nature medicine

LETTERS

https://doi.org/10.1038/s41591-020-0870-z

Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network

Sushravya Raghunath[1], Alvaro E. Ulloa Cerna[1], Linyuan Jing[1], David P. vanMaanen[1], Joshua Stough[1,2], Dustin N. Hartzel[3], Joseph B. Leader[3], H. Lester Kirchner[4], Martin C. Stumpe[5], Ashraf Hafez[5], Arun Nemani[5], Tanner Carbonati[5], Kipp W. Johnson[5], Katelyn Young[6], Christopher W. Good[7], John M. Pfeifer[8], Aalpen A. Patel[9], Brian P. Delisle[10], Amro Alsaid[7], Dominik Beer[7], Christopher M. Haggerty[1,7,11] and Brandon K. Fornwalt[1,7,9,11] ✉

The electrocardiogram (ECG) is a widely used medical test, consisting of voltage versus time traces collected from surface recordings over the heart[1]. Here we hypothesized that a deep neural network (DNN) can predict an important future clinical event, 1-year all-cause mortality, from ECG voltage–time traces. By using ECGs collected over a 34-year period in a large regional health system, we trained a DNN with 1,169,662 12-lead resting ECGs obtained from 253,397 patients, in which 99,371 events occurred. The model achieved an area under the curve (AUC) of 0.88 on a held-out test set of 168,914 patients, in which 14,207 events occurred. Even within the large subset of patients ($n = 45,285$) with ECGs interpreted as 'normal' by a physician, the performance of the model in predicting 1-year mortality remained high (AUC = 0.85). A blinded survey of cardiologists demonstrated that many of the discriminating features of these normal ECGs were not apparent to expert reviewers. Finally, a Cox proportional-hazard model revealed a hazard ratio of 9.5 ($P < 0.005$) for the two predicted groups (dead versus alive 1 year after ECG) over a 25-year follow-up period. These results show that deep learning can add substantial prognostic information to the interpretation of 12-lead resting ECGs, even in cases that are interpreted as normal by physicians.

The prediction of risk is fundamental to the practice of medicine. Within cardiovascular medicine, for example, risk scoring systems are widely used to support a variety of important individual patient care decisions, such as determining the intensity of medical therapy, helping to decide whether invasive therapies are warranted, deciding on the use of advanced heart failure therapies and guiding an ischemic workup before elective surgery[2–7] (Supplementary Table 1). Risk prediction is also important at the population level, particularly within accountable care organizations and other value-based care models. These systems must find ways to reduce costs while maintaining quality[8], and risk prediction is critical to this effort by helping to optimally direct resources to the patients who need them most. Despite these needs, existing risk prediction tools are far from perfect[9], and cardiovascular outcomes could be improved with better risk prediction.

One option for improving risk prediction in cardiovascular disease is through enhanced use of the 12-lead ECG[10,11], which is one of the most widely used cardiovascular diagnostic tests. However, despite widespread use, the ECG has not been well adopted as a prognostic tool[11]. In fact, most cardiac scoring systems do not use the ECG as a variable. Those that do (TIMI[2] and GRACE[3], for example) make very limited use, assessing only for the presence of ST segment deviation. Automated approaches to analyzing ECG data to provide enhanced prognostic capabilities may therefore have tremendous impact on cardiovascular disease outcomes.

There has been steady improvement in ECG signal processing over several decades[12,13], with automated techniques for feature extraction[14–16], morphology detection for heartbeat classification[17] and diagnostic capabilities for a range of conditions, such as arrhythmias[18,19]. Recently, the emergence of large clinically acquired ECG datasets combined with exponential growth in computational power and improvements in DNNs has enabled considerable advancement in the automated interpretation of ECGs[20–22]. Machine learning methods, including neural networks, have been explored for automated and accurate measurement of intervals and for feature extraction[23–26]. Deep learning in particular has recently shown promise for diagnosing abnormal heart rhythms[27], identifying acute findings[28], identifying asymptomatic cardiac dysfunction[29,30] and detecting the electrocardiographic signature of paroxysmal atrial fibrillation in patients currently in sinus rhythm[31]. However, there are no reports of an automated method to predict clinically relevant future events, such as short-term mortality, directly from ECGs.

We hypothesized that a DNN could learn novel features in resting 12-lead ECG voltage–time data to directly predict 1-year mortality. We leveraged nearly 2.3 million ECGs (an order of magnitude more than in previous studies) and DNNs to show that this hypothesis is

[1]Department of Translational Data Science and Informatics, Geisinger, Danville, PA, USA. [2]Department of Computer Science, Bucknell University, Lewisburg, PA, USA. [3]Phenomic Analytics and Clinical Data Core, Geisinger, Danville, PA, USA. [4]Department of Population Health Sciences, Geisinger, Danville, PA, USA. [5]Tempus Labs, Inc., Chicago, IL, USA. [6]Department of Internal Medicine, Geisinger, Danville, PA, USA. [7]Heart Institute, Geisinger, Danville, PA, USA. [8]Heart and Vascular Center, Evangelical Hospital, Lewisburg, PA, USA. [9]Department of Radiology, Geisinger, Danville, PA, USA. [10]Department of Physiology and Cardiovascular Research Center, University of Kentucky, Lexington, KY, USA. [11]These authors contributed equally: Christopher M. Haggerty, Brandon K. Fornwalt. ✉e-mail: bkf@gatech.edu

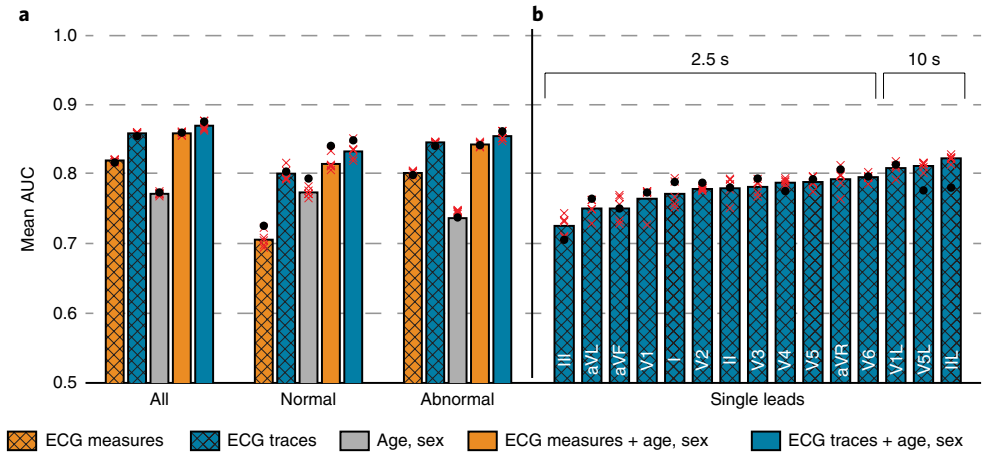


Fig. 1 | Summary of model performance as area under the receiver operating characteristic curve for predicting 1-year mortality. a, The mean AUC for the indicated input data, including (i) clinically acquired ECG measures (9 numerical values and 31 diagnostic labels), (ii) ECG voltage–time traces only, (iii) age and sex alone, (iv) ECG measures with age and sex, and (v) ECG voltage–time traces with age and sex. Models for (i), (iii) and (iv) used XGB, and models for (ii) and (v) used a DNN. 'Normal' refers to the ECGs in the test set labeled as normal by the original interpreting physician at the time of ECG acquisition, 'abnormal' refers to any ECGs not identified as normal in the test set and 'all' includes both normal and abnormal ECGs in the test set. **b**, The relative performance of the DNN models using single leads as input (sorted by increasing performance). The mean AUC of models M1–M5 (derived from fivefold cross-validation) are shown as the bar heights, while individual data points for each of the five models are shown as a red 'x'; black dots represent the AUC of model M0 (trained on 60% of the data and tested on the 40% holdout set). '2.5 s' and '10 s' refer to the duration of the voltage–time traces used for the model. Performance as a measure of AUPRC is shown in Extended Data Fig. 3.

true. Second, we demonstrated that a DNN has higher accuracy for the prediction of 1-year mortality than a model that uses the routine measurements and patterns identified by modern automated ECG systems and confirmed by cardiologists. Third, we showed that the predictive accuracy of a DNN is preserved even in the large subset of ECGs interpreted as normal by physicians. Finally, we showed that the model retains predictive ability for decades, well beyond 1 year.

We extracted all 12-lead ECGs from the electronic records of a large regional US health system (Geisinger). A standard 12-lead ECG has 15 voltage–time traces, including traces of 2.5 s in duration for all 12 leads and traces of 10 s in duration for leads V1, II and V5. After certain exclusions (illustrated in Extended Data Fig. 1), there were 2,338,833 ECGs from 536,661 patients available for the study. We split the dataset into a cross-validation (CV) dataset (60% of the data) and holdout test dataset (40% of the data). The CV dataset included 1,169,662 ECGs from 253,397 patients used for training. The holdout test dataset consisted of 168,914 ECGs with 14,207 events, where a single random ECG was chosen for each patient. We trained a DNN to aggregate the spatial and temporal features of the voltage–time signals to predict 1-year mortality. The DNN architecture is illustrated in Extended Data Fig. 2. A single model (M0) was trained on the CV dataset and evaluated on the holdout set. Additionally, fivefold cross-validation was performed within the CV dataset to evaluate performance (models M1–M5). Patients were not shared between training and test sets (details in the Methods).

The area under the receiver operating characteristic curve (AUC) for predicting 1-year all-cause mortality was 0.855 (model M0; CV (M1–M5): 0.859 ± 0.001) when using the ECG voltage–time traces alone. Performance improved to 0.876 (CV: 0.870 ± 0.006) when age and sex were added as additional input features (Fig. 1a, blue bars). The model trained with all 15 ECG voltage–time traces together provided the best AUC in comparison to models derived from any single lead (Fig. 1b). Models derived from the 10-s tracings had higher AUCs than models derived from the 2.5-s tracings, demonstrating that longer duration of the data likely provides more informative features to the model (Fig. 1b).

We compared the model performance of the DNN trained using voltage–time traces to that of baseline models trained using only age and sex as features and traditional interval measurements and patterns/diagnoses (ECG measures) as features. ECG measures were derived from the clinical reports, including 9 continuous numerical measurements (for example, QRS duration) and 31 categorical patterns (for example, left bundle branch block) (complete list in the Methods). The baseline models with tabular features were trained with an XGBoost (XGB) classifier[32,33] to predict 1-year mortality. The AUC of the XGB model with age and sex was 0.774 (CV: 0.772 ± 0.002) (Fig. 1a, gray bars). The AUC of the XGB model with ECG measures was 0.817 (CV: 0.820 ± 0.002), which improved to 0.860 (CV: 0.859 ± 0.002) with the addition of age and sex (Fig. 1a, orange bars). Both of these results were significantly lower than the AUCs of the corresponding DNN models ($P < 0.05$ from confidence intervals of average bootstrapped difference in AUC), although the good performance of the ECG measures model is notable (Fig. 1a, orange bars) and could have potential value for scenarios lacking digitized traces. To compare with common clinical risk scoring methods, we implemented and evaluated the predictive performance of the Framingham risk score (FRS)[34] and the Charlson comorbidity index (CCI)[35]. The AUCs for these models were 0.648 and 0.816, respectively, again demonstrating inferiority to the DNN model ($P < 0.05$ from confidence intervals of average bootstrap difference in AUC when compared to the DNN models).

The model was also able to predict 1-year mortality from ECGs without easily recognizable high-risk features. To show this, we evaluated the performance of the model in ECGs interpreted as normal by a physician. This set included only ECGs lacking any diagnostic abnormalities. Note that an ECG interpreted as normal does not necessarily imply that the ECG was collected from a patient without cardiovascular disease. There were 42,285 normal and 124,378 abnormal ECGs in the holdout test set. For normal ECGs, the DNN model (trained to predict 1-year mortality from all the ECGs) yielded AUCs of 0.804 (CV: 0.801 ± 0.009) and 0.849 (CV: 0.833 ± 0.001) for ECG traces alone and with the addition of age and sex, respectively; for abnormal ECGs, the model yielded mean AUCs of 0.841 (CV: 0.846 ± 0.001) and 0.862 (CV: 0.855 ± 0.006),

Table 1 | Summary of the performance of the DNN model (M0) trained with voltage–time traces, as well as age and sex, in different scenarios

Data	ECGs (%)	Model performance					
		All		Normal		Abnormal	
		AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
Holdout set (168,914 ECGs)							
All	100	0.876	0.425	0.849	0.196	0.862	0.442
Clinical context (168,914 ECGs available)							
Inpatient	19	0.791	0.546	0.791	0.351	0.779	0.555
Emergency	8	0.862	0.337	0.825	0.191	0.852	0.357
Outpatient	38	0.843	0.237	0.779	0.075	0.836	0.256
Unknown	35	0.882	0.336	0.864	0.125	0.867	0.351
Comorbidities (167,816 ECGs available)							
Coronary artery disease	8	0.811	0.488	0.769	0.185	0.804	0.495
Hypertension	39	0.850	0.415	0.821	0.176	0.837	0.430
Heart failure	2	0.734	0.605	0.779	0.555	0.729	0.606
Diabetes	13	0.825	0.426	0.812	0.220	0.807	0.437
Without the above phenotypes	28	0.891	0.417	0.861	0.208	0.880	0.436

The holdout set corresponds to all data available in the 40% of unique patients held out before the beginning of the study; clinical context corresponds to the patient setting at the time of ECG acquisition; and comorbidities corresponds to patients with known clinical comorbidities at the time of ECG acquisition. Data shown for both clinical context and comorbidities refer to the patients in the holdout set. Patient demographics are shown in Table2. DNN, deep neural network; AUC, area under the receiver operating characteristic curve; AUPRC, area under the precision recall curve; ECG, electrocardiogram.

respectively (Fig. 1a). The DNN using ECG traces generally had better performance than the XGB model using ECG measures for both normal and abnormal ECGs, although the confidence intervals of the mean AUCs did overlap for the case of normal ECGs with age and sex. Performance as a measure of area under the precision recall curve (AUPRC) is shown in Extended Data Fig. 3.

To determine whether the model performance was robust with respect to patient comorbidities, we completed subanalyses across a range of cardiovascular diagnoses and clinical contexts within the holdout test set. Generally, model performance was consistently maintained across these different scenarios (Table 1; additional details available in Table 2). We also approximated a test in an independent clinical setting by dividing all data by hospital or clinic of acquisition and separately training and testing across sites. The model trained with data from our largest hospital (567,788 ECGs from 161,152 patients) and tested on data from the rest of the Geisinger system (292,671 patients with a single random ECG chosen for each patient) yielded AUCs of 0.852 (without age and sex included as variables) and 0.870 (with age and sex included as variables).

Although the DNN model demonstrated good performance, model interpretability is a challenge. We used two techniques to highlight features used by the model that may correlate with clinically known ECG patterns that are useful for the prediction of mortality. First, we summarized patient demographics and the distribution of ECG measures by class predictions in the holdout test set (based on the M0 model) in Table 2. The table demonstrates that ECG findings that are correlated with a higher risk of mortality (for example, atrial fibrillation, left bundle branch block and prior infarct) are much more common in those predicted to die within 1 year. Second, we implemented a guided gradient class activation mapping (Grad-CAM)[36,37] method to display the neural-network-based model activations that contributed to a prediction of mortality within 1 year. As an example, we hypothesized that ECGs showing acute anterior ST segment elevation myocardial infarction (STEMI) would show higher saliency at the sites of ST segment elevation. Extended Data Fig. 4 illustrates that indeed elevated ST segments of leads V2 and V3 were highlighted as salient, in conjunction with a high likelihood risk score prediction for 1-year mortality, for three patients with anterior STEMI who died within 1 year of the ECG. These findings demonstrate initial promising results for saliency maps based on the guided Grad-CAM approach (details in the Methods).

To further investigate predictive performance within the overall dataset and the subsets of normal and abnormal ECGs, a Kaplan–Meier survival analysis was performed using follow-up data available in the electronic health record (EHR) for the chosen operating point on the ROC curve (Fig. 2 and Table 2). The performance characteristics of the model at different operating points are summarized in Supplementary Table 2. For normal ECGs, the median survival times of the groups predicted alive and dead at 1 year (model M0 on the holdout set) were >25 and 7 years, respectively, and, for abnormal ECGs, these were 21 and 4 years, respectively (Fig. 2b). A Cox proportional-hazard regression model was fit, and the hazard ratios with 95% confidence intervals were 8.1 (7.9–8.2) (CV: 8.1 (7.4–9.0)) in all ECGs, 6.8 (6.6–6.9) (CV: 6.8 (6.2–7.6)) in abnormal ECGs and 9.5 (9.0–10.0) (CV: 9.1 (7.3–11.0)) in normal ECGs (all $P < 0.005$) when comparing those predicted by the DNN to be alive versus dead at 1 year after ECG. Thus, the hazard ratio was largest in the subset of normal ECGs, and the prediction of 1-year mortality from the DNN was a significant discriminator of long-term survival for up to 25 years after the initial ECG.

To further explore the predictive characteristics within the subset of normal ECGs, we hypothesized that, among patients with normal ECGs who died within 1 year, the burden of cardiac-related mortality would be higher in cases correctly predicted to die by the model than in those who died within 1 year despite being predicted to survive. We completed a blinded physician chart review to classify cause of death in 266 age- and sex-matched patients from these two groups within the holdout test set to define the primary cause of death as cardiac, non-cardiac or unknown. We did not find evidence that the proportion of cardiac-related deaths was different between these groups (3.8% versus 4.5% for predicted dead and alive, respectively), suggesting that the predictive features may extend beyond cardiac disease alone.

Finally, we investigated whether the features learned by the model are visually apparent to cardiologists and whether they are

Table 2 | Patient demographics and summary of data distribution across the predicted groups for the DNN model (M0) trained with ECGs, as well as age and sex

	Holdout test dataset (total)	Holdout test dataset	
		Predicted dead	Predicted alive
ECGs (*n*)	168,914	38,702	130,212
Events (*n*)	14,207	11,004	3,203
Age (years)	58 ± 18	73 ± 14	53 ± 17
Sex (male in %)	47	50	46
Patterns (31 categorical variables) as % of ECGs			
Ventricular tachycardia	0.04	0.2	0
SVT	0.3	0.9	0.1
Atrial flutter	0.6	1.9	0.2
Atrial fibrillation	4.4	14.5	1.5
Complete block	0.04	0.1	0.02
Pacemaker	1.7	5.4	0.6
Left BBB	1.7	4.8	0.7
Incomplete left BBB	0.2	0.6	0.1
Second-degree AV block	0.08	0.2	0.05
Intraventricular block	1.3	3.2	0.7
Fascicular block	2.2	5.4	1.3
PVC	4.5	10.5	2.7
Sinus tachycardia	7.9	17.9	5.0
Ischemia	6.0	13.4	3.8
Right BBB	3.9	8.6	2.5
PAC	3.7	8.3	2.4
Left axis deviation	6.3	13.5	4.2
Prolonged QT	3.4	6.7	2.3
Low QRS voltage	3.9	7.8	2.8
Prior infarct	13.6	24.8	10.2
First-degree AV block	4.2	7.2	3.4
Acute MI	0.6	1.0	0.5
Nonspecific ST abnormality	7.5	12.3	6.0
Right axis deviation	2.0	3.0	1.7
LVH	6.8	9.7	5.9
Nonspecific T-wave abnormality	10.5	14.9	9.2
Other bradycardia	0.08	0.1	0.08
Incomplete right BBB	3.4	2.9	3.6
Sinus bradycardia	14.6	5.8	17.3
Normal	26.8	6.3	32.9
Early repolarization	0.6	0.1	0.7
ECG measurements (9 continuous variables)			
QRS duration (ms)	93 ± 20	101 ± 28	91 ± 15
QT (ms)	393 ± 44	388 ± 60	395 ± 38
QTC (ms)	436 ± 34	458 ± 42	429 ± 28
PR interval (ms)	155 ± 40	151 ± 60	156 ± 31
Ventricular rate (bpm)	77 ± 19	88 ± 23	73 ± 16
Average RR interval (ms)	824 ± 183	728 ± 193	852 ± 170
P axis	47 ± 25	48 ± 31	47 ± 23
R axis	27 ± 43	16 ± 58	30 ± 37
T axis	45 ± 42	63 ± 63	40 ± 31

Predicted groups were identified using an optimal threshold on the receiver operating characteristic curve from the training data with the highest point on the iso-accuracy line. Continuous measurements are presented as the mean ± s.d. *n*, number of samples; ms, milliseconds; bpm, beats per minute; PVC, premature ventricular contractions; PAC, premature atrial contractions; SVT, supraventricular tachycardia; MI, myocardial infarction; BBB, bundle branch block; LVH, left ventricular hypertrophy; AV, atrioventricular.

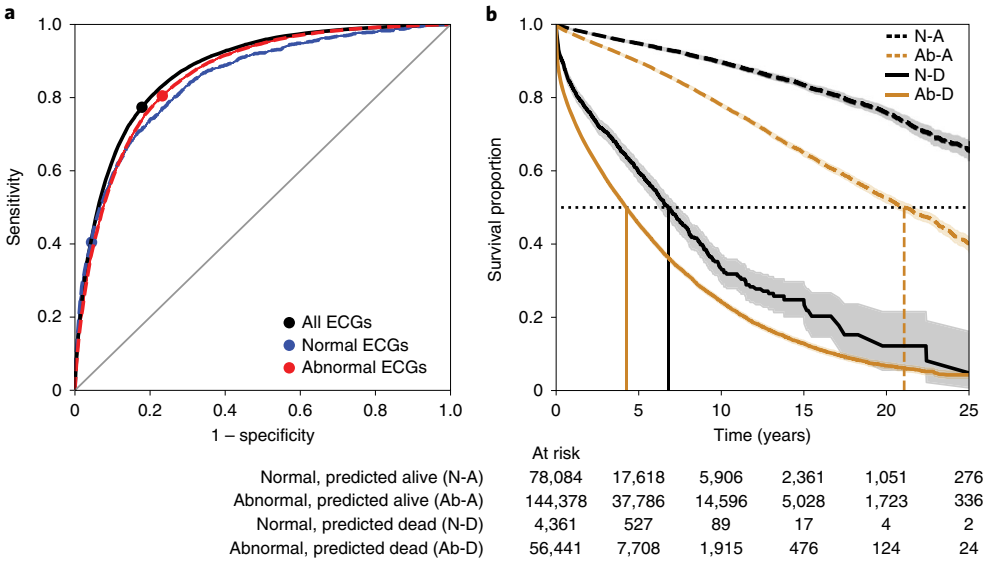


Fig. 2 | Receiver operating characteristic curves indicating the optimal operating point and corresponding Kaplan–Meier survival curves. a, ROC curves of the trained M0 models with operating points marked for all the data (black circle), the normal ECG subset (blue circle) and the abnormal ECG subset (red circle). **b**, Kaplan–Meier curves for predicted alive and dead groups in the normal and abnormal ECG subsets at the operating points in **a**. The shaded area is the 95% confidence interval. The table shows the at-risk population for the given time intervals in the holdout test data.

clinically interpretable. To do this, we randomly chose 100 sets of paired normal ECGs from the holdout test set. Each pair consisted of a true positive (normal ECG from a patient correctly predicted by the model to die within 1 year) and a true negative (normal ECG from a patient correctly predicted by the model to be alive at 1 year), matched for age and sex. We surveyed ten cardiologists, asking them to identify which ECG of each pair was linked to 1-year mortality. The cardiologists had accuracies of 61–78% (22–56% above random chance) for this initial task. After allowing each cardiologist to study a separate dataset of 100 paired ECGs labeled to show the outcome, their prediction accuracy upon repeating the original blinded survey of 100 paired ECGs was 60–93% (20–86% above random chance), demonstrating an average relative improvement in performance of 13% after seeing the model results ($P=0.017$). The cardiologists reported that three features were useful for the prediction of mortality: higher heart rates, the quality of the ECG baseline and evidence of slight left atrial enlargement (Extended Data Fig. 5).

We chose to predict 1-year all-cause mortality as a target endpoint because it is well defined, readily available and clinically meaningful. We believe that mortality prediction is useful in numerous clinical contexts. In the outpatient setting, primary care physicians and cardiologists use risk prediction tools to assess the safety of elective surgical procedures and to guide therapy for both primary and secondary prevention of cardiovascular disease. A 'normal' ECG may be giving false reassurance in these settings. In the emergency department and inpatient setting, the model may be able to help stratify patients by risk who present with chest pain or nonspecific symptoms that might be an anginal equivalent. For elderly patients and those with multiple comorbidities approaching the end of life, a prediction of 1-year mortality may help guide decisions surrounding palliative care and the use of invasive and/or high-risk procedures. Finally, when applied at the population level, our model could enable health systems and insurance providers to better understand and optimally deploy resources to their patient population. The Kaplan–Meier curves show that the model clearly identifies a high-risk population, even in cases of false positives (that is, despite being a 'false positive' for 1-year mortality, these patients maintain significantly elevated risk of death over 25 years). This high-risk population may benefit from targeted interventions with the goal of improving outcomes. In all of these hypothetical settings, we expect the model to complement, not replace, physician judgment.

The findings from our physician survey have two insights to offer in this regard: (1) while the physicians were able to correctly identify many of the normal ECGs linked to mortality in this idealized setup, not all cases were visually apparent and thus the model supplemented physician insight and (2) physician performance improved slightly after reviewing the model predictions, suggesting the potential for complementary gains through human–machine interaction. In the clinical scenarios listed above, the model could be used as a standalone tool, integrated with existing risk scores or further developed into a new comprehensive risk scoring system (for example, by adding additional input features). Future work is needed to ensure model generalizability to these different clinical contexts, but evidence presented in this study suggests that it will prove useful in varied clinical settings.

There are several limitations to acknowledge. We applied minimal restrictions on the included ECG data, such as removing poor tracings and studies without the standard three 10-s rhythm strips. These exclusions may have introduced bias; however, this risk is minimized by the large size of the dataset and the demonstrated performance across various clinical contexts (Table 1). Additionally, the neural network design currently precludes full interpretability, so the ability to define the basis for model predictions is limited. Available techniques, including Grad-CAM, provided evidence that the model was identifying clinically relevant features (Extended Data Fig. 4), but full interpretability will be a focus of future work.

In summary, we leveraged a large set of ~2.3 million ECGs collected from 536,661 patients over a period of 34 years to demonstrate the potential for DNNs to automatically predict a highly clinically relevant endpoint (1-year mortality) directly from 12-lead ECG voltage–time data. This potential is evident through several important findings. First, the DNN model using voltage–time traces alone outperformed traditional clinical risk scores such as FRS and CCI, as well as another machine learning models that used a collection of 40 clinical ECG features (including both numerical measurements and diagnostic patterns). This suggests that the DNN is

able to identify novel patterns of prognostic relevance from voltage–time traces. Second, in addition to predicting 1-year mortality with an AUC of 0.876, the model showed substantial predictive ability beyond the 1-year mark. Third, despite conventional wisdom regarding the negative predictive value of ECGs[38,39], we found that prediction accuracy remained high even in the large subset of ECGs interpreted as normal by a cardiologist. We have shown that deep learning has the potential to add significant prognostic information to one of the most widely used medical tests, the 12-lead ECG, which with further study could prove useful in a clinical context, both for risk prediction and to improve outcomes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-0870-z.

Received: 29 March 2019; Accepted: 1 April 2020;
Published online: 11 May 2020

References

1. Fye, W. B. A history of the origin, evolution, and impact of electrocardiography. *Am. J. Cardiol.* **73**, 937–949 (1994).
2. Chesebro, J. H. et al. Thrombolysis in Myocardial Infarction (TIMI) trial, phase I: a comparison between intravenous tissue plasminogen activator and intravenous streptokinase. Clinical findings through hospital discharge. *Circulation* **76**, 142–154 (1987).
3. Eagle, K. A. et al. A validated prediction model for all forms of acute coronary syndrome. *JAMA* **291**, 2727–2733 (2004).
4. Kenchaiah, S. et al. Obesity and the risk of heart failure. *N. Engl. J. Med.* **347**, 305–313 (2002).
5. Levy, W. C. et al. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation* **113**, 1424–1433 (2006).
6. Goldman, L. et al. Multifactorial index of cardiac risk in noncardiac surgical procedures. *Surv. Anesthesiol.* **22**, 482 (1978).
7. Lloyd-Jones, D. M. et al. Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease. *Circulation* **139**, e1162–e1177 (2019).
8. Hwang, W., Chang, J., LaClair, M. & Paz, H. Effects of integrated delivery system on cost and quality. *Am. J. Manag. Care* **19**, e175–e184 (2013).
9. Motwani, M. et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur. Heart J.* **52**, 468–476 (2016).
10. Curry, S. J. et al. Screening for cardiovascular disease risk with electrocardiography: US Preventive Services Task Force recommendation statement. *JAMA* **319**, 2308–2314 (2018).
11. Lanza, G. A. The electrocardiogram as a prognostic tool for predicting major cardiac events. *Prog. Cardiovasc. Dis.* **50**, 87–111 (2007).
12. Pan, J. & Tompkins, W. J. A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* **32**, 230–236 (1985).
13. Belforte, G., De Mori, R. & Ferraris, F. A contribution to the automatic processing of electrocardiograms using syntactic methods. *IEEE Trans. Biomed. Eng.* **26**, 125–136 (1979).
14. Madeiro, J. P. V., Cortez, P. C., Marques, J. A. L., Seisdedos, C. R. V. & Sobrinho, C. R. M. R. An innovative approach of QRS segmentation based on first-derivative, Hilbert and wavelet transforms. *Med. Eng. Phys.* **34**, 1236–1246 (2012).
15. Köhler, B. U., Hennig, C. & Orglmeister, R. The principles of software QRS detection. *IEEE Eng. Med. Biol. Mag.* **21**, 42–57 (2002).
16. Addison, P. S. Wavelet transforms and the ECG: a review. *Physiol. Meas.* **26**, R155–R199 (2005).
17. Acharya, U. R. et al. A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* **89**, 389–396 (2017).
18. LeBlanc, A. R. Quantitative analysis of cardiac arrhythmias. *Crit. Rev. Biomed. Eng.* **14**, 1–43 (1986).
19. Luz, E. J., Schwartz, W. R., Cámara-Chávez, G. & Menotti, D. ECG-based heartbeat classification for arrhythmia detection: a survey. *Comput. Methods Programs Biomed.* **127**, 144–164 (2016).
20. Rahhal, M. M. A. et al. Deep learning approach for active classification of electrocardiogram signals. *Inf. Sci.* **345**, 340–354 (2016).
21. Liu, W. et al. Real-time multilead convolutional neural network for myocardial infarction detection. *IEEE J. Biomed. Health Inform.* **22**, 1434–1444 (2017).
22. Goodfellow, S. D. et al. Towards understanding ECG rhythm classification using convolutional neural networks and attention mappings. in *Machine Learning for Healthcare Conf.* 83–101 (2018).
23. Yu, S. N. & Chen, Y. H. Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recog. Lett.* **48**, 1142–1150 (2007).
24. Asl, B. M., Setarehdan, S. K. & Mohebbi, M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artif. Intell. Med.* **44**, 51–64 (2008).
25. Karpagachelvi, S., Arthanari, M. & Sivakumar, M. Classification of ECG signals using extreme learning machine. *Comput. Inf.* Sci. https://doi.org/10.5539/cis.v4n1p42 (2014).
26. Kampouraki, A., Manis, G. & Nikou, C. Heartbeat time series classification with support vector machines. *IEEE Trans. Inf. Technol. Biomed.* **13**, 512–518 (2009).
27. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
28. Smith, S. W. et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. *J. Electrocardiol.* **52**, 88–95 (2019).
29. Attia, Z. I. et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med.* **25**, 70–74 (2019).
30. Attia, Z. I. et al. Prospective validation of a deep learning ECG algorithm for the detection of left ventricular systolic dysfunction. *J. Cardiovasc. Electrophysiol.* **30**, 668–674 (2019).
31. Attia, Z. I. et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* **6736**, 1–7 (2019).
32. Chen, T. & Guestrin, C. XGBoost: reliable large-scale tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
33. Chen, T. & He, T. Higgs Boson discovery with boosted trees. *JMLR Work. Conf. Proc.* **42**, 69–80 (2015).
34. D'Agostino, R. B. et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
35. Quan, H. et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **43**, 1130–1139 (2005).
36. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
37. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: the all convolutional net. Preprint at https://arxiv.org/abs/1412.6806 (2014).
38. Davie, A. P. et al. Value of the electrocardiogram in identifying heart failure due to left ventricular systolic dysfunction. *Br. Med. J.* **312**, 222 (1996).
39. Hedberg, P. et al. Electrocardiogram and B-type natriuretic peptide as screening tools for left ventricular systolic dysfunction in a population-based sample of 75-year-old men and women. *Am. Heart J.* **148**, 524–529 (2004).

Methods

Detailed information on the experimental design can also be found in the Nature Research Reporting Summary.

ECG and patient data. The institutional review board approved this study with a waiver of consent, in conjunction with our institutional patient privacy policies. We extracted 2.7 million standard 12-lead ECG traces from our institutional clinical MUSE (GE Healthcare) database, acquired between 1984 and 2019. We retained only the resting 12-lead ECGs with voltage–time traces of 2.5 s for 12 leads and 10 s for 3 leads (V1, II, V5) that did not have significant artifacts and were associated with at least 1 year of follow-up or death within 1 year. 'Significant artifacts' were defined as being reported as a poor tracing by the automated machine analysis or the interpreting cardiologist, with findings statements such as 'Poor data quality, interpretation may be adversely affected'. After exclusions, 2.3 million ECGs remained, with 60% of them stored at 500 Hz and the remaining ECGs stored at 250 Hz. All data were resampled to 500 Hz by linear interpolation. Sixty percent of the data were selected as a 'CV' (for cross-validation) set, with the remaining 40% used as a holdout test set (Extended Data Fig. 1).

The quantitative measurements and findings within the final ECG clinical reports were parsed to identify 31 diagnostic pattern classes and 9 continuous ECG measurements (all detailed below). An ECG was defined to be 'abnormal' if the pattern label was flagged for at least one diagnostic abnormality. The nine ECG measurements were the measurements logged in the confirmed ECG reports that included QRS duration, QT, QTc, PR interval, ventricular rate, average RR interval, and P-, Q- and T-wave axes. Data for the nine variables were 88–100% complete. Missing values were imputed using multiple imputations by chained equations[40]. Patterns included normal, left bundle branch block, incomplete left bundle branch block, right bundle branch block, incomplete right bundle branch block, complete heart block, atrial fibrillation, atrial flutter, acute myocardial infarction, left ventricular hypertrophy, premature ventricular contractions, premature atrial contractions, first-degree atrioventricular block, second-degree atrioventricular block, fascicular block, sinus bradycardia, other bradycardia, sinus tachycardia, ventricular tachycardia, supraventricular tachycardia, prolonged QT, pacemaker, ischemia, low QRS voltage, intraventricular block, prior infarct, nonspecific T-wave abnormality, nonspecific ST abnormality, left axis deviation, right axis deviation and early repolarization. The 31 clinical diagnosis patterns were parsed from the structured findings statements on the basis of the key phrases that are standard within the MUSE system. This was performed by identifying the key phrases by thorough examination of the lookup table of MUSE codes and the corresponding 'findings' text strings that correspond to each pattern and, additionally, by identifying any variations in the findings text when the physician manually entered 'free' text into the findings field: for example, 'Atrial flutter is present' or 'Atrial flutter now present' for a positive flag and 'Atrial flutter' string match along with 'Atrial flutter has resolved' for a negative flag. Such a string-based rule set was defined for each pattern, and labels for the 31 patterns were generated for all ECGs. The labels were iteratively reviewed for hundreds of cases to ensure accuracy.

For patient demographic data, the survival time and patient age were calculated with reference to the date of ECG acquisition and only patients above 18 years of age at the time of ECG were included in this study. Patient status (dead/alive) was defined through a combination of EHRs and monthly updates from the Social Security Death Index. Moreover, data for alive patients were censored at the patient's last known physical alive encounter to limit bias from incomplete records. Sex was also extracted from the EHR data.

ECG sampling strategy. To avoid over-representing sicker patients who undergo more ECGs, we selected a single ECG per patient in the test set. Rather than selecting the most recent (last) ECG (that is, the ECG closest to death for patients who died), we randomly sampled one ECG from among all the ECGs available for a given patient. This latter strategy was considered to be most representative of deploying the model on a given ECG from a new patient, which in each case will be at a random time point in that patient's life. To test the consistency of the random sampling strategy, we repeated 50 different random selections and demonstrated consistency in the model performance (data not shown).

Model development and evaluation, including statistical analysis. We designed a convolutional neural network (model architecture illustrated in Extended Data Fig. 2) using five branches with the input of three leads as channels that are concurrent in time (branch 1: leads I, II, III; branch 2: leads aVR, aVL, aVF; branch 3: leads V1, V2, V3; branch 4: leads V4, V5, V6; branch 5: leads V1-long, II-long, V5-long). Note that each branch represents the three leads that were acquired at the same time (during the same heartbeats), for a duration of 2.5 s. In a typical 12-lead ECG, four of these groups of three leads are acquired over a duration of 10 s. Concurrently, the 'long leads' are recorded over the entire 10-s duration. Thus, the architecture was designed to account for these details, because arrhythmias in particular cause the traces to change morphology throughout the standard clinical acquisition.

A convolutional block consisted of a one-dimensional convolution layer followed by batch normalization and rectified linear unit (ReLU)[41] activations. The first four branches and the last branch consisted of four and six convolutional blocks, respectively, followed by a global average pooling (GAP)[42] layer. The outputs of all of the branches were then concatenated and connected to a series of six dense layers of 256 (with dropout), 128 (with dropout), 64, 32, 8 and 1 unit(s) with a sigmoid function as the final layer. We used the Adam[43] optimizer with a learning rate of 1×10^{-5}, and batch size was set to 2,048; otherwise, all hyperparameters were set to default. Each model branch was computed in parallel on a separate GPU for faster computation. Replacing the GAP layer with long short-term memory[44] gave similar performance for substantially longer run times; hence, the final model used GAP layers.

We developed a model (M0) trained on the entire CV set and five models (M1–M5) from fivefold cross-validation within the CV set. In a fivefold cross-validation, the data are split into five folds: that is, 80% of the data are used for training, and the remaining 20% are the test set for evaluation. The five models are tested on the respective unique 20% test sets. Five percent of the training set was identified as an internal validation set before training for tracking the validation loss during training to avoid overfitting by early stopping[45]. The models were evaluated by AUC, which is a robust metric of model performance that represents the ability to discriminate between two classes. Higher AUC suggests higher performance, with an AUC of 0.5 being equivalent to a random chance guess and an AUC of 1.0, representing perfect discrimination. The AUC for model M0 was reported on the holdout test set. Additionally, we also reported the mean and s.d. of the AUC for CV models M1–M5 to evaluate model generalizability and stability. The AUCs were compared by bootstrapping 1,000 instances (using random and variable sampling with replacement). Differences between models were thus defined to be statistically significant if the absolute difference in the 95% confidence intervals was greater than zero. We also evaluated the models using AUPRC, computing average precision score as a weighted average of the precisions achieved at each threshold by the increase in recall. The operating point was optimized for maximum accuracy in the training set and applied on the respective test set. The predictions from model M0 were used to choose candidates for the clinical chart review for cause of death and for the data chosen for the visual survey by cardiologists. To compare the prognostic efficacy of the ECG voltage–time traces to the corresponding clinically reported ECG measures, we cross-validated an XGB classifier with exactly the same training, internal validation and test sets used for the DNNs.

The models were trained with all of the available ECGs for patients in the training set with their corresponding survival time frames. The test set included one randomly chosen ECG for each patient. The data were split such that the same patient was not in both the training and test sets. We evaluated loss (binary cross-entropy) on the internal validation set (which was approximately balanced for outcome class) for each epoch. The loss function was weighted to compensate for the imbalance in the proportion of output labels (alive/dead) during training. The training was terminated if the internal validation loss did not decrease for 10 epochs (early-stopping criteria), and the maximum number of epochs was set to 500.

The model was implemented using Keras (version 2.1.6-tf) with a TensorFlow backend (version 1.9.0) in Python (version 3.5.2), and default training parameters were used except where specified. For single leads as input, a single branch of the above-mentioned model was used. When demographic variables (age and sex) were added to the model, a 64-hidden-unit layer following the input layer was concatenated with the other branches. All training was performed on an NVIDIA DGX2 platform with 16 available V100 GPUs and 32 GB of RAM per GPU. When fit with five GPUs, each model took ~10 min per epoch.

Phenotypes. We defined four comorbidities or phenotypes, including coronary artery disease, hypertension, heart failure and type 2 diabetes, for patients in the holdout set at the time of ECG. The phenotype definitions were as follows:

1. Coronary artery disease was defined as any of the following: a discharge diagnosis of STEMI or non-STEMI; a discharge diagnosis of unstable angina with negative biomarkers (troponin or creatine kinase myocardial band) and with evidence of either coronary artery disease or inducible ischemia; a history of percutaneous coronary intervention or coronary artery bypass grafting procedures; or exertional angina with evidence of coronary artery disease or inducible ischemia.
2. Hypertension was defined as satisfying at least two of the following criteria: (i) two instances of antihypertensive medication orders; (ii) at least three high-blood-pressure readings (≥140 systolic or ≥90 diastolic) each over 7 d apart; or (iii) a diagnosis code (ICD-9 or ICD-10) for hypertension either on the patient problem list or on at least two clinic encounters within 2 years.
3. Heart failure was defined using the 'definite' category of the eMERGE phenotype[46]. This was implemented and validated by blinded chart review within Geisinger EHR data.
4. Type 2 diabetes was defined as satisfying at least two of the following conditions: (i) two abnormal lab values (outpatient fasting blood glucose result of ≥125 mg/dl, outpatient blood glucose result of >200 mg/dl or outpatient HbA1c result of ≥6.5%) at least 90 d apart; (ii) two prescriptions of antidiabetic medication at least 90 d apart; or (iii) a diagnosis of type 2 diabetes from either the patient problem list or a completed office visit.

Additional validation, simulating external dataset. The ECGs were classified by the location where the ECG was taken as 'GMC' or 'non-GMC'. GMC represented patients who had ECGs in Geisinger Medical Center (GMC) in Danville, PA, and non-GMC represented ECGs acquired at other facilities within the Geisinger system, comprising a mix of hospital and community clinic settings. All patients in the non-GMC group who were also in the GMC group were removed from the non-GMC group, such that there was no overlap of patients between the two groups. A model was trained with all ECGs relative to their events from the GMC group (567,788 ECGs from 161,152 patients) and tested on the non-GMC group (292,671 ECGs from 292,671 patients). A single random ECG for a patient was chosen in the test set for evaluation. The test set had 17,279 events in 292,671 ECGs, with 1,338 events among 86,878 normal ECGs and 15,941 events among 205,793 abnormal ECGs.

Model interpretations. Interpretable models are important to bolster transparency in predictions and to increase clinical acceptance. To create interpretable outputs for model predictions, we used a technique called guided Grad-CAM to objectively determine the salient regions of a particular lead contributing toward a mortality prediction. We achieved this by first calculating the gradients for each prediction score with respect to the feature maps of the final convolution layer for each branch[36]. Next, these gradients were global average pooled to calculate feature importance weights, which were then used to form a partial linearization of weights from each branch to give the Grad-CAM. We also calculated rectified gradients across each individual lead with Guided-Backpropagation[37] and finally multiplied pointwise with the Grad-CAM output to give the ECG's guided Grad-CAM results. Via peak-to-peak detection[47], we calculated the average heartbeat along with the respective guided Grad-CAM values for each lead with a total time step of 600 ms, to create the 'average' ECG waveform for each lead and patient. Finally, these maps were averaged across patients to generate saliency maps.

Some limitations of using the guided Grad-CAM approach include generalization of this approach across large cohorts where the saliencies across patients and varying average heartbeat waveforms can nullify confounding activations. This may result in average class activation maps with lost pertinent information, providing limited interpretations. The current methods for interpretability based on class activations are tuned for individual patients, while future work can include techniques that generalize model interpretability well among large patient cohorts and across multiple disease states.

Survival analysis. We performed Kaplan–Meier survival analysis[48] with the available follow-up data stratified by the DNN model prediction, using an optimal operating point for the models. The data were censored on the basis of the most recent encounter. We fit a Cox proportional-hazard model[49] regressing mortality on the DNN-model-predicted classification of alive or dead in the subset of normal ECGs and in the subset of abnormal ECGs. The hazard ratios with 95% confidence intervals were reported for all data, the normal subset and the abnormal subset for models M0 and M1–M5 (mean value with lower and upper bounds, of 95% confidence interval, in this case). The lifelines package (version 0.21.5) in Python was used for survival analysis.

Chart review, including statistical analysis. We selected all patients from the holdout test set with a normal ECG who were incorrectly predicted by the model to survive 1 year (that is, the patient died within 1 year despite the model predicting that he/she would survive) and matched them by age and sex with patients from the holdout set with a normal ECG who were correctly predicted to die in 1 year. A comprehensive single physician chart review was performed, with blinding to group status, using the available EHR data to define cause of death as cardiac, non-cardiac or unknown. Cause of death was designated as cardiac if there was sufficient evidence that any of the following disease processes led to mortality: heart failure, cardiomyopathy, coronary artery disease, arrhythmias, valvular heart disease or pericardial effusion. Cause of death was listed as unknown if there was inadequate documentation to determine an exact cause and for cases where medical complexity precluded a definite classification.

For deaths occurring in the hospital, notes from the three most recent outpatient visits, admission notes and MD progress notes were used to adjudicate cause of death. For deaths occurring outside the hospital, notes from the three most recent outpatient visits and social work notes were used, if available. If there was a hospital admission note available within 30 d of death with no useful data afterward, we based the cause of death on the available admission note rather than assigning an unknown status. If there were no available records within 60 d of the presumed date of death, the case was classified as unknown. Official death certificates were also used, if available.

We compared the frequency of cardiac-related mortality between groups by Fisher's exact test.

Cardiologist survey, including statistical analysis. In an effort to identify the differential clinical features between true positives and true negatives in ECGs interpreted as normal by a physician, we designed a series of surveys for ten independent cardiologists. Pairs of ECGs, including one true positive (correctly predicted by the model to die within 1 year) and one true negative (correctly predicted by the model to survive 1 year), matched for age and sex, were presented to each cardiologist, with blinding to the model outcome. The cardiologists were aware that each pair contained a true positive and a true negative. They were presented with 100 pairs of ECGs and asked to determine which patient died within 1 year. In addition to the ECG tracing, they were shown patient age, sex and nine computed ECG measurements. Next, they were shown 100 different ECGs with survival status labeled. Finally, they were shown the original survey again and asked to predict survival status. The cardiologists did not know their performance on the initial survey before completing the final survey. We compared the performance on the surveys before and after seeing the model results by two-tailed paired *t* test.

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

Data availability

All requests for raw and analyzed data and related materials, excluding programming code, will be reviewed by our legal department to verify whether the request is subject to any intellectual property or confidentiality constraints. 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 for non-commercial research purposes.

Code availability

Programming code related to data preprocessing and model specification will be made available under GNU General Public License version 3 upon request to the corresponding author.

References

40. van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, https://doi.org/10.18637/jss.v045.i03 (2011).
41. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. *Proceedings of the ICML* **27**, 807–814 (2010).
42. Lin, M., Chen, Q. & Yan, S. Network in network. Preprint at https://arxiv.org/abs/1312.4400 (2013).
43. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
44. Hochreiter, S. & Schmidhuber, J. J. Long short-term memory. *Neural Comput.* **9**, 1–32 (1997).
45. Prechelt, L. Early stopping—but when? in *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science* **7700**, 55–69 (Springer, 2012).
46. Shah, S. J. et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* **131**, 269–279 (2015).
47. Carreiras, C. et al. BioSPPy: biosignal processing in Python https://github.com/PIA-Group/BioSPPy (2018).
48. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Am. Stat. Assoc.* **53**, 457–481 (1958).
49. Cox, D. R. Regression models with life tables. *J. R. Stat. Soc. B* **34**, 187–220 (1972).

Acknowledgements

The authors would like to acknowledge C. Nevius and B. McCarty for their help in submitting the IRB approval for the study and developing a scheduler for efficient computational scheduling of the study experiments. The authors also acknowledge the time and contribution of the following cardiologists who performed the survey reported in the paper: N. Mead, B. Carry, G. Yost, S. Siddiqi, T. Rizwan and B. Durr.

Author contributions

S.R., C.M.H. and B.K.F. conceived the study and designed the experiments. S.R. conducted all the experiments. S.R., A.E.U.C. and J.S. contributed to the code base and deep learning framework used for the experiments. S.R., A.U.E.C., L.J., D.P.v.M., J.B.L. and D.N.H. assembled the data. H.L.K., B.P.D., A.A.P. and J.S. contributed to many discussions on experimental design. S.R. and D.P.v.M. designed the web application to perform the blinded surveys of cardiologists. C.W.G., J.M.P., A.A. and D.B. are the cardiologists who completed the survey and provided clinical insights. S.R., J.M.P., A.N., M.C.S., T.C., A.H. and K.W.J. contributed to the model interpretation with guided Grad-CAM. S.R., C.M.H. and K.Y. contributed to clinical chart review for cause-of-death analysis. All authors critically revised the manuscript.

Competing interests

This work was supported in part by funding from the Pennsylvania Department of Health (SAP 4100070267), an American Heart Association Competitive Catalyst Award (17CCRG33700289), the Geisinger Health Plan and Clinic, and Tempus. The content of this article does not reflect the views of the funding sources. Geisinger receives funding from Tempus for ongoing development of predictive modeling technology and

commercialization. Tempus and Geisinger have jointly applied for a patent related to the work. None of the Geisinger authors has ownership interest in any of the intellectual property resulting from the partnership. A.N., T.C., A.H., M.C.S. and K.W.J. are employees of Tempus.

Additional information

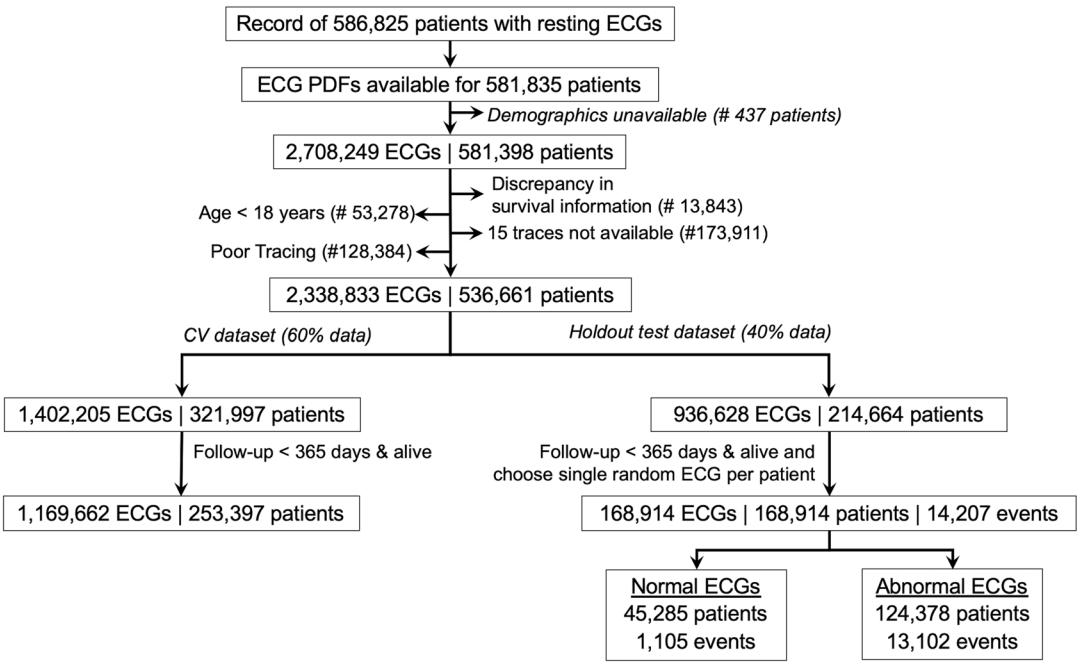
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-0870-z.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0870-z.

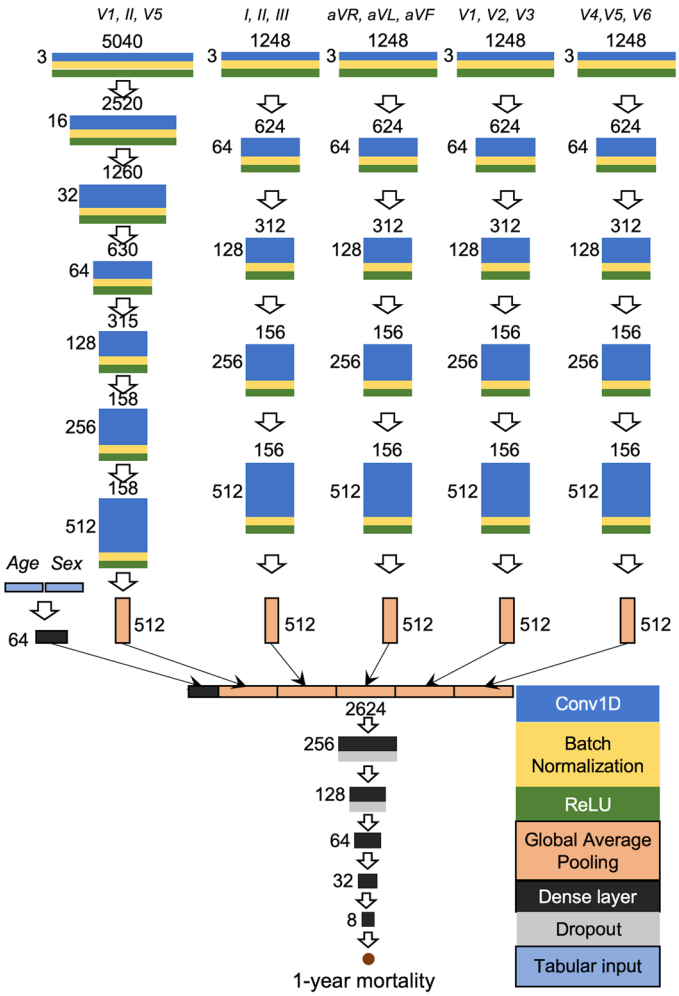
Correspondence and requests for materials should be addressed to B.K.F.

Peer review information Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

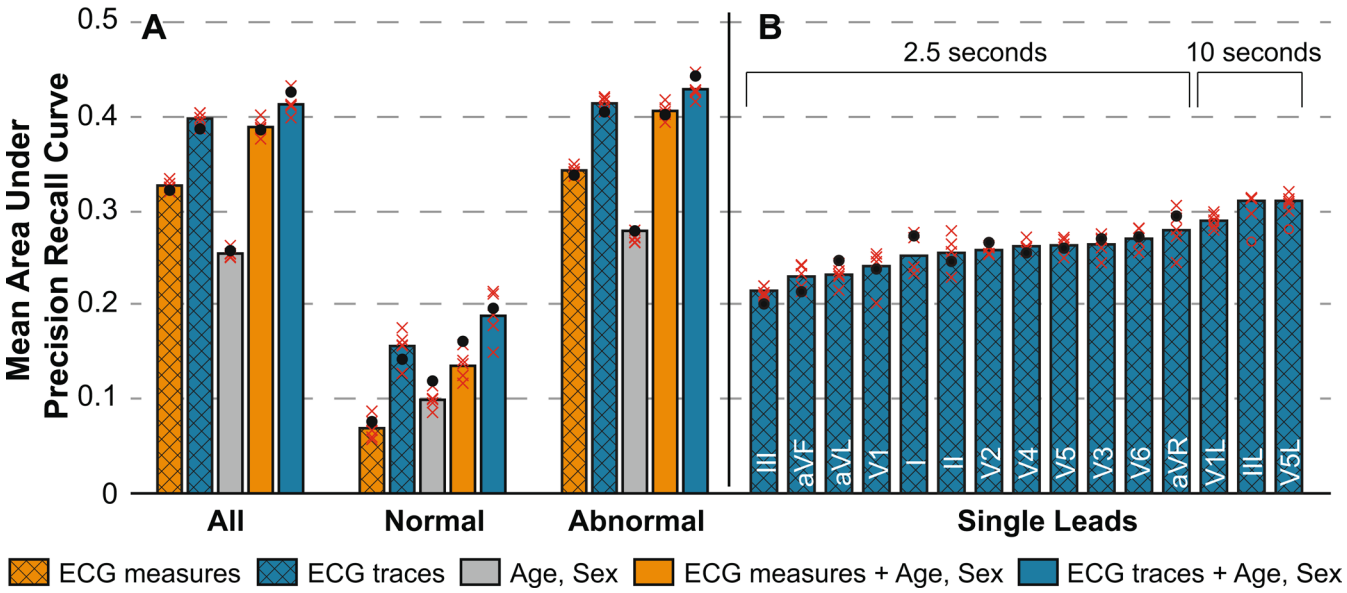
Reprints and permissions information is available at www.nature.com/reprints.



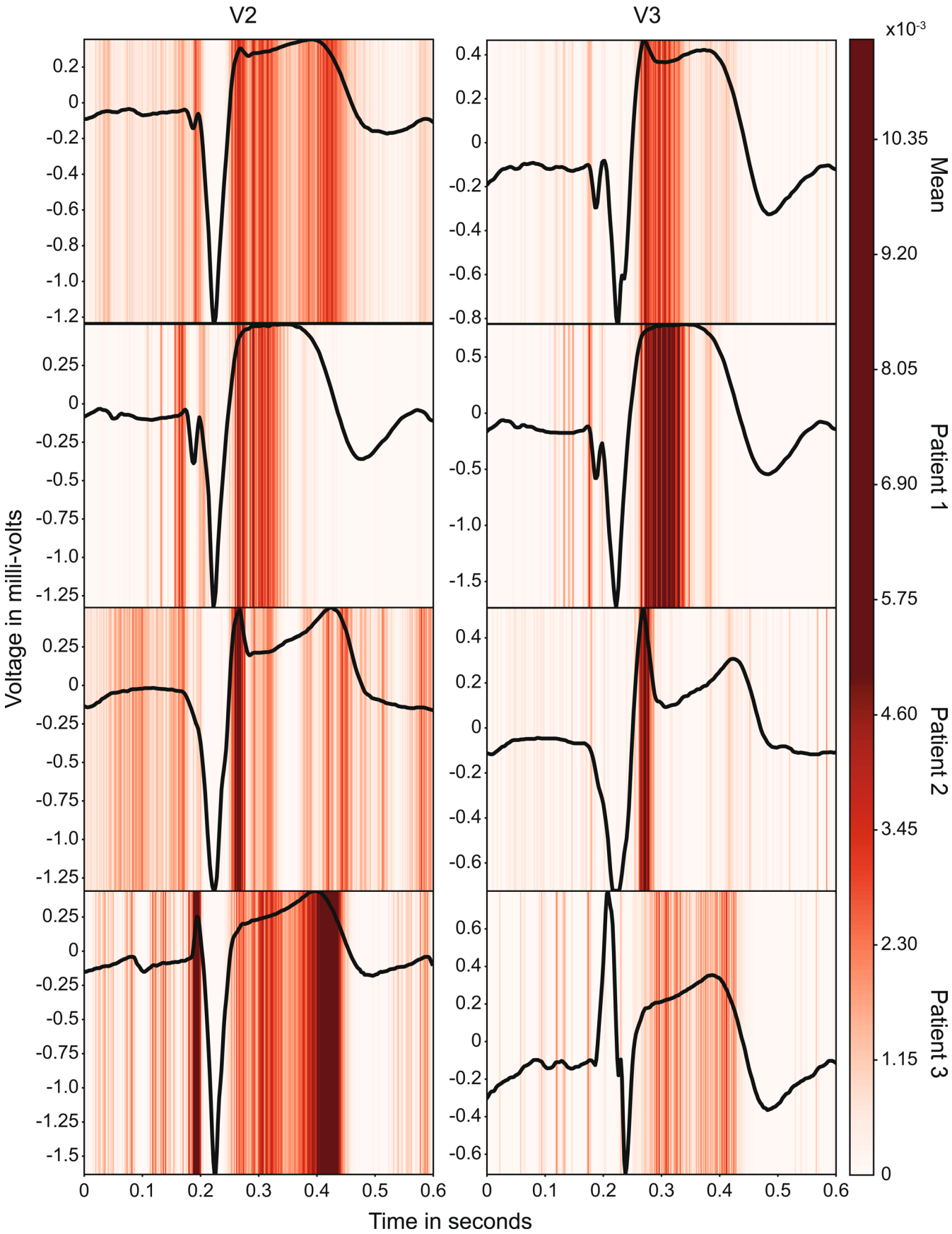
Extended Data Fig. 1 | Summary of the data used in the study. Summary of data used in the study. Note that '15 traces' means the standard 12 'short duration' leads (2.5 seconds of voltage data for each) plus 3 'long duration' leads (10 seconds of voltage data for each). PDF = portable document format. CV = cross-validation.



Extended Data Fig. 2 | Model Architecture. Model architecture used in the study.

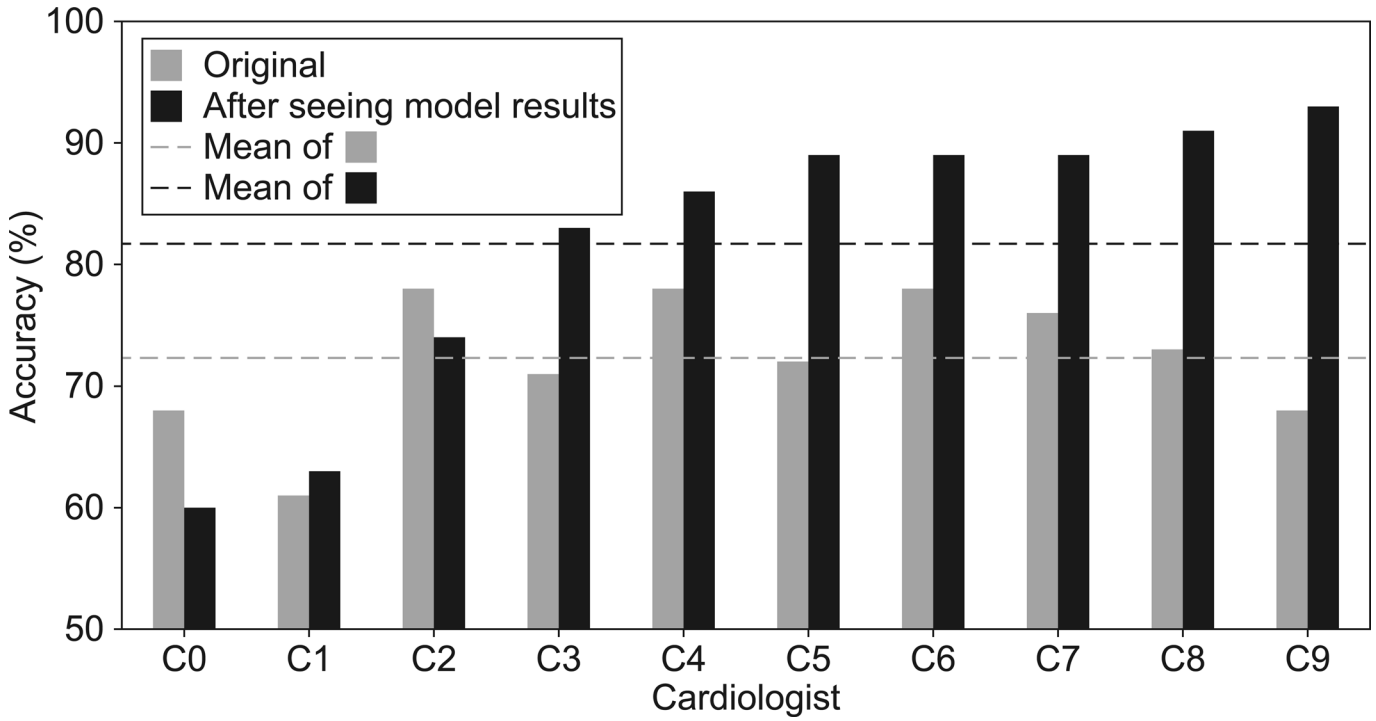


Extended Data Fig. 3 | Model performance as area under precision recall curve. Summary of model performance as area under precision recall curve (AUPRC) to predict one-year mortality. (A) The mean AUPRC for the indicated input data, including (i) clinically-acquired ECG measures (9 numerical values and 31 diagnostic labels), (ii) ECG voltage-time traces only, (iii) age and sex alone, (iv) ECG measures with age and sex, and (v) ECG voltage-time traces with age and sex. Models for (i), (iii) & (iv) used XGBoost and models for (ii) & (v) used a DNN. 'Normal' refers to the ECGs in the test set labeled as normal by the original interpreting physician at the time of ECG acquisition, 'abnormal' refers to any ECGs not identified as normal in the test set and 'all' includes both normal and abnormal ECGs in the test set. (B) The relative performance of the DNN models using single leads as input (sorted by increasing performance). The mean AUPRC of the models M1-M5 (derived from 5-fold cross-validation, see text) are shown as the bar heights while individual data points for each of the 5 models are shown as a red 'x'; black dots represent the AUPRC of model M0 (trained on 60% of the data and tested on the 40% holdout set). '2.5 seconds' and '10 seconds' refers to the duration of the voltage-time traces used for the model (see text for details).



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Model explainability with GRAD-CAM. The guided gradient class activation maps (guided Grad-CAM) overlaid on signal-averaged waveforms for three patients (bottom 3 rows) as well as mean signal and activation across patients (top row) for leads V2 and V3. Clinical ECG findings for all three patients reported anterior acute myocardial infarction with apparent ST segment elevations. Note that these patients were predicted high risk by the model, and all died within a year after this ECG (that is, they were considered 'true positives'). The overlay of the saliency map from guided Grad-CAM highlights the regions deemed salient (darker red regions) by the model towards prediction of high likelihood of mortality in a year, which coincided with the ST segment.



Extended Data Fig. 5 | Cardiologist visual survey. Accuracy for the ten cardiologists to correctly identify the true positive ECG (dead within a year) when presented with two 'normal' ECGs corresponding to a paired set of a true positive and true negative (n=100) (gray bars). Accuracy is also shown (black bars) for the same survey after being shown an independent set of paired ECGs (n=100) with outcomes labeled. All ECG pairs presented were matched for age and sex.

nature research

Corresponding author(s): Brandon K. Fornwalt MD PhD

Last updated by author(s): March 23, 2020

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.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom code was written in python (version: 3.5.2) to extract and parse the data from the clinical database.
Data analysis	Custom code in Python 3.5.2 was used to perform all the experiments and analyses. The packages and versions used for analyses are: Keras (version: 2.1.6-tf) with a TensorFlow backend (version: 1.9.0) and lifelines package (version: 0.21.5).

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. 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

All requests for raw and analyzed data and related materials, excluding programming code, will be reviewed by our legal department to verify whether the request is subject to any intellectual property or confidentiality constraints. 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 for non-commercial research purposes.

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	No sample size calculation was done for the main findings in the paper as the available data was large and was determined to be adequate. For a sub-analysis of review with cardiologists, one-sample binomial test power calculation was performed to determine the sample size.
Data exclusions	We excluded data with poor tracing, missing information and data from patients less than 18 years of age at the time of study (Extended Data Figure 1).
Replication	The experiments were performed with 5-fold cross-validation and successfully tested on true hold-out dataset.
Randomization	The train and test subsets were generated randomly with similar distribution of outcomes for model training and evaluation.
Blinding	For the cardiologist review of a subset of results, they were blinded to the results and process.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

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

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	The mean age of the population was 58 years (s.d: 18 years). Population consisted of 47% of males and 27% of the patients had normal ECG. (ECG measurement and abnormal ECG pattern distribution in Table 2).
Recruitment	This was a retrospective study without consent.
Ethics oversight	Geisinger institutional IRB approved this retrospective study with a waiver of consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N/A this was an analysis of retrospective data performed with a waiver of consent.
Study protocol	N/A this was not a trial
Data collection	No prospective recruitment was performed for this retrospective study.
Outcomes	The outcome studied was 1-year mortality, studied in retrospect, as detailed in the paper.