https://doi.org/10.1038/s41591-018-0240-2

Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram

Zachi I. Attia[1], Suraj Kapa[1], Francisco Lopez-Jimenez[1], Paul M. McKie[1], Dorothy J. Ladewig[2], Gaurav Satam[2], Patricia A. Pellikka[1], Maurice Enriquez-Sarano[1], Peter A. Noseworthy[1], Thomas M. Munger[1], Samuel J. Asirvatham[1], Christopher G. Scott[3], Rickey E. Carter[4] and Paul A. Friedman[1]*

Asymptomatic left ventricular dysfunction (ALVD) is present in 3–6% of the general population, is associated with reduced quality of life and longevity, and is treatable when found[1–4]. An inexpensive, noninvasive screening tool for ALVD in the doctor's office is not available. We tested the hypothesis that application of artificial intelligence (AI) to the electrocardiogram (ECG), a routine method of measuring the heart's electrical activity, could identify ALVD. Using paired 12-lead ECG and echocardiogram data, including the left ventricular ejection fraction (a measure of contractile function), from 44,959 patients at the Mayo Clinic, we trained a convolutional neural network to identify patients with ventricular dysfunction, defined as ejection fraction ≤35%, using the ECG data alone. When tested on an independent set of 52,870 patients, the network model yielded values for the area under the curve, sensitivity, specificity, and accuracy of 0.93, 86.3%, 85.7%, and 85.7%, respectively. In patients without ventricular dysfunction, those with a positive AI screen were at 4 times the risk (hazard ratio, 4.1; 95% confidence interval, 3.3 to 5.0) of developing future ventricular dysfunction compared with those with a negative screen. Application of AI to the ECG—a ubiquitous, low-cost test—permits the ECG to serve as a powerful screening tool in asymptomatic individuals to identify ALVD.

ALVD is present in 1.4–2.2% of the population (9% in the elderly) and is associated with reduced quality of life and increased morbidity and mortality[1]. Once ALVD is identified, medical treatments (angiotensin-converting enzyme inhibitors, angiotensin receptor, and beta blockers) and device implantation (implantable cardioverter-defibrillators and cardiac resynchronization systems) are effective in prevention of progression to symptomatic heart failure and reduce mortality[2–4]. While strategies for early identification of ALVD may prevent progression to symptomatic heart failure, a noninvasive and low-cost screening tool does not currently exist. As a result, several groups have sought to identify less costly and minimally invasive or noninvasive approaches to identifying patients with ALVD[5,6]. Currently, the best-studied test for screening is B-type natriuretic peptide (BNP) levels, but studies on BNP have been disappointing, and the test requires invasive blood draws[5,6].

AI using neural networks has been applied to sophisticated recognition of subtle patterns in digital data in numerous fields, including image recognition, self-driving automobiles, lesion identification in pathological specimens, speech recognition, language translation, and automated detection of mammographic lesions[7–11]. We hypothesized that the metabolic and structural derangements associated with the cardiomyopathic process would result in ECG changes that could be reliably detected by a properly trained neural network. To test this hypothesis, we created, trained, validated, and then tested a large neural network.

A total of 625,326 patients with paired ECG and transthoracic echocardiogram (TTE) were screened to identify the study cohort selected for analysis (Fig. 1). The first ECG–TTE data pair from patients with ECG and echocardiogram performed within a 2 week interval constituted the analysis data set, which consisted of 97,829 patients: 35,970 in the training set, 8,989 in the validation set, and 52,870 in the holdout testing set. No patient was in more than one group (Fig. 1). The overall patient population had a mean age of 61.8 ± 16.5 years, and 7.8% of the population had an ejection fraction (EF) of ≤35%. Table 1 shows patient characteristics for the training, validation, and test sets. In the testing data set, 4,131 patients (7.8%) had an EF of 35% or less, 6,740 patients (12.7%) had an EF greater than 35% and less than 50%, and 41,999 patients (79.5%) had an EF of 50% or higher. Over 89% of the TTEs were performed within 1 d of the index ECG.

In the test data (that is, data not used to train the algorithm), the algorithm provided a high degree of discrimination between EF ≦ 35% and EF > 35% (area under the curve (AUC), 0.93; Fig. 2a). When selecting a threshold with no preference for sensitivity, the overall accuracy was 85.7%, with a specificity of 85.7% and sensitivity of 86.3%, an F_1 score of 49.5%, and a negative predictive value of 98.7%. Using a threshold to yield a 90% sensitivity on the validation set and applying the algorithm to the testing data set, the sensitivity was 89.1%, specificity 83%, overall accuracy 83.5%, and negative predictive value 98.9%. When patients with no known comorbidities (Table 1) were separately analyzed by the network, the AUC increased to 0.98, with a sensitivity of 95.6%, specificity of 92.4%, negative predictive value of 99.8%, and accuracy of 92.5%. The identical AUCs among the training, validation, and test data sets demonstrate the robustness of the algorithm to different data sets. The network performance was strong across all age and sex groups (Fig. 2b); however, significant differences were noted in the strength of association ($P < 0.001$).

[1]Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA. [2]Business Development, Mayo Clinic, Rochester, MN, USA. [3]Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [4]Health Sciences Research, Mayo Clinic, Jacksonville, FL, USA. *e-mail: friedman.paul@mayo.edu

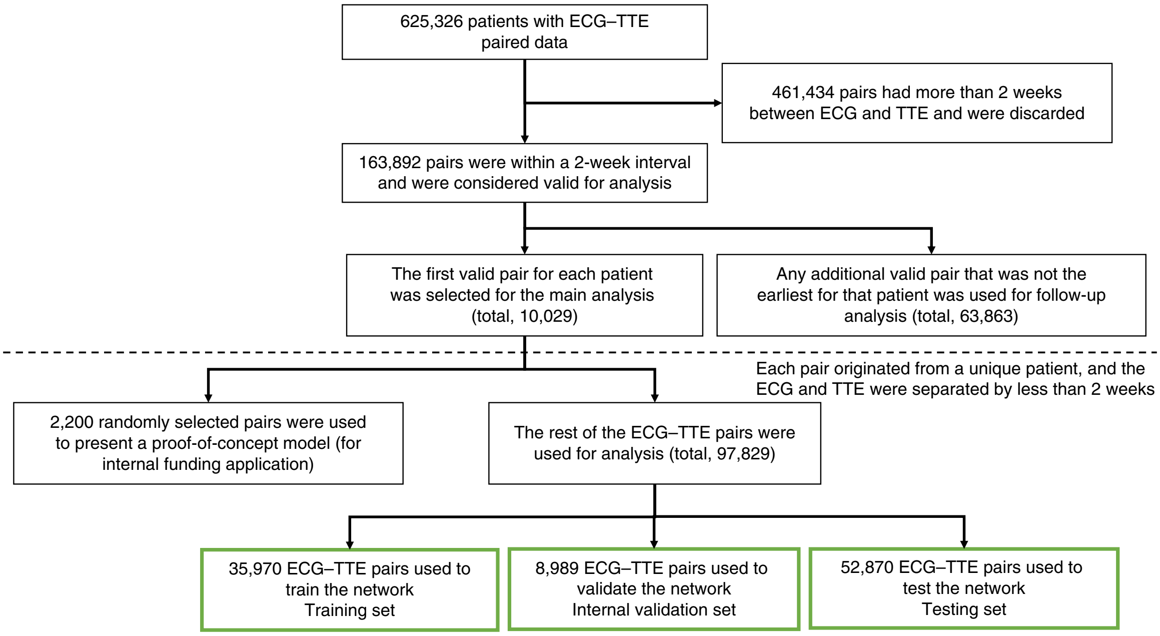
625,326 patients with ECG–TTE paired data
461,434 pairs had more than 2 weeks between ECG and TTE and were discarded
163,892 pairs were within a 2-week interval and were considered valid for analysis
The first valid pair for each patient was selected for the main analysis (total, 10,029)
Any additional valid pair that was not the earliest for that patient was used for follow-up analysis (total, 63,863)
Each pair originated from a unique patient, and the ECG and TTE were separated by less than 2 weeks
2,200 randomly selected pairs were used to present a proof-of-concept model (for internal funding application)
The rest of the ECG–TTE pairs were used for analysis (total, 97,829)
35,970 ECG–TTE pairs used to train the network Training set
8,989 ECG–TTE pairs used to validate the network Internal validation set
52,870 ECG–TTE pairs used to test the network Testing set

Fig. 1 | Creation of the study data sets. Schematic of the data set creation and analysis strategy, which was devised to assure a robust and reliable data set for training, validating, and testing of the network. Once a patient's data were placed in one of the data sets, that individual's data were used only in that set, avoiding 'cross-contamination' among the training, validation, and test data sets. Patients for whom data at more than one time point were available were used for the follow-up analysis, to determine whether an index abnormal AI screen with a normal EF (apparent false positive) was associated with future risk of a low EF. The details of the flow chart and how each of the data sets was used are described in the Methods. The final data sets used in the analysis for training, validating, and testing of the network are depicted as green boxes.

Table 1 | Patient characteristics and comorbidities at enrollment

	Training set (*n* = 35,970)	Validation set (*n* = 8,989)	Test set (*n* = 52,870)	*P* value
Age, years	61.6 (16.5)	61.8 (16.5)	61.8 (16.5)	0.44
Age groups, *n* (%)				0.86
<40	4,046 (11%)	1,008 (11%)	5,861 (11%)	
40–49	3,875 (11%)	942 (10%)	5,599 (11%)	
50–59	6,376 (18%)	1,587 (18%)	9,341 (18%)	
60–69	8,559 (24%)	2,110 (23%)	12,649 (24%)	
70–79	8,573 (24%)	2,158 (24%)	12,550 (24%)	
80+	4,541 (13%)	1,184 (13%)	6,870 (13%)	
Sex, *n* (%)				0.64
Female	15,358 (43%)	3,821 (43%)	22,704 (43%)	
Male	20,612 (57%)	5,168 (57%)	30,166 (57%)	
Mean EF	56.3 (11.9)	56.1 (12.1)	56.2 (12.0)	0.45
Heart failure, *n* (%)	10,365 (20%)	7,003 (19%)	1,803 (20%)	0.45
Diabetes mellitus, *n* (%)	8,458 (24%)	2,079 (23%)	12,433 (24%)	0.71
Hypercholesterolemia, *n* (%)	15,593 (43%)	3,851 (43%)	23,059 (44%)	0.35
Renal disease, *n* (%)	6,929 (19%)	1,685 (19%)	10,219 (19%)	0.43
Hypertension, *n* (%)	16,831 (47%)	4,163 (46%)	24,643 (47%)	0.69
Coronary artery disease, *n* (%)	13,563 (38%)	3,380 (38%)	20,040 (38%)	0.77
Myocardial infarction, *n* (%)	4,556 (13%)	1,111 (12%)	6,770 (13%)	0.48

Enrollment refers to the time of the initial ECG–TTE data pair acquisition. Numbers in parentheses refer to the standard deviation (for age and sex) or to the percentage of individuals randomly assigned to a given subset (all other rows). *P* values refer to two-tailed analysis of variance or χ^2 tests for a difference in the distribution of values between the training, validation, and test sets.

When selecting a threshold with no preference for sensitivity (that is, a threshold that will yield an equal sensitivity and specificity using validation data), 10,544 patients (19.9%) in the test set were identified by the network as having a low EF. Of these, 33.8% had an EF of 35% or less, 29.5% had an EF of 36–50%, and 36.6% had a normal EF. In the group that the network identified as normal, 1.3% had an EF of 35% or less and 8.6% had EF of 36–50%; the rest (90.1%) had a normal EF (Supplementary Fig. 1).

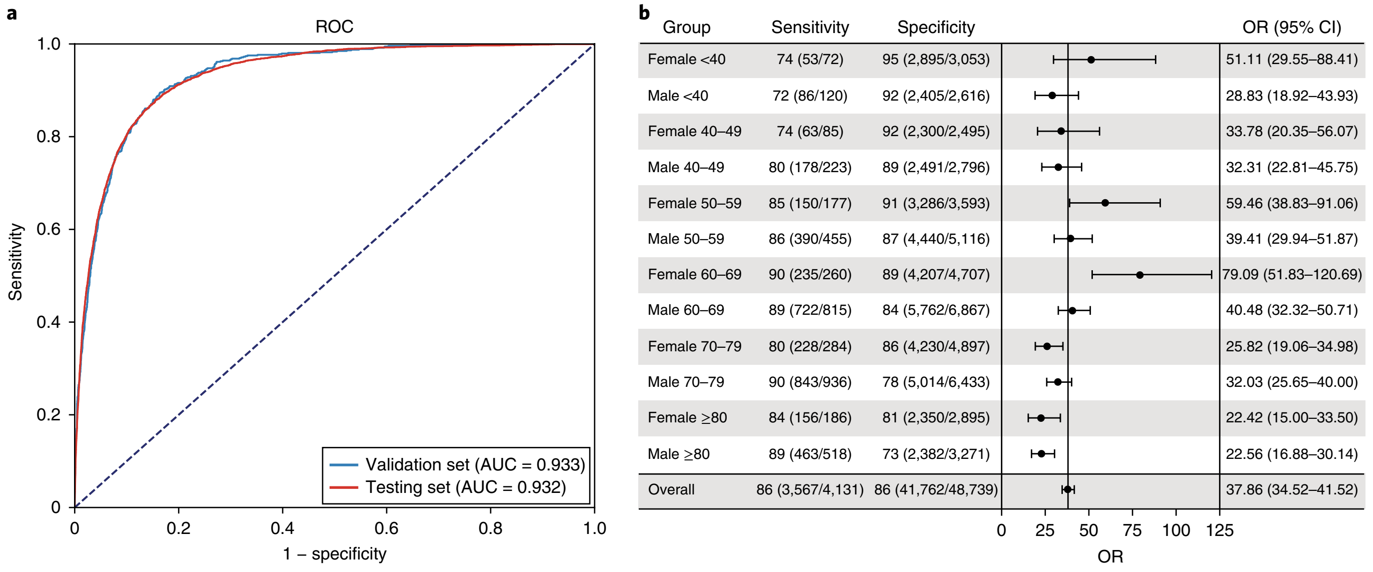

Fig. 2 | Network ROC and sensitivity and specificity across age and gender subsets. a, The ROC of the convolutional neural network used to identify patients with an EF of $\leq$35%. The ROC curve and AUC were calculated using the validation and testing (holdout) data sets. **b**, The convolutional neural network's sensitivity and specificity to detect EF $\leq$ 35% is tabulated across a range of age and gender combinations. The diagnostic OR, which is the ratio of positive likelihood ratio (sensitivity / (1 – specificity)) to the negative likelihood ratio ((1 – sensitivity) / specificity), as well as the associated 95% CI, is shown for each age and gender combination and for the overall study sample. The Breslow–Day test of homogeneity showed significant variation across the age and gender combinations ($P < 0.0001$).

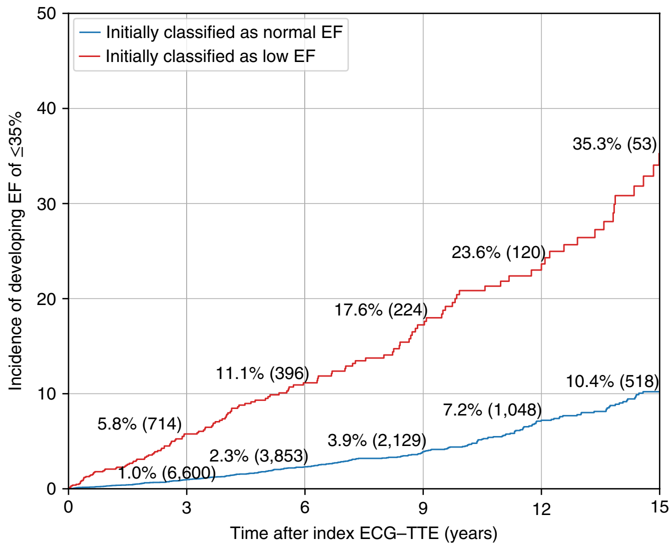

Fig. 3 | Long-term incidence of developing an EF of $\leq$35% in patients with an initially normal EF stratified by AI classification. Long-term outcome of patients with an echocardiographic EF of $\geq$50% at the time of initial classification, stratified by the initial network classification. The ordinate shows the cumulative incidence of developing a low EF (35%), and the abscissa indicates years from the time of index ECG–TTE evaluation. A fourfold risk of future low EF was present when the AI algorithm defined the ECG as abnormal (age- and sex-adjusted HR, 4.1 (95% CI, 3.3–5.0), $P < 0.001$), compared with patients with an echocardiographic normal EF who were classified as having a normal EF by the ECG network. The numbers reported along the cumulative incidence curves reflect the estimated cumulative incidence (and number at risk, in parentheses) for each group at the times indicated along the axis.

Of the patients identified by the network as having a normal EF who also had a confirmatory normal contemporaneous EF by echocardiography ('true negative'), 11,515 had a follow-up echocardiogram. Of these true negative patients, 302 developed a low EF over a median follow-up period of 3.8 years (interquartile range, 1.4–7.5 years) (Fig. 3, 1.8% and 4.4% for 5 and 10 year incidence, respectively). In contrast, 1,335 patients were labeled by the network as having a low EF, but the contemporaneous echocardiogram demonstrated a normal EF ('false positive'). Of these patients with an initial false positive result, 147 developed left ventricular dysfunction during a median follow-up period of 3.4 years (interquartile range, 1.2–6.8) (Fig. 3, 9.5% and 20.8% for 5 and 10 year incidence, respectively). This represented a fourfold risk of future low EF when the AI algorithm defines the ECG as abnormal (age and sex adjusted hazard ratio (HR), 4.1 (95% CI, 3.3–5.0), $P < 0.001$), suggesting that the network identified ECG abnormalities before overt ventricular dysfunction became manifest.

Left ventricular systolic dysfunction is associated with impaired quality of life, increased morbidity, and increased mortality[1]. Furthermore, ALVD affects >7 million people in the United States, and many more individuals globally. Major cardiovascular professional societies have endorsed evidence-based therapies that improve symptoms and survival once left ventricular dysfunction is detected[12,13]. However, effective population-wide screening for ventricular dysfunction is lacking[1,6]. We found that the application of AI using a convolutional neural network to the standard 12-lead ECG—an inexpensive, widely available, common clinical test—enabled detection of left ventricular dysfunction with an AUC of 0.93. The performance of this test compares favorably with other common screening tests, such as prostate-specific antigen for prostate cancer (AUC, 0.92), mammography for breast cancer (AUC, 0.85), and cervical cytology for cervical cancer (AUC, 0.70). In contrast to BNP, accuracy was not affected by age or sex. Importantly, in addition to effectively identifying individuals with ventricular systolic dysfunction, the network also identified those patients with initially normal left ventricular function who are at risk of subsequently developing a low EF. Such 'false positive' patients, with an abnormal network screen but a normal EF, had a fourfold increased risk of developing ventricular dysfunction over the next 5 years (10% risk at 5 years). This suggests the network detected early, subclinical, metabolic or structural abnormalities that manifest in the

ECG. Whether this group would benefit from serial screening or medical therapy to prevent the development of ventricular dysfunction, however, is unknown.

Congestive heart failure afflicts more than 5 million people and consumes more than $30 billion in health-care expenditures annually in the United States alone[14,15]. Early detection and prevention are a health-care imperative. We found that 6% of patients in our population had low EF, consistent with previous studies[1]. BNP and N-terminal pro b-type natriuretic peptide (NTproBNP) have been proposed for detection of left ventricular systolic dysfunction. Bhalla et al.[16] assessed BNP to screen for systolic and diastolic dysfunction and found an AUC of 0.60 for BNP and 0.70 for NTproBNP; results improved with the addition of impedance cardiography to 0.70 and 0.73, respectively. A Mayo Clinic study from Olmsted County assessing individuals aged 45 and older found the AUC was greater for individuals with more severe systolic dysfunction (0.82 to 0.92) than than for those with any systolic dysfunction (0.51 to 0.74)[17]. Moreover, optimal discriminatory levels for BNP varied with age and sex. In contrast, we found excellent AI network performance across all age and sex strata. This capability appears to be unique to the neural network screen. The specific ECG characteristics used by our convolutional neural network to classify individuals as having or lacking a low EF are not known, owing to the nature of neural networks, although we presume it is detecting the known pathological effects of heart failure on the ECG (Supplementary Fig. 2). However, by training the network with a large cohort of approximately 45,000 ECG and EF data pairs, the network was exposed to a sufficient number of ECG variants to classify with certainly individuals with a low EF, as demonstrated by the AUC of 0.93 obtained when testing in a population of 52,870 individuals. In contrast to previous applications of neural networks in medicine, such as for identification of mammographic lesions[9], we expand the use of AI to extend beyond the capacity of human skills.

Another important characteristic of our network is that it uses an inexpensive, standardized, ubiquitous test as its input—the 12-lead ECG. In many rural areas in the United States[18] and in developing countries, access to cardiological care and imaging is limited. The availability of a portable, inexpensive test for ventricular systolic dysfunction permits optimal utilization of limited imaging resources, while also enabling individuals to benefit from early institution of effective therapies such as beta adrenergic blockers, angiotensin receptor antagonists, and, where available, implantable devices (defibrillators and cardiac resynchronization systems)[13,19]. With the emergence of smartphone-enabled electrodes[20], mobile applications may permit ECG use in resource-constrained regions. The software-based nature of test 'samples' for our network also enables continuous feedback and refinement, with rapid distribution of system improvements.

Our work is best understood in the context of its limitations. The accuracy, sensitivity, and specificity of the networks were all excellent, but the test's positive predictive value was only 33.8%. This is in part because we selected an EF cutoff of 35%. We selected this detection threshold owing to the well-established outcome and therapeutic implications of this value[19]. However, identification of an EF of <50% is still clinically significant, as this reflects an abnormal EF. When considering a higher cut-off for abnormal function of <50%, the positive predictive value of the network was 63.4%. Of the 'false positives', 45% had an EF of <50% but >35% (29.5% of all positives). While statistically patients with an EF in this range are considered false positives, medically this value is actionable, and an echocardiogram is justified. An interesting finding was that in patients with a false positive result with a network-predicted low EF and an echocardiographic normal EF at the time of screening—a group that assessed with current technology would be considered healthy—there was a fourfold increase in risk of developing a low EF in the future compared with patients with a negative AI screen.

Another limitation is that the ECG–TTE pairs were not simultaneously acquired. However, since we included over 100,000 ECG–TTE paired data sets with more than 89% of TTEs performed within 24 h of the ECG, the likelihood of inaccuracy related to temporal delay is small.

In conclusion, applying AI to the ECG—a ubiquitous, low-cost test—permits the ECG to serve as a powerful tool to screen for left ventricular dysfunction and furthermore to identify individuals at increased risk for its development in the future.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0240-2.

Received: 15 March 2018; Accepted: 1 October 2018;
Published online: 7 January 2019

References

1. McDonagh, T. A., McDonald, K. & Maisel, A. S. Screening for asymptomatic left ventricular dysfunction using B-type natriuretic Peptide. *Congest. Heart Fail.* **14**, 5–8 (2008).
2. Dargie, H. J. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. *Lancet* **357**, 1385–1390 (2001).
3. Pfeffer, M. A. et al. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. *N. Engl. J. Med.* **327**, 669–677 (1992).
4. Priori, S. G. et al. 2015 ESC guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: the Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). *Eur. Heart J.* **36**, 2793–2867 (2015).
5. Betti, I. et al. The role of N-terminal PRO-brain natriuretic peptide and echocardiography for screening asymptomatic left ventricular dysfunction in a population at high risk for heart failure. The PROBE-HF study. *J. Card. Fail.* **15**, 377–384 (2009).
6. Redfield, M. M. et al. Plasma brain natriuretic peptide to detect preclinical ventricular systolic or diastolic dysfunction: a community-based study. *Circulation* **109**, 3176–3181 (2004).
7. Kim, J. H., Kwon, H. S. & Seo, H. W. Evaluating a pivot-based approach for bilingual lexicon extraction. *Comput. Intell. Neurosci.* **2015**, 434153 (2015).
8. Pasquier, M., Quek, C. & Toh, M. Fuzzylot: a novel self-organising fuzzy-neural rule-based pilot system for automated vehicles. *Neural Netw.* **14**, 1099–1112 (2001).
9. Salazar-Licea, L. A., Pedraza-Ortega, J. C., Pastrana-Palma, A. & Aceves-Fernandez, M. A. Location of mammograms ROI's and reduction of false-positive. *Comput. Methods Programs Biomed.* **143**, 97–111 (2017).
10. Wingfield, C. et al. Relating dynamic brain states to dynamic machine states: human and machine solutions to the speech recognition problem. *PLoS Comput. Biol.* **13**, e1005617 (2017).
11. Yoshida, H. et al. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastric Cancer* **21**, 249–257 (2018).
12. Al-Khatib, S. M. et al. 2017 AHA/ACC/HRS guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: executive summary. A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. *Circulation* **138**, e210–e271 (2018).
13. Yancy, C. W. et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **62**, e147–e239 (2013).
14. Heidenreich, P. A. et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. *Circ. Heart Fail.* **6**, 606–619 (2013).
15. Mozaffarian, D. et al. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. *Circulation* **131**, e29–e322 (2015).
16. Bhalla, V. et al. Diagnostic ability of B-type natriuretic peptide and impedance cardiography: testing to identify left ventricular dysfunction in hypertensive patients. *Am. J. Hypertens.* **18**, 73S–81S (2005).

17. Costello-Boerrigter, L. C. et al. Amino-terminal pro-B-type natriuretic peptide and B-type natriuretic peptide in the general community: determinants and detection of left ventricular dysfunction. *J. Am. Coll. Cardiol.* **47**, 345–353 (2006).
18. Gruca, T. S., Pyo, T. H. & Nelson, G. C. Providing cardiology care in rural areas through visiting consultant clinics. *J. Am. Heart Assoc.* **5**, e002909 (2016).
19. Yancy, C. W. et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *J. Am. Coll. Cardiol.* **70**, 776–803 (2017).
20. Yasin, O. Z. et al. Noninvasive blood potassium measurement using signal-processed, single-lead ECG acquired from a handheld smartphone. *J. Electrocardiol.* **50**, 620–625 (2017).

Acknowledgements

The study was conceived, funded, and executed entirely by Mayo Clinic. There was no industry support of any kind.

Author contributions

D.J.L. and G.S. contributed to the literature search, study coordination, data management, and data collection. Z.I.A., P.A.F., S.K., and F.L.-J. contributed to the study design. C.G.S., Z.I.A., R.E.C., and F.L.-J. contributed to the data analysis. C.G.S., Z.I.A., R.E.C., P.A.N., P.A.P., M.E.-S., P.A.F., S.K., P.M.M., T.M.M., and S.J.A. contributed to data interpretation. Z.I.A., R.E.C., F.L.-J., P.A.F., P.A.N., and P.A.P. contributed to the writing of the manuscript. R.E.C., P.A.N., P.A.P., M.E.-S., F.L.-J., S.K., P.M.M., T.M.M., and S.J.A contributed to the critical review and editing.

Competing interests

Mayo Clinic has licensed the underlying technology to EKO, a maker of digital stethoscopes with embedded ECG electrodes. Mayo Clinic may receive financial benefit from the use of this technology, but at no point will Mayo Clinic benefit financially from its use for the care of patients at Mayo Clinic. P.A.F., F.L.-J., S.K., and Z.I.A. may also receive financial benefit from this agreement.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0240-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.A.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data sources and study population. Following institutional review board approval, we obtained data from the Mayo Clinic digital data vault. We identified 163,892 adult patients (18 years old or older) with at least one digital, standard 10-s 12-lead ECG acquired in the supine position between January 1994 and February 2017 and at least one TTE obtained within 14 d of the index ECG (Fig. 1). For patients with multiple ECG and TTE data sets meeting these criteria, the earliest pair was used for network creation, validation, or testing, and subsequent TTE data were used for analysis of follow-up. We performed a preliminary proof-of-concept assessment to obtain internal funding using 2,200 ECG–TTE data pairs, which were excluded from the present analysis, leaving a cohort of 97,829 patients whose first ECG–TTE paired data sets were used for the primary analysis.

ECGs were acquired at a sampling rate of 500 Hz using a Marquette ECG machine (GE Healthcare) and stored using the MUSE data management system for later retrieval. Comprehensive 2-D or 3-D and Doppler echocardiography were available for all patients. Quantitative data were recorded at the time of the acquisition in a Mayo Clinic–developed, custom database (Echo Image Management System). EF is routinely measured or estimated using standardized methodologies, and in most reports, more than one value may be recorded. For the purpose of this study, the EF value used in the models was the first available from a standard hierarchical sequence: EF determined using 3-D echocardiography[21], a biplane approach using the Simpson method, a 2-D method[22], or M-mode and, in the absence of any of the preceding, the reported visually estimated EF. If the estimation was a range, we used the midpoint as a single EF value. EF was classified as low (≤35%), mildly depressed (35–49%), or normal (≥50%).

Primary and secondary outcomes. The primary outcome was the ability of the AI network to identify patients with an EF of 35% or less using the ECG signal alone. This value was selected owing to its clear-cut clinical and therapeutic importance[23]. The secondary outcome was the ability of the AI network to identify individuals with a normal EF at the time of screening, but with an increased risk of subsequent low EF during follow-up.

Overview of AI model development. We implemented a convolutional neural network using the Keras framework with a Tensorflow (Google) backend and Python[24]. Convolutional neural networks, which are commonly applied to images, operate such that the convolutions can be used to extract very subtle patterns in a data set. Each 12-lead ECG was considered a 12 × 5,000 (that is, 12 leads by 10-s duration sampled at 500 Hz) matrix, where the first dimension represents a spatial dimension and the second represents a temporal one[24]. We used the internal validation set to optimize the network architecture, and hyperparameters such as batch and step size. Multiple networks were tested, and the network that yielded the highest AUC of the receiver operator curve (ROC) for the validation data set was selected. In cases in which more than one network had similar results when tested using the validation set, the 'simpler' network—the one with fewer parameters or layers—was selected[24].

ECG analysis is mostly a visual task. While the signal is a time series, it is pseudocyclical, and its main features are morphologic[25–27]. To enable detection of patterns in these features, we used architectures that were based on convolutional layers for feature extraction.

The main architecture tested was a convolutional neural network[28,29] in which the convolutional blocks were followed by two fully connected layers. The selected network (Supplementary Fig. 3) was composed of six convolutional layers, each of which was followed by a nonlinear 'Relu' activation function, a batch-normalization layer[30], and a max-pooling layer[31]. As these first six layers were designed to learn features within each lead, each convolutional layer had a $K \times 1$ filter shape with N filters. K started at 5 and was decreased to 3 after the third layer, while N started at 16 and was increased to 32 in the third layer and to 64 in the fifth layer. After each 'temporal' layer, there was a Relu activation layer, a batch-normalization layer, and max-pooling layer (4 × 1 after the third and fifth layers and 2 × 1 after other temporal convolutional layers). Following the last temporal convolutional layer, a 'spatial' convolutional layer was used. In this layer, the filter shape was 12 × 1, allowing it to fuse data from the different leads (hence the term 'spatial'). The reason a convolutional filter as opposed to a pooling layer was used for lead fusion is that the ECG is structured and the spatial structure of a 12-lead ECG is consistent among patients and adjusted per patient[32,33]. After this spatial layer, the data were fed to a fully connected network with two hidden layers (64 and 32 nodes) followed by Relu activation layers, batch-normalization layers, and dropout layers to avoid overfitting. The output layer had two classes and was activated using the 'Softmax' function[34].

AI model training. Owing to the large size of our data set, we used approximately 50% of the data set for training the network. This left us with a very large data set to test the network to better assess its robustness (Fig. 1). After the initial split into the development and testing (holdout) data sets, the development data set was further divided into the training data (80% of the development set) and the internal validation data (20%). As we filtered the data set to include only one ECG per patient, none of the patients used for model development were used for validation or testing.

For training, ECGs were fed to the network, and the network weights were updated using the Adam optimize[29] with categorical cross-entropy as the loss function. After each epoch, the network was tested using the internal validation data set, and training was stopped after it was optimized. The network hyperparameters were also tuned during this process, and the network with the lowest loss value was selected.

Optimizing the architecture and hyperparameters. The convolutional neural network framework provides immense flexibility with structure, and as a result, turning the algorithm involves modification of both the architecture and the hyperparameters associated with each aspect of the model specification. We evaluated combinations of architecture and hyperparameters using an iterative process, including a formalized grid search for hyperparameters. In terms of architecture, we evaluated the change in model performance based on adding neurons to the fully connected layers. Increasing the number of neurons yielded lower performance (for example, 0.912 using 256 hidden nodes). Removing two of the pooling layers to allow more temporal features (by a factor of 16 overall) and replacing the fully connected layers with long short term memory (LSTM) layers to allow real temporal feature detection (128 and 32 LSTM cells in each layer) did not significantly improve the validation AUC (0.936) and increased the number of parameters (to 560,000).

To exploit the pseudocyclical nature of the signal and reduce the number of parameters, we also tested the same architecture on a shorter segment of 2 s (1,000 samples zero-padded to 1,024) with an overlap of 1 s (500 samples). While the AUC per segment was lower (0.91), the average score of all nine segments (due to the overlap) per ECG yielded an AUC of 0.934. The architecture reduced the number of parameters by half, to 159,000, which helped optimize the training. All results presented in this manuscript were created using this specific architecture. Computer code is available upon request.

AI-augmented ECG to identify a low EF. After selecting the optimal network using validation data, we created an ROC using the same validation set and measured its AUC as a primary assessment of network strength. We used the validation data set ROC to select two thresholds for the probability of having a low EF: the first was selected by giving an equal weight to sensitivity and specificity, and the second was selected to yield a sensitivity of 90% on the validation data set. The convolutional neural network model was then used on the test data to test its ability to predict a low EF. The two thresholds were used to calculate sensitivity, specificity, and accuracy in the test data, which had not been used for model training or threshold selection. To investigate network performance differences by age and sex, odds ratios (ORs) were calculated within groups. The Breslow–Day test was used to test for homogeneity of ORs across the groups defined on age and sex. Sensitivity and specificity were also calculated within these groups.

AI-augmented ECG to predict a future low EF. We hypothesized that early in the course of any disease that impacts EF, ECG signals would show subtle abnormal patterns due to metabolic and structural derangements that had not affected a sufficient quantity of myocardium to cause a reduction in EF. We further hypothesized that the convolutional neural network would classify some of these cases as abnormal, giving the initial appearance of a false positive test (that is, an individual classified as having a low EF but reported as normal) that with time would become a true positive test. To test this hypothesis, we designed a substudy to identify patients that met the following conditions: (i) the network predicted the patient had a low EF; (ii) the individual had an echocardiogram performed within 14 d that demonstrated a normal EF (≥50%), indicating a false positive finding by the algorithm; and (iii) the individual had at least one additional echocardiogram (not used for training or testing) available at a later date. A control group (EF ≥ 50%) was created using the true negative cases (algorithm and clinical determinations were both consistent with not having a low EF). A Kaplan–Meier analysis was used to depict the incidence of low EF for the true negatives versus the false positives over time. Subsequently, Cox proportional hazards regression was used to estimate the hazard for low EF after adjusting for age and sex.

Statistical considerations. For measures of diagnostic performance (AUC of ROC, sensitivity), the sample sizes are so large that normal CIs are expected to have a width of <0.5%. As such, the CIs are not reported alongside the estimated values due to the values' high precision. Continuous data are presented as mean ± s.d. Cox models are presented with two-sided P values without any correction for multiple testing. The effect size of the Cox model is represented by the HR and its associated 95% CI. Survival analyses were conducted using SAS version 9.4. The convolutional neural network was trained using Keras (version 2.0.3) and TensorFlow (version 1.0.1).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Programming code related to the data preprocessing and Keras model specification will be made available under the GNU General Public License version 3 upon request to Z.I.A. (attia.itzhak@mayo.edu).

Data availability

All requests for raw and analyzed data and related materials, excluding programming code, will be reviewed by the Mayo Clinic legal department and Mayo Clinic Ventures to verify whether the request is subject to any intellectual property or confidentiality obligations. Requests for patient-related data not included in the paper will not be considered. Any data and materials that can be shared will be released via a Material Transfer Agreement.

References

21. Yamani, H., Cai, Q. & Ahmad, M. Three-dimensional echocardiography in evaluation of left ventricular indices. *Echocardiography* **29**, 66–75 (2012).
22. Quinones, M. A. et al. A new, simplified and accurate method for determining ejection fraction with two-dimensional echocardiography. *Circulation* **64**, 744–753 (1981).
23. Russo, A. M. et al. ACCF/HRS/AHA/ASE/HFSA/SCAI/SCCT/SCMR 2013 appropriate use criteria for implantable cardioverter-defibrillators and cardiac resynchronization therapy: a report of the American College of Cardiology Foundation appropriate use criteria task force, Heart Rhythm Society, American Heart Association, American Society of Echocardiography, Heart Failure Society of America, Society for Cardiovascular Angiography and Interventions, Society of Cardiovascular Computed Tomography, and Society for Cardiovascular Magnetic Resonance. *J. Am. Coll. Cardiol.* **61**, 1318–1368 (2013).
24. van Rossum, G. *Python Tutorial, Technical Report CS-R9526* (CWI, Amsterdam, 1995).
25. Gholam-Hosseini, H. & Nazeran, H. Detection and extraction of the ECG signal parameters. In *Proc. 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 127–130 (IEEE, 1998).
26. Sugrue, A. et al. Identification of concealed and manifest long QT syndrome using a novel T wave analysis program. *Circ. Arrhythm. Electrophysiol.* **9**, e003830 (2016).
27. Couderc, J. P. et al. T-wave morphology abnormalities in benign, potent, and arrhythmogenic I_{kr} inhibition. *Heart Rhythm* **8**, 1036–1043 (2011).
28. Krizhevsky A., S I., Hinton G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105 (Neural Information Processing Systems, 2012).
29. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
30. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/abs/1502.03167 (2015).
31. Nagi, J. et al. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *Proc. 2011 IEEE International Conference on Signal and Image Processing Applications* 342–347 (IEEE, 2011).
32. Wilson, F. N. et al. The precordial electrocardiogram. *Am. Heart. J.* **27**, 19–85 (2004).
33. Khan, G. M. A new electrode placement method for obtaining 12-lead ECGs. *Open Heart* **2**, e000226 (2015).
34. Cristianini, N. & Shawe-Taylor, J. An Introduction to Support Vector Machines and other Kernel-based Learning Methods (Cambridge University Press, New York, 2000).

nature research

Corresponding author(s): NMED-A90984A

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection: All data utilized in this study were electronically retrieved. SAS, Python, SQL and other core systems of our data warehouse were utilized.

Data analysis: SAS version 9.4 (Phreg and general descriptive analyses); Keras version 2.0.3 and TensorFlow 1.0.1 (CNN training); Python 2.7

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Computer code is available upon request. The raw patient data is not publicly available to institutional policy and human subjects approval.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study does not follow the traditional null hypothesis testing strategy that lends itself to power calculation (i.e., there is no null hypothesis for the algorithm). Instead, we abstracted all of the cases that met inclusion criteria and utilized measures of statistical precision. For sample sizes over 2000, which is this case for this study, precision of proportions (i.e., AUC, accuracy, sensitivity, etc) have precision <1 percentage point. Thus, the sample size of the study was sufficiently sized to support the research objectives.
Data exclusions	This study design required that the electrocardiogram (ECG) and echocardiogram (echo) be obtained within two weeks. As a result, 461,434 of the available patients with both an ECG and echo were excluded. Further eliminations were utilized to ensure that each patient was only present in the dataset once. This further eliminated 63,863 potentially eligible ECG-echo pairings. With the selection of the first observed paired assessment, a total of 97,829 were available for the primary analysis
Replication	The study design included separate training and test data. The test data was not used to determine model estimates. The replication of the findings from the training data was tested on this withheld data. In addition, a validation dataset was used to optimize model configuration as the model was being trained. This validation data was not utilized to assess the replication of the algorithm.
Randomization	This was a retrospective chart review study. There was no intervention or treatment assignment as a result. Therefore, randomization is not applicable.
Blinding	As noted in the randomization section above, there was no randomization or treatment assignment. Accordingly, blinding was not applicable to this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Unique biological materials
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☐	☒	Human research participants

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	This study population consisted of patients 18 years of age or older seen at Mayo Clinic in Rochester, MN. Eligible patients were required to have at least one digital, standard 10-second 12-lead ECG acquired in the supine position between January 1994 and February 2017 and at least one transhoracic echocardiogram obtained within 14 days of the ECG. The overall patient population had a mean age of 61.8+/- 16.5 years, and 7.8% of the population had a left ventricular ejection fraction of <35%
Recruitment	N/A - records based research