

RESEARCH

Open Access



# Deep survival analysis from adult and pediatric electrocardiograms: a multi-center benchmark study

Platon Lukyanenko<sup>1,2</sup>, Joshua Mayourian<sup>2,3</sup>, Mingxuan Liu<sup>4</sup>, John K. Triedman<sup>2,3</sup>, Sunil J. Ghelani<sup>2,3</sup> and William G. La Cava<sup>1,2\*</sup>

\*Correspondence:

William G. La Cava  
william.lacava@childrens.harvard.edu

<sup>1</sup>Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA

<sup>2</sup>Department of Pediatrics, Harvard Medical School, Boston, MA, USA

<sup>3</sup>Department of Cardiology, Boston Children's Hospital, Boston, MA, USA

<sup>4</sup>National University of Singapore, University Hall, Singapore

## Abstract

**Background** Artificial intelligence applied to electrocardiography (AI-ECG) has recently shown potential for mortality prediction, but heterogeneous approaches and private datasets have limited generalizable insights into AI methodologies fit for this purpose. To address this, we systematically evaluated model design choices across three large medical center cohorts: Beth Israel Deaconess (MIMIC-IV;  $n = 795,546$  ECGs, United States), Telehealth Network of Minas Gerais (Code-15;  $n = 345,779$ , Brazil), and Boston Children's Hospital (BCH;  $n = 255,379$ , United States).

**Results** We comprehensively evaluates models to predict all-cause mortality, comparing horizon-based classification and deep survival methods various neural architectures including convolutional neural networks and transformers. We also benchmarked against demographic-only and gradient boosting baselines. Top models yielded good performance (median concordance, Code-15: 0.83; MIMIC-IV: 0.78; BCH: 0.81). Incorporating age and sex improved performance across all datasets. Classifier-Cox models exhibited site-dependent sensitivity to horizon choice (median Pearson's R, Code-15: 0.35; MIMIC-IV:  $-0.71$ ; BCH: 0.37). External validation reduced concordance, and in some cases, demographic-only models outperformed externally trained AI-ECG models on Code-15. However, models trained on multi-site data outperformed site-specific models by margins ranging from 5% to 22%.

**Conclusions** These findings highlight several key factors for robust AI-ECG deployment. Deep survival methods consistently provided advantages over horizon-based classifiers, while inclusion of demographic covariates such as age and sex improved predictive performance across sites. The sensitivity of classifier-based models to horizon selection underscores the need for site-specific calibration. The multi-site experiment reveals that cross-cohort training, even between adult and pediatric cohorts, can substantially improve performance on those cohorts compared to cohort-specific training. Together, these results emphasize the importance of model type, demographic features, and training data diversity in developing AI-ECG models that can be reliably applied across populations.



**Keywords** Electrocardiography, AI-ECG, Convolutional neural networks, Survival analysis

## Introduction

Electrocardiography (ECG) measures the electric activity of the heart. Abnormal 12-channel ECGs often indicate cardiovascular pathology and are thus a marker for disease and mortality. AI-ECG refers to the application of AI or machine learning to ECG interpretation. A common AI-ECG task is risk prediction, which can be framed as predicting event occurrence (e.g. mortality or passing a diagnostic threshold). Several studies have recently applied AI-ECG to predict patient outcomes including mortality risk [1–5], ventricular hypertrophy [6], and ventricular dysfunction in both adults [7] and children [8]. While these studies demonstrate AI-ECG’s potential value, most use private data and only report a single, or a small set, of modeling approaches. Consequently, there is little consensus on best approaches to AI-ECG model development in general, including basic task formulation (e.g., classification or time-to-event prediction), the choice of model architecture, covariates, and other training procedures.

Most studies develop AI-ECG models on single, large, private ECG datasets (e.g., ECGs: 2.4M [9], 2.3M [4]; 1.2M [10]); the majority report results for a single AI-ECG model configuration (e.g [1, 2, 9, 10, 15]), and still fewer make models publicly available (e.g [1, 15]). Another notable challenge in the mortality prediction setting is for site-specific models to generalize to new sites; Sau et al. [10] and Hughes et al. [1] report model performance drops (AUROC) of 6–20% in external testing.

Verifiably evaluating different modeling approaches and creating reproducible public ECG models requires large, public ECG datasets linked to high-quality patient-level data. Fortunately, such data has recently become more available with the public releases of the MIMIC-IV-ECG dataset (~ 800k ECGs) in 2023 [13, 14] and the Code-15 dataset (~ 345k ECGs) in 2020 [15], vastly expanding on pre-2020 datasets [16].

Given the high interest in risk prediction from ECGs and the availability of these new datasets, the primary goal of this study is to benchmark state-of-the-art deep learning architectures and survival modeling approaches to inform AI-ECG best practices. To do so, we develop AI-ECG models of all-cause mortality using Code-15, MIMIC-IV-ECG, and a pediatric dataset from Boston Children’s Hospital (BCH). We quantify model performance as a function of model architecture, survival analysis approach, and covariate inclusion choice, across datasets, classifier time horizons, and patient subgroups. We create robust baseline models to contextualize AI-ECG performance relative to traditional machine learning approaches and simple baselines. Finally, we evaluate the impact of cross-site training on model generalization and site-specific performance.

## Background

Most AI applications to ECG signals have considered simpler prediction tasks than mortality; for example, many studies focus on arrhythmia detection, where classical AI methods leveraging feature extraction regularly report accuracies exceeding 99% [31]. While such performance is, in part, made possible by multi-center ECG benchmarking challenges [32], these have yet to explore mortality and tend to focus on longer signals with lower patient counts.

Mortality prediction from ECGs is a more complex task, given the myriad pathologies that influence risk, yet has considerable utility from a clinical point of view, as it provides a general way for providers to risk-stratify patients and manage treatment. Recent studies have demonstrated 5-year mortality prediction from ECGs, beginning with Raghunath et al. in 2020 [4] and, more recently Hughes et al.'s work in 2023 [1]. Many studies utilize the time-series-adapted residual network that was proposed by Ribeiro et al. in 2020 [15]. However, these works did not fully consider mortality as a time-to-event prediction (i.e., as a survival analysis problem), nor did they utilize state of the art approaches to deep learning for survival analysis [20]. Most recently, Sau et al. [10] demonstrated survival analysis from ECGs using a deep survival approach called DeepHit [24]. However, the study does not provide code or sufficient methods details to support future research, which is one of our motivations here.

Our study covers these methods and benchmarks several additional approaches. For example, we benchmark a recent transformer-based method called ECGTransform [30], as well InceptionTime, which is considered a state-of-the-art time series classification architecture [26]. We benchmark three additional deep survival approaches. In contrast to prior studies, we evaluate the impact of cross-site training, benchmark on two large, publicly available databases, and make our code and model benchmarks publicly available to support downstream scientific research.

## Methods

### Patient populations

Code-15 [15] is an ECG dataset from the Telehealth Network of Minas Gerais, a Brazilian public agency providing telehealth services to Minas Gerais and Amazonian and Northeast states. Patient ECGs were recorded in primary care facilities by technicians and examined remotely by a cardiologist. The publicly available dataset includes 345,779 ECGs collected from April to September of 2018. Code15 ECGs are 7–10 s signals sampled at 400 Hz, centered and padded with zeros to total a length of 4096. The Code-15 Dataset provides multiple ECGs per subject and indicates the patient's age and the follow-up time after the patient's final ECG. We only use the one entry per subject that provides a specific follow-up time.

MIMIC-IV-ECG [13], referred to as simply MIMIC-IV for brevity, includes 795,546 ECGs from 159,608 patients collected between 2008 and 2019 at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Patient ECGs were recorded in various settings, including emergency settings, hospitals, and outpatient care centers. Each ECGs is a 10 s signal sampled at 500 Hz.

MIMIC-IV tracks date-of-death with state and hospital records, and censors deaths one year after a final recorded hospital visit. These signals have channels 4 and 5 switched compared to Code-15 and are altered to match the Code-15 format. Machine measures are used "as is" when generating models, as neither normalization nor indicator variables improved results in preliminary analyses.

BCH includes 225,379 ECGs from 79,568 patients collected 1990 to 2018 at Boston Children's Hospital in Boston, Massachusetts from emergency, hospital, and outpatient care settings [17]. In contrast to the adult cohorts, this dataset predominantly represents a pediatric congenital heart disease cohort. BCH ECGs are 8-second signals sampled at 250 Hz. Death was tracked by an internal institutional database. Contrasting with adult

data, pediatric ECGs have higher heart rates, a relatively larger right ventricle, and a strong age dependency; pediatric mortality is also more likely to have a congenital cause.

### Dataset preparation

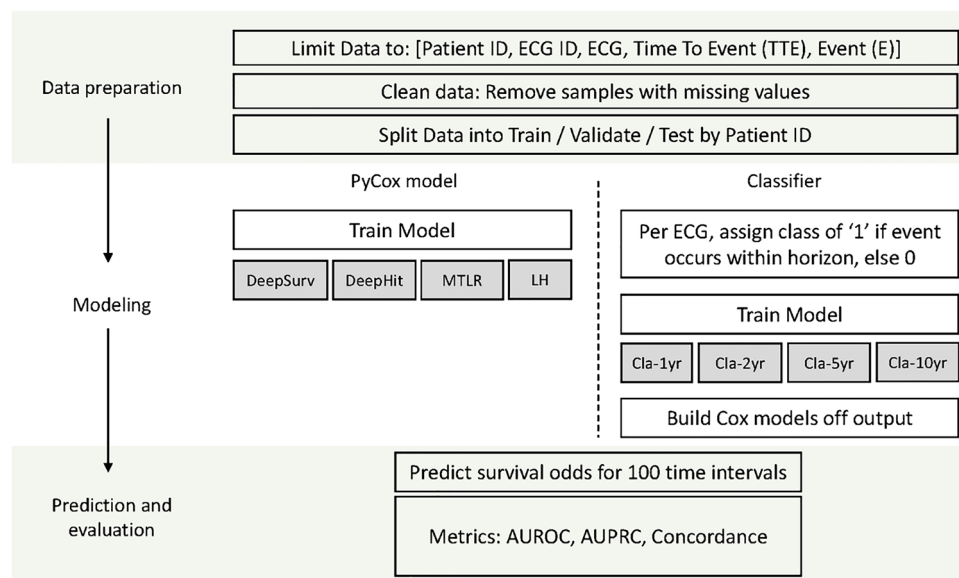
Dataset preparation is shown in Fig. 1. Data is limited to [Patient ID, Time-To-Event, Event, ECG, Age, Sex, Machine Measures (MIMIC-IV only)], and entries with missing ECG samples or Time-To-Event are excluded. Time-To-Event is set to a minimum of 1 day. All ECGs are limited to 7 s, resampled to 400 Hz, and symmetrically padded with zeros for an end shape of  $12 \times 4096$ . ECGs are split 64/16/20 into train/validation/test sets by patient ID (Figure S1). We do not limit the number of ECGs per patient. The random seed determines the Training/Validation split while the Test set remains fixed. Before model training or evaluation, all ECGs are z-score normalized per ECG channel based on the training set.

### Survival analysis

Survival analysis builds survival functions,  $S(t)$ , denoting the probability of not having experienced an event by a time  $t$ . Survival analysis uses data in the form [Time-To-Event, Event], which denotes event state at a follow-up time. This leverages information even when final event time is unknown: a device that works for two years before being lost still provides two years of evidence of non-breakage. ‘Censoring’ refers to not knowing event outcomes for some subjects.

The most common survival functions are Kaplan-Meier curves which display survival over time from a tracked population. If the population is clustered into groups, an individual’s trajectory can be estimated from their cluster’s survival function. Otherwise, survival functions usually fit an exponential decay or Weibull function to regressors (demographics, measurements, etc.).

The Cox proportional hazards model [18] models hazard,  $h(t) = d/dt (1 - S(t))$ , as an unknown base function scaled by exponential decays:  $h(t) = h_0(t)\exp(\beta_1 \times x_1 + \beta_2 \times x_2 + \dots)$



**Fig. 1** Flow diagram describing the experimental design

$\beta_m x_m$ ) where  $\{x_1, x_m\}$  are regressors and  $\{\beta_1, \beta_m\}$  are learned weights. The Breslow estimator is often used to fit  $h_0(t)$  [19].

The Cox regression assumes that *log-risk* ( $\beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \beta_m x_m$ ) is *proportional* (e.g. smoker risk / non-smoker risk = constant) and a *linear* function of regressors that is *time-constant* (there are no time-regressor interactions).

Broadly, there are two survival modeling approaches with neural networks:

**Classifier-Cox (CC)** The first approach converts inputs into a set of values through a neural network. These values are then treated as generic markers to fit a Cox regression [18]. The neural networks are typically trained to return a single value classifying inputs by whether an event occurs within a time horizon or not. This approach is easily implemented and can account for new covariates in the Cox regression stage but must handle data censoring when assigning classifier labels. We tune models to minimize binary cross-entropy loss at a fixed horizon time, sweeping several horizons.

**Deep survival (DS)** The second approach trains neural networks with survival-specific loss functions to directly generate survival curves. We evaluate four different Deep Survival approaches from the PyCox [20] package. DeepSurv [21] models *log-risk* as a *time-constant, proportional, non-linear* function. LogisticHazard (LH) [22] models *log-risk* as a *time-varying, non-proportional, non-linear* function and attempts to eliminate batch size effects. MTLR [23] models *risk* with a *time-varying, non-proportional, logistic regression* on a *non-linear* function. DeepHit [24] models *risk* as a *time-varying, non-proportional, non-linear* function and allows for several competing risks – i.e., multiple possible outcome events – which we do not exploit in this study. These survival models were chosen for their variety and standardized implementation with PyCox, which minimizes loss functions customized to each approach.

### AI-ECG survival models

Several recent studies predict mortality from ECGs. These are either built completely on large, private, ECG banks [3–5, 10, 16, 25] or are fine-tuned in the context of a particular setting or population [2, 11, 12]. Some studies use Deep Survival [3, 10, 11], and some use Classifier-Cox approaches [4, 5, 25]. Most studies use a convolutional network to interpret ECG, with a slightly-more-popular choice being the ResNet architecture from Ribeiro et al. [5, 10, 15, 17, 25].

### Benchmark procedure

Benchmark parameters are in Table 2. In total, more than 3,000 AI-ECG models are analyzed. We use two convolutional neural network (CNN) architectures, InceptionTime

**Table 1** Settings for the mortality prediction benchmark experiment

Setting	Values
Survival Analysis	
Deep-Survival	DeepHit, DeepSurv, LogisticHazard, Multi-Task Logistic Regression
Classifier-Cox	Classification mortality time horizon (years): 1,2,5,10
AI-ECG Architecture	InceptionTime (CNN), Resnet (CNN), ECGTransformer
Covariate options	None, Age + Sex, Age + Sex + ECG Machine Measure (MIMIC-IV only)
Demographic-only models	XGBoost, Feedforward Network
Normalization	z-score per channel, based on model's training data

[26] and a modified ResNet architecture [15], and one transformer architecture, ECG-Transform [30]. For each architecture, we train four deep survival models and four classifier-Cox models where neural net classifiers predict mortality up to one, two, five, and ten-year horizons. Models are trained with and without age and sex covariates. Additional modeling details are in the Supplement. We make this resource publicly available (see Data and Resource Availability Section).

### Neural network architectures

Our neural networks had three modules: an ECG-processing module, a covariate-processing module (three feedforward-ReLU layers with output dimension 32), and a fusion module connecting the other two pieces (three feedforward-ReLU layers with output dimension 128) that heads into a final linear layer (dimension 1 for Classifier-Cox models and DeepSurv, dimension 100 for LH, MTLR, and DeepHit – one per time bin). When covariates are not included, the covariate-processing module is skipped. When ECG is excluded for covariate-only baselines, the ECG module returns a one-dimensional value of '0', resulting in a feed-forward network.

Along with the survival approaches, we benchmark three neural network architectures. “ResNet” is a multi-channel time series [15, 25] adaptation of the original ResNet [27]. Architecture parameters were kept at defaults tuned to the full Code dataset. This model has 6.9-7.5 M parameters. “InceptionTime” [26] adapts AlexNet [28] by widening convolutional kernel widths and including channel-wise bottlenecks to control model complexity. This architecture performs well on many small time-series classification benchmarks and has recently been used in fetal heart rate monitoring [29]. Architecture parameters were kept at publication defaults (kernel widths 11, 21, 41). This model has 510-530k parameters. “ECGTransForm” is a recent transformer architecture combining multi-scale convolutions with a bidirectional transformer. It reports state-of-the-art performance in arrhythmia classification and has 2.6 M parameters [30]. These architectures were chosen for their use in ECG or physiological time series analysis and adapted from existing repositories.

### Baselines

We generate baseline models by training XGBoost and feedforward networks (‘FF’) on age and sex covariates. In addition, we leverage the automated machine measures available in MIMIC-IV (Axes – P, QRS, T; Durations – P, PQ, QRS, QT, RR.). We generate a baseline model using these measures, along with age and sex, as a baseline feature extraction modeling approach. We additionally test providing measures to AI-ECG models.

### Outcome measures and statistics

Our primary outcome measure is the Concordance Index (concordance). Like area under the receiver operating characteristic curve (AUROC), concordance evaluates the ability of a model to rank patients correctly by risk, but also takes into account the timing of the event. At the time of an event, a subject should be at a higher risk than any other subject still under observation; concordance is the fraction of correctly ranked subject-subject comparisons. Additional measures (AUROC, AUPRC, time-censored concordance) are available in the Supplemental Material.

Unless otherwise mentioned, paired comparisons use the Wilcoxon test and unpaired comparisons use the Mann-Whitney test. Multiple comparisons are adjusted for with Benjamini-Hochberg corrections at a 5% false discovery rate.

## Results

### Patient populations

The patient cohorts resulting from our data preparation and inclusion criteria are summarized in Table 1. Overall, our experiment is conducted over 1,200,658 ECGs from 443,216 patients.

### Local model concordance by architecture, survival approach

The concordance of site-specific models, across model configurations, input and datasets, is summarized in Fig. 2. Because transformers showed markedly worse performance, their results, along with additional baselines, are detailed in Table S1. The following analysis, statistics, and discussion focus on the CNN and competitive baseline models.

The model configurations with best test set performance on each dataset are shown in Table 3, including performance at the 1 and 5 year marks. For cross-reference, chance mortality rates per year are in Table S2. For MIMIC-IV-ECG, appending automatic machine measurements as covariates did not noticeably improve model performance.

### Classifier-Cox performance is sensitive to time horizon

We find that classifier-Cox horizon choice (the time at which binary mortality is determined) correlates with median model concordance in all cases (Table 4). This correlation can be positive (Code-15, BCH) or negative (MIMIC-IV), hindering the selection of a generally appropriate value.

### Deep survival approaches outperform Classifier-Cox approaches

Among local models trained with ECG and covariates, Deep Survival (DS) models, overall, show statistically higher concordance than Classifier-Cox (CC) models across sites ( $p < 1.98E-03$ , Table S3), although the performances differences are seldom large. On cross-evaluation, DS outperforms CC in 4/6 cases and CC outperforms DS in 1/6.

### CNN model architectures perform similarly, and outperform transformer architectures

For models trained with ECG and covariates, InceptionTime (IT) outperforms ResNet (RN) on the BCH dataset; on cross-evaluation, RN outperforms IT in 3/6 cases and IT outperforms RN in 1/6 cases (Table S4). Transformers, even on top individual runs (concordance, Code-15: 0.793, MIMIC-IV: 0.762, BCH: 0.774) do not exceed median Deep Survival CNN model concordance (Table S1).

### Including demographic covariates improves model performance

In paired evaluations, AI-ECG models that included age and sex covariates always outperform ECG-only models ( $p < 2.05E-4$ , Table S5; Fig. 2), improving concordance an average of 5.2%, and as much as 11.7%.

**Table 2** Population characteristics for the three sites used for benchmarking

Site	Measure	All	Event (Mortality)	Non-Event	p value	
Code-15 (2018)	ECGs, N (%)	233,647	8,341 (3.6)	225,306 (96.4)		
	Patients, N (%)	233,647	8,341 (3.6)	225,306 (96.4)		
	Sex, N (%)	Female	138,911 (59.4)	3,853 (2.8)	135,058 (97.2)	<< 1E-5
		Male	947,36 (40.6%)	4,488 (4.7)	90,248 (95.3)	
	Mean Age, Yr (Std)	50.7 (19.8)	70.4 (14.0)	50.0 (19.6)		<< 1E-5
	Mean Follow-up Time, Yrs (Std)	3.7 (1.9)	2.0 (1.6)	3.7 (1.8)		<< 1E-5
MIMIC-IV (2008–2019)	ECGs, N (%)	785,035	215,039 (27.4)	569,996 (72.6)		
	Patients, N (%)	159,122	25,107 (15.8)	134,015 (84.2)		
	Sex, N (%)	Female	384,923 (49.0)	99,170 (25.8)	285,753 (74.2)	<< 1E-5
		Male	400,112 (51.0)	115,869 (29.0)	284,243 (71.0)	
	Mean Age, Yrs (Std)	62.3 (17.1)	70.4 (14.2)	59.2 (17.1)		<< 1E-5
	Mean Follow-up Time, Yrs (Std)	2.6 (2.6)	2.0 (2.4)	2.9 (2.6)		<< 1E-5
BCH (1990–2018)	ECGs, N (%)	181,976	12,638 (6.9)	169,338 (93.1)		
	Patients, N (%)	50,447	16,71 (3.3)	48,776 (96.7)		
	Sex, N (%)	Female	86,141 (47.3)	6,048 (7.0)	80,093 (93.0)	0.23
		Male	95,835 (52.7)	6,590 (6.9)	89,245 (93.1)	
	Mean Age, Yrs (Std)	12.8 (11.9)	21.5 (16.7)	12.2 (11.2)		<< 1E-5
	Mean Follow-up Time, Yrs (Std)	8.9 (6.7)	7.2 (6.1)	9.0 (6.7)		<< 1E-5

Categorical p-value comparisons use the Chi-square test; numerical comparisons use Student's t-test

### Multi-site model training improves both global and local model performance

Figure 3 shows the performances of ResNet AI-ECG models in local and external evaluation across sites. As expected, models perform best when trained, at least partially, on data generated from the same population as their respective test sets. Covariate-only models (age and sex models) provide surprisingly strong baselines and generalize well between sites. This is especially noteworthy for Code-15: covariate-only models generalize to Code-15 cohort better than CNNs trained on other single site data.

The rightmost column of Fig. 3 also shows the performance of ResNet models trained on data from all sites. These provide maximum, or near-maximum, performance on all datasets, across modeling approaches. In Fig. 4, we more directly compare the performance of locally and globally trained ResNet models on each site, which demonstrates that globally trained models outperform site-specific models, even on test data from the sites on which the local models are trained.

We investigate site-specific model performances across test cohorts on age and sex subgroups in Figure S4. Although cohorts differ by age and sex which may make generalization challenging, we do not find clear evidence that performance differences are explained by these age or sex differences.

### Survival models capture expected survival rates

We compare Kaplan-Meier curves for Code-15 and MIMIC-IV to mean predicted population survival from ECG-only models in Figure S2, confirming that model estimates closely match expected survival rates throughout the study period.

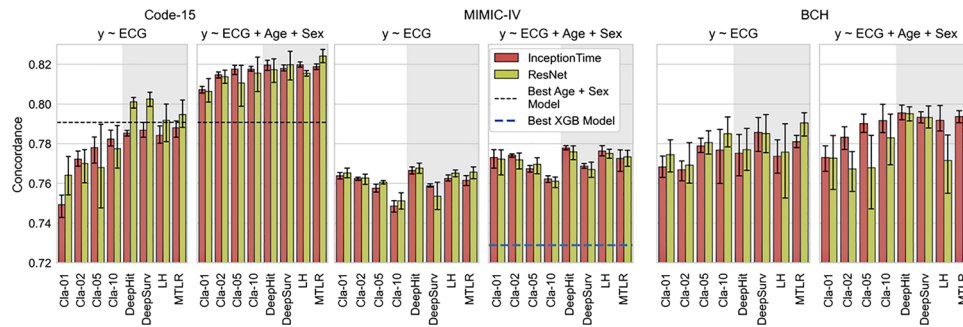
### Model explanations exhibit salient risk factors for mortality

We investigate model explainability with a median-waveform analysis [8] in Fig. 5. The median high-risk waveforms show lower amplitude and QRS complexes, as well as flattened/inverted lateral precordial T waves indicating several pathological findings (delayed myocardial activation, possibly myocardial strain). The QRS complex is most salient, suggesting focus mostly on myocardial activation (heartbeat dynamics). High salience of the V2 lead indicates anteroseptal activation and aligns with analyses of left ventricular (LV) dysfunction. Model explainability across various heart lesions for a ResNet Classifier-Cox approach is also explored in prior work [17].

## Discussion

### Contribution

This work is the first to benchmark AI-ECG survival modeling approaches in all-cause mortality prediction on three broad datasets: Code-15, MIMIC-IV, and BCH. We provide verifiable results and baselines that inform ECG-processing and survival-analysis approaches for future AI-ECG modeling. We also release our code, which can be adapted to evaluate new modeling approaches or used to train AI-ECG survival-analysis models. This can also be applied beyond mortality modeling to model event occurrence (e.g. first diagnostic threshold crossing) from data in the [time-to-event, event] format. Our work may support future studies that develop, implement, and test AI-ECG models that prospectively inform patient care and case management.



**Fig. 2** Model concordances across datasets (subplots), covariates (columns), survival modeling approaches (x-axis), and architectures (color). Hatched bars indicate deep survival approaches. “Cla-X” marks X-year horizon Classifier-Cox models. “Best XGB Model” is a Cla-2 model trained on demographics and automatic ECG machine measures. Error bars denote 95% confidence intervals of mean concordance over 5 random seeds. Error bars denote 95% confidence interval

**Table 3** Concordance and AUPRC of top-performing, site-specific models, per dataset and input configuration

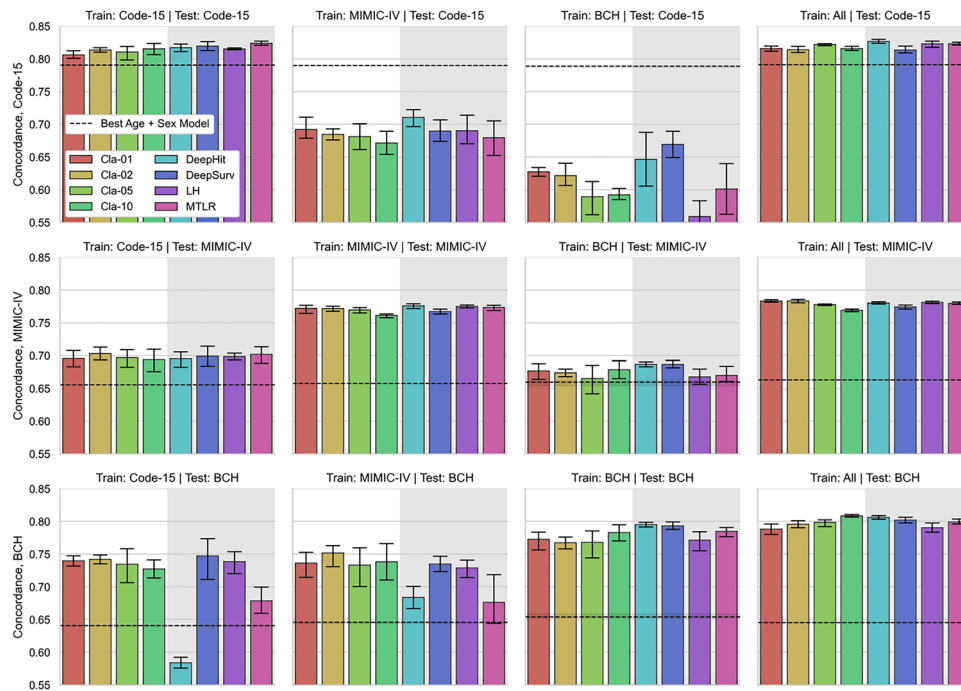
Dataset	Model Inputs	Approach	Architecture	Concordance	1 Year AUPRC	5 Year AUPRC
Code15	y ~ ECG	DeepSurv	ResNet	0.80 (0.80–0.81)	0.07 (0.07–0.07)	0.14 (0.14–0.14)
	y ~ ECG + Age + Sex	MTLR	ResNet	0.83 (0.82–0.83)	0.07 (0.07–0.07)	0.14 (0.14–0.15)
MIMIC-IV	y ~ ECG	DeepHit	ResNet	0.77 (0.77–0.77)	0.42 (0.42–0.42)	0.49 (0.49–0.49)
	y ~ ECG + Age + Sex	DeepHit	InceptionTime	0.78 (0.78–0.78)	0.45 (0.44–0.45)	0.52 (0.51–0.52)
	y ~ ECG + Age + Sex + Machine Measures	DeepHit	ResNet	0.77 (0.77–0.78)	0.43 (0.43–0.44)	0.46 (0.45–0.48)
BCH	y ~ ECG	MTLR	ResNet	0.79 (0.79–0.80)	0.05 (0.04–0.05)	0.13 (0.13–0.14)
	y ~ ECG + Age + Sex	DeepHit	ResNet	0.79 (0.79–0.80)	0.05 (0.05–0.06)	0.14 (0.13–0.14)

For cross-reference, mortality rates per year are in Table S2

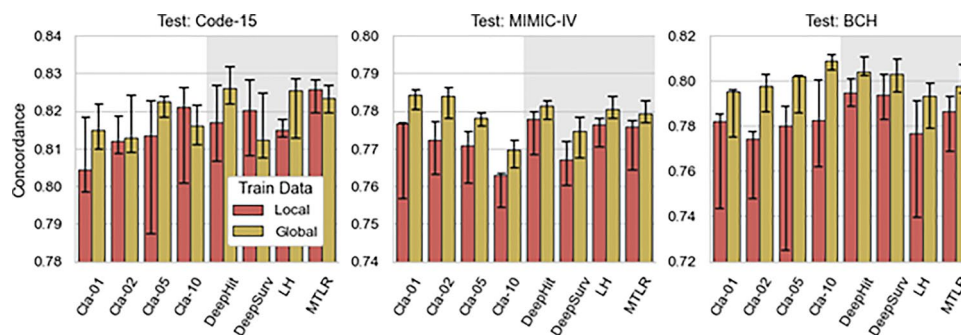
**Table 4** Classifier-Cox models are sensitive to the choice of time horizon

Dataset	ECG + Demographic		ECG Only	
	Pearson R	p value	Pearson R	p value
Code-15	0.35	2.66E-02	0.41	8.89E-03
MIMIC-IV	-0.71	2.66E-07	-0.88	1.08E-13
BCH	0.37	1.95E-02	0.47	2.35E-03

Pearson’s R correlation between classification model concordances and the choice of time horizon used in training. The direction of the effect is positive in BCH and Code-15 and negative in MIMIC-IV



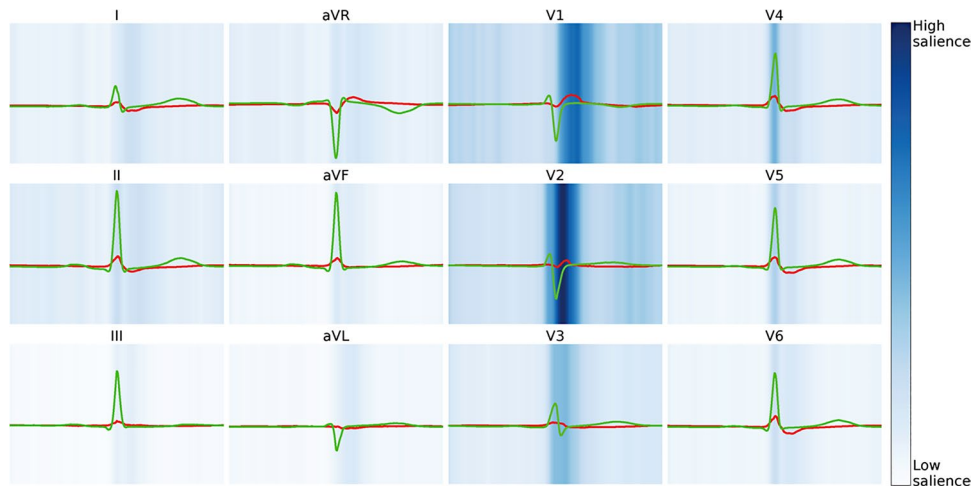
**Fig. 3** Model concordance in cross-evaluation. Rows denote test performance on Code-15, MIMIC-IV, and BCH, respectively. Columns denote which site the models were trained on. The rightmost column shows concordance for models trained across the training sets of all three sites. Bar color denotes the survival analysis approach; hatches denote deep survival approaches. The performance of the best demographic only baseline model (trained on age + sex) is shown by a dotted line. Results here are for the ResNet model architectures trained with ECGs and demographics. Error bars denote 95% confidence interval



**Fig. 4** Multi-site training improves local performance. Median concordance of mortality models trained on site-specific data (light brown) versus training data from each site (red). Subplots denote the site-specific performance of these models, from left to right, on data from Code-15, MIMIC-IV-ECG, and BCH, respectively. X-axis denotes the modeling approach; all approaches use the ResNet architecture and include demographic covariates. Error bars denote 95% confidence interval

**Deep survival approaches outperform Classifier-Cox approaches**

In this study we find Deep Survival approaches work better than Classifier-Cox approaches in two respects. First, the median Deep Survival models significantly increase model concordance over Classifier-Cox models. Second, deep survival approaches do not require a time horizon to be specified, and classifier-cox models exhibit performance sensitivity to this horizon choice. Our results indicate that horizon choice is both dataset-specific, with opposing trends for Code-15 and MIMIC-IV, and can substantially impact performance with the wrong choice. For example, short-horizon data includes



**Fig. 5** Median heartbeats across 12-lead ECG and SHAP analysis for a MIMIC-IV DeepHit ResNet. Green: median heartbeat for 100 ECGs with lowest-risk-estimate (mean prediction probability: 0%). Red: median heartbeat for 100 ECGs with highest-risk-estimate (mean prediction probability: 69%). Blue shading: SHAP saliency. Model explanations highlight dampened QRS activity, indicative of LV dysfunction, in high-risk patients. For details of median waveform analysis, see [8]

few positive-event samples but assumptions about censored patients have little effect. Meanwhile, long-horizon data has more positive-event samples but the assumptions about censored patients have additional time to compound. Furthermore, when training classifiers it is necessary to combine (Time-to-Event, Event) data into a single variable to train a classifier (we discuss our approach in the Supplement). This choice can have substantial effects on what data models use (e.g. MIMIC-IV data is censored to 1 year unless linked to follow-up measures or a mortality). Deep Survival modeling does not require such a choice.

#### Comparisons to past work and model generalizability

Our results confirm the performance of ECG mortality risk prediction models from past studies at the one-year mark (0.81–0.85 concordance) [4, 5] and five-year mark (concordance: 0.78–0.83) [1, 5]. Additionally, our ResNet DeepHit, ECG-only MIMIC-IV results (concordance: 0.77) concur with a recent study that trained a ResNet DeepHit model, dubbed “AIRE” [10], on 1.2 M, 10-second ECGs from Beth-Israel Deaconess Medical Center in Boston, MA (concordance: 0.775). The results stand to reason as we expect this dataset to largely overlap with MIMIC-IV.

Prior work developing AI-ECG mortality prediction models reports performance deterioration of different degrees across sites, from 0 to 6% AUROC differences (SEER [1]), to 1–18% concordance drops (AIRE [10]). Follow-up evaluations of SEER indicate larger performance deterioration on the UK Biobank cohort (UKB AUROC 0.57 [10], vs. 0.83 in Stanford cohort [1]). Our results confirm that AI-ECG models trained on site-specific data struggle to generalize to new sites, across AI-ECG models, and further suggest that these differences are not fully explained by demographic differences between cohorts (Figure S3) nor mortality rate differences (Table S2). However, by training on multiple sites, our study demonstrates that performance improves across cohorts and, importantly, outperforms site specific models.

### **The importance of baseline models and demographic covariates**

Past work rarely compares AI-ECG to simpler models, such as demographic-only models, as we do here, yet AI-ECG's value depends on its favorable performance to less complex baselines. Our results have important implications for studies reporting generalization performance on Code-15 [15], where demographic-only baseline models (median concordance: 0.79) outperform the AI-ECG models tested here, as well as those in prior work (e.g. AIRE concordance: 0.76 [10]).

We conjecture that AI-ECG models may have more value in acute care settings (i.e., MIMIC-IV, BCH) than in outpatient settings (Code-15) due to the clinical context. In general, our results strongly support the include of age and sex covariates in all AI-ECG mortality prediction models.

### **Towards clinical translation**

We hope that this benchmark can assist in AI-ECG modeling by characterizing higher-performing ECG mortality modeling approaches, giving expected ranges for in-dataset and cross-dataset evaluation, and enabling easier model training and evaluation through publicly released code. In turn, these models could help develop tools for automated patient risk estimation: just as current systems flag potential anomalies, patients could be flagged for high estimated risk, potentially informing intervention or follow-up timing (i.e., determining how closely to follow patients).

### **Limitations**

While we conducted broad sweeps of model configurations in this study, they are not exhaustive. We cannot rule out that other model architectures, or fine-grained tuning of the architectures we benchmark, may improve performance on specific datasets. Furthermore, our results are specific to mortality risk prediction from AI-ECGs, which is one of several prediction targets of interest. Future work could expand the clinical interpretation of model behavior to understand specific pathologies leading to mortality, which we do not inspect deeply here. As a retrospective benchmark, this study may not fully capture deployment considerations; the ethical deployment of survival models in a clinical setting requires that model performance be evaluated in those contexts, with appropriate oversight and a focus on how model use impacts patient care in specific clinical workflows.

### **Conclusions and summarized recommendations**

Based on this study, we recommend the following for AI-ECG model development: (1) including demographic covariates; (2) using deep survival, rather than classification, approaches; (3) the 1-D ResNet architecture; (4) supplementing local model training with data from additional sites. These model choices generally provide better performance, avoid classifier difficulties related to horizon choice and data censoring, and rely on a well-vetted architecture. Taken together, these design choices can be responsible for most of the improvements AI-ECG offers over simpler baseline models, which we measure to be 5–25% median concordance improvement, depending on cohort. We hope this study and its recommendations provide a valuable resource for future ECG and ECG-mortality studies.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13040-025-00510-4>.

Supplementary Material 1

### Acknowledgements

We would like to thank Lorenzo Peracchio for help with data validation.

### Author contributions

Conceptualization, P.L., J.M., W.G.L.; methodology, P.L., W.G.L.; investigation, P.L., W.G.L.; writing—original draft, P.L., W.G.L.; writing—review & editing, J.M., S.G., J.K.T., W.G.L.; funding acquisition, J.K.T.; resources, W.G.L.; supervision, J.M., J.K.T., S.G., W.G.L.

### Funding

P.L., and W.G.L. were partially supported by the Kostin Innovation Fund at Boston Children’s Hospital. W.G.L. was partially supported by National Institutes of Health grant R01LM014300.

### Data availability

This paper analyzes two existing, publicly available datasets, accessible at:- MIMIC-IV-ECG: [<https://doi.org/10.13026/4nqg-sb35>] under the Open Data Commons Open Database License v1.0- Code-15: [<https://doi.org/10.5281/zenodo.4916206>], under Creative Commons license CC-BY 4.0- Both datasets need preprocessing to align to the 7-second Code-15 format, available in the repository below.- The BCH data reported in this study cannot be deposited in a public repository because of Institutional Review Board policies to protect pediatric patient privacy. To request access, contact the corresponding author.- Benchmarking and data preprocessing scripts are available at <https://github.com/cavalab/ecg-survival-benchmark> and is publicly available under the Gnu Public License v3.- The top deep survival models for MIMIC-IV and Code-15 cohorts are available on Zenodo at <https://doi.org/10.5281/zenodo.1687773>.

### Declarations

#### Ethics approval and consent to participate

The BCH data study used here was approved for a waiver of informed consent by Boston Children’s Hospital Institutional Review Board (IRB-P00044967).

#### Consent for publication

N/A.

#### Competing interests

The authors declare no competing interests.

Received: 12 September 2025 / Accepted: 12 December 2025

Published online: 17 December 2025

### References

1. Weston Hughes J, et al. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *Npj Digit Med*. 2023;6(1):169. <https://doi.org/10.1038/s41746-023-00916-6>. September 2023.
2. Salah S, Al-Zaiti, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nat Med*. 2023;29(7):1804–13. <https://doi.org/10.1038/s41591-023-02396-3>. July 2023.
3. Shaan Khurshid, et al. ECG-Based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*. 2022;145(2):122–33. <https://doi.org/10.1161/circulationaha.121.057480>.
4. Sushravya, Raghunath, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*. 2020;26(6):886–91. <https://doi.org/10.1038/s41591-020-0870-z>. June 2020.
5. Weijie, Sun et al. (2023). Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *Npj Digit Med*. 2023;6(1):21. <https://doi.org/10.1038/s41746-023-00765-3>.
6. Takahiro, Kokubo, et al. Automatic detection of left ventricular dilatation and hypertrophy from electrocardiograms using deep learning. *Int Heart J*. 2022;63(5):939–47. <https://doi.org/10.1536/ihj.22-132>.
7. Demilade, Adedinsewo et al. (2020). Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with Dyspnea. *Arrhythmia Electrophysiol*. 2020;13(8):e008437. <https://doi.org/10.1161/circep.120.008437>.
8. Joshua, Mayourian et al. (2024). Pediatric ECG-based deep learning to predict left ventricular dysfunction and re- modeling. *Circulation*. 2024;149(12):917–931. <https://doi.org/10.1161/circulationaha.123.067750>.
9. Hongling, Zhu et al. (2024). Four-channel ECG as a single source for early diagnosis of cardiac hypertrophy and dilation — a deep learning approach. *NEJM AI* 2024-09-26. <https://doi.org/10.1056/Aloa2300297>.
10. Sau A, et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. *Lancet Digit Health*. 2024;6(11):e791–e802. [https://doi.org/10.1016/S2589-7500\(24\)00172-9](https://doi.org/10.1016/S2589-7500(24)00172-9).
11. Van De Leur RR, et al. Electrocardiogram-based mortality prediction in patients with COVID-19 using machine learning. *Neth Heart J*. 2022;30(6):312–8. <https://doi.org/10.1007/s12471-022-01670-2>. June 2022.
12. Weimann K, Tim OF, Conrad. (2021). Transfer learning for ECG classification. *Sci Rep*. 2021;11(1):5251. <https://doi.org/10.1038/s41598-021-84374-8>.
13. Brian, Gow et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset, 2023. <https://doi.org/10.13026/4nqg-sb35>.

14. Goldberger L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215–e220.
15. Antônio H, Ribeiro et al. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun*. 2020;11(1):1760. <https://doi.org/10.1038/s41467-020-15432-4>.
16. Elena Merdjanovska and Aleksandra Rashkovska. (2022). Comprehensive survey of computational ECG analysis: Databases, methods and applications. *Expert Systems with Applications*, 203:117206, October 2022. <https://doi.org/10.1016/j.eswa.2022.117206>.
17. Joshua, Mayourian et al. (2024). Electrocardiogram-based deep learning to predict mortality in paediatric and adult congenital heart disease. *Eur Heart J*. 2024. <https://doi.org/10.1093/eurheartj/ehae651>.
18. David RC. Regression models and life-tables. *J R Stat Soc Series B Methodol*. 1972;34(2):187–202.
19. Terry M, Therneau. Modeling survival data: Extending the Cox Model. In: Bickel P, Diggle P, Fienberg S, Krickeberg K, Olkin I, Wermuth N, et al., editors. *Proceedings of the First Seattle Symposium in Biostatistics*. Vol. 123. New York, NY: Springer US; 1997. p. 51–84. <https://doi.org/10.1007/978-1-4757-3294-8>.
20. Håvard Kvamme Ørnulf, Borgan I, Scheel. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res*. 2019;20(129):1–30. <https://doi.org/10.48550/arXiv.1907.00825>.
21. Jared L, Katzman et al. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24. <https://doi.org/10.1186/s12874-018-0482-1>.
22. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ*. 2019. <https://doi.org/10.7717/peerj.6257>. 7:e6257, January 2019.
23. Stephane Fotso. (2018). Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework, January 2018. arXiv:1801.05512[cs, stat]. <https://doi.org/10.48550/arXiv.1801.05512>.
24. Lee C, Zame W, Yoon J, Mihaela Van Der Schaar. April, and. DeepHit: a deep learning approach to survival analysis with competing risks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1). <https://doi.org/10.1609/aaai.v32i1.11842>.
25. Emily M, Lima, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nat Commun*. 2021;12(1):5117. <https://doi.org/10.1038/s41467-021-25351-7>.
26. Hassan Ismail, Fawaz, et al. InceptionTime: Finding AlexNet for time series classification. *Data Mining Knowl Discov*. 2020;34(6):1936–62. arXiv:1909.04939[cs, stat]. <https://doi.org/10.48550/arXiv.1909.04939>.
27. He K, Zhang X, Ren S, Sun J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway (NJ): IEEE; 2016. p. 770–8. <https://doi.org/10.1109/CVPR.2016.90>.
28. Krizhevsky A, Sutskever I, Hinton GE. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Vol. 25. La Jolla (CA): Curran Associates, Inc.; 2012. <https://doi.org/10.1145/3065386>.
29. Jennifer A, McCoy et al. (2024). Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. *Am J Obstet Gynecol*. 2024. <https://doi.org/10.1016/j.ajog.2024.04.022>.
30. Hany El-Ghaish and Emadeldeen Eldele. (2024). ECGTransForm: Empowering adaptive ECG arrhythmia classification framework with bidirectional transformer, *Biomedical Signal Processing and Control*, 2024. <https://doi.org/10.1016/j.bspc.2023.105714>.
31. Luz EJdaS, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: a survey. *Comput Methods Programs Biomed*. 2016;127:144–64. <https://doi.org/10.1016/j.cmpb.2015.12.008>.
32. Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, Ghanbari H, Li Q, Sharma A, Clifford GD. Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in cardiology challenge 2021. *Comput Cardiol*. 2021;48:1–4.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.