concealed, treatable cardiovascular disease who may need to undergo echocardiography or other diagnostic testing after a more careful clinical evaluation. There is a concern that the AI-ECG model may identify 'cardiac disease' in general as opposed to being specific to a certain condition. When we applied the AI-ECG model developed for screening reduced LVEF to our testing population, its AUC was 0.59 (not shown) in detecting moderate to severe AS, suggesting that our model is indeed learning features specific to AS. Such a broad strategy may improve cardiac health, augment adoption of evidence-based treatments, and optimize utilization of expensive imaging resources in the community.

Finally, our model was developed in patients who were referred to clinic, thus, all ECGs and echocardiography exams were performed for a clinical reason. Since an extremely large number of patients were involved in this study, various ECG morphologies and echocardiography findings are expected to be included and thus learned by the model. Therefore, our AI-ECG model can be used for population-based screening for AS.

Limitations

This is a study from three tertiary referral medical centres and thus may reflect a referral-biased population. However, our medical centre has three geographically separated locations (Minnesota,

Arizona, and Florida), thus covering diverse patient populations. Some important clinical information such as cardiac murmur is lacking in this study to see how our AI-ECG model performs in patients with or without cardiac murmur. Since we use large dataset, it is difficult to review all medical record to check the presence of murmur. However, of 6351 patients with referral reason of cardiac murmur for TTE, only 7% were found to have moderate or severe AS. AI-ECG will be thus clinically helpful for excluding significant AS even in patients with cardiac murmur, although some of them may have another valvular heart disease. Although negative predictive value was high as 99% in our model indicating excellent accuracy of negative results, false negative was present in 1%. Clinical characteristics of the false-negative patients were similar to those of true-positive patients, but significantly different from the true-negative group (*Table 4*). The false-negative group had a higher proportion of moderate AS and lower aortic valve velocity compared with the true-positive group, which might have been responsible for less characteristic ECG changes of AS (*Table 3*). We will therefore need further investigation to explain why ECG morphology in false-negative patients resembles that of the true-negative group. We did not test our AI-ECG model in an external group to assess its generalizability in this study. Additional testing of our AI-ECG in different races and different parts of the world is currently planned. AI-ECG has varying sensitivities and specificities for different sex, age, and comorbidities. We will need to refine our AI-ECG model based on patient's sex, age, and clinical conditions sharing ECG changes seen in AS to increase the screening diagnostic accuracy even higher. In this study, only one type of ECG machine was used and other ECG machine was not available in the current study. Therefore, we cannot discuss how different types of ECG machine impact on the results. Lastly, haemodynamics of AS depends on the flow status and there are several different types of AS but it was not addressed in the current study. Stress test is helpful in reclassifying the severity of AS and also risk stratification based on the extent of increase in pressure gradient as well as other clinical and haemodynamic data.[34] However, we did not include stress test result for this study.

Conclusion

Application of AI in the form of a CNN to a 12-lead ECG—an inexpensive, ubiquitous, commonly used test—enables it to serve as a powerful screening tool for the detection of patients with moderate to severe AS.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

References

1. Baumgartner H, Falk V, Bax JJ, De Bonis M, Hamm C, Holm PJ, Iung B, Lancellotti P, Lansac E, Rodriguez Munoz D, Rosenhek R, Sjogren J, Tornos Mas P, Vahanian A, Walther T, Wendler O, Windecker S, Zamorano JL, ESC Scientific Document Group. 2017 ESC/EACTS Guidelines for the management of valvular heart disease. *Eur Heart J* 2017;**38**:2739–2791.

2. Lancellotti P, Magne J, Dulgheru R, Clavel MA, Donal E, Vannan MA, Chambers J, Rosenhek R, Habib G, Lloyd G, Nistri S, Garbi M, Marchetta S, Fattouch K, Coisne A, Montaigne D, Modine T, Davin L, Gach O, Radermecker M, Liu S, Gillam L, Rossi A, Galli E, Ilardi F, Tastet L, Capoulade R, Zilberszac R, Vollema EM, Delgado V, Cosyns B, Lafitte S, Bernard A, Pierard LA, Bax JJ, Pibarot P, Oury C. Outcomes of patients with asymptomatic aortic stenosis followed up in heart valve clinics. *JAMA Cardiol* 2018;**3**:1060–1068.

3. Kang DH, Park SJ, Lee SA, Lee S, Kim DH, Kim HK, Yun SC, Hong GR, Song JM, Chung CH, Song JK, Lee JW, Park SW. Early surgery or conservative care for asymptomatic aortic stenosis. *N Engl J Med* 2020;**382**:111–119.

4. Strange G, Stewart S, Celermajer D, Prior D, Scalia GM, Marwick T, Ilton M, Joseph M, Codde J, Playford D; National Echocardiography Database of Australia contributing sites. Poor long-term survival in patients with moderate aortic stenosis. *J Am Coll Cardiol* 2019;**74**:1851–1863.

5. Ito S, Miranda WR, Nkomo VT, Boler AN, Pislaru SV, Pellikka PA, Crusan DJ, Lewis BR, Nishimura RA, Oh JK. Prognostic risk stratification of patients with moderate aortic stenosis. *J Am Soc Echocardiogr* 2021;**34**:248–256.

6. Kattoor AJ, Shanbhag A, Abraham A, Vallurupalli S. Clinical context and detection of the murmur of advanced aortic stenosis. *South Med J* 2018;**111**:230–234.

7. Carabello BA, Paulus WJ. Aortic stenosis. *Lancet* 2009;**373**:956–966.

8. Vranic II. Electrocardiographic appearance of aortic stenosis before and after aortic valve replacement. *Ann Noninvasive Electrocardiol* 2017;**22**:e12457.

9. Oh JK, Taliercio CP, Holmes DR Jr, Reeder GS, Bailey KR, Seward JB, Tajik AJ. Prediction of the severity of aortic stenosis by Doppler aortic valve area determination: prospective Doppler-catheterization correlation in 100 patients. *J Am Coll Cardiol* 1988;**11**:1227–1234.

10. 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.

11. Thaden JJ, Nkomo VT, Lee KJ, Oh JK. Doppler imaging in aortic stenosis: the importance of the nonapical imaging windows to determine severity in a contemporary cohort. *J Am Soc Echocardiogr* 2015;**28**:780–785.

12. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;**35**:1285–1298.

13. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019;**25**:70–74.

14. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, Pellikka PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Kapa S. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol* 2019;**12**:e007284.

15. SciPy.Org. scipy.signal.resample. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.resample.html (10 January 2020).

16. Huang G, Liu Z, Pleiss G, Van Der Maaten L, Weinberger K. Convolutional networks with dense connectivity. *IEEE Trans Pattern Anal Mach Intell* 2019; doi: 10.1109/TPAMI.2019.2918284.

17. Kuhn MJK. *Applied Predictive Modeling*. New York: Springer; 2013.

18. Keras Visualization Toolkit. https://raghakot.github.io/keras-vis/vis.visualization/#visualize_saliency (5 December 2020).

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–845.

20. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, Gachuhi N, Wilson B, Jaiswal MS, Befano B, Long LR, Herrero R, Einstein MH, Burk RD, Demarco M, Gage JC, Rodriguez AC, Wentzensen N, Schiffman M. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst* 2019;**111**:923–932.

21. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GC, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S. International evaluation of an AI system for breast cancer screening. *Nature* 2020;**577**:89–94.

22. Chiang SJ, Daimon M, Miyazaki S, Kawata T, Morimoto-Ichikawa R, Maruyama M, Ohmura H, Miyauchi K, Lee SL, Daida H. When and how aortic stenosis is first diagnosed: a single-center observational study. *J Cardiol* 2016;**68**:324–328.

23. Pellikka PA, Sarano ME, Nishimura RA, Malouf JF, Bailey KR, Scott CG, Barnes ME, Tajik AJ. Outcome of 622 adults with asymptomatic, hemodynamically

significant aortic stenosis during prolonged follow-up. *Circulation* 2005;**111**: 3290–3295.

24. Baumgartner H, Iung B, Otto CM. Timing of intervention in asymptomatic patients with valvular heart disease. *Eur Heart J* 2020;**41**:4349–4356.

25. Bing R, Cavalcante JL, Everett RJ, Clavel MA, Newby DE, Dweck MR. Imaging and impact of myocardial fibrosis in aortic stenosis. *JACC Cardiovasc Imaging* 2019;**12**:283–296.

26. Musa TA, Treibel TA, Vassiliou VS, Captur G, Singh A, Chin C, Dobson LE, Pica S, Loudon M, Malley T, Rigolli M, Foley JRJ, Bijsterveld P, Law GR, Dweck MR, Myerson SG, McCann GP, Prasad SK, Moon JC, Greenwood JP. Myocardial scar and mortality in severe aortic stenosis. *Circulation* 2018;**138**:1935–1947.

27. Ito S, Miranda WR, Nkomo VT, Connolly HM, Pislaru SV, Greason KL, Pellikka PA, Lewis BR, Oh JK. Reduced left ventricular ejection fraction in patients with aortic stenosis. *J Am Coll Cardiol* 2018;**71**:1313–1321.

28. Choi B, Lee S, Kim SM, Lee EJ, Lee SR, Kim DH, Jang JY, Kang SW, Lee KU, Chang EJ, Song JK. Dipeptidyl peptidase-4 induces aortic valve calcification by inhibiting insulin-like growth factor-1 signaling in valvular interstitial cells. *Circulation* 2017;**135**:1935–1950.

29. Lee S, Lee SA, Choi B, Kim YJ, Oh SJ, Choi HM, Kim EK, Kim DH, Cho GY, Song JM, Park SW, Kang DH, Song JK. Dipeptidyl peptidase-4 inhibition to prevent progression of calcific aortic stenosis. *Heart* 2020;**106**:1824–1831.

30. Bergmark BA, O'Donoghue ML, Murphy SA, Kuder JF, Ezhov MV, Ceska R, Gouni-Berthold I, Jensen HK, Tokgozoglu SL, Mach F, Huber K, Gaciong Z, Lewis BS, Schiele F, Jukema JW, Pedersen TR, Giugliano RP, Sabatine MS. An exploratory analysis of proprotein convertase subtilisin/kexin type 9 inhibition and aortic stenosis in the FOURIER trial. *JAMA Cardiol* 2020;**5**:709–713.

31. 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.

32. Hata E, Seo C, Nakayama M, Iwasaki K, Ohkawauchi T, Ohya J. Classification of aortic stenosis using ECG by deep learning and its analysis using grad-CAM. *Annu Int Conf IEEE Eng Med Biol Soc* 2020;**2020**:1548–1551.

33. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;**394**:861–867.

34. Marechaux S, Hachicha Z, Bellouin A, Dumesnil JG, Meimoun P, Pasquet A, Bergeron S, Arsenault M, Le Tourneau T, Ennezat PV, Pibarot P. Usefulness of exercise-stress echocardiography for risk stratification of true asymptomatic patients with aortic valve stenosis. *Eur Heart J* 2010;**31**: 1390–1397.